

**CLiC-it 2020
Italian Conference
on Computational Linguistics**

**Proceedings
of the
Seventh Italian Conference
on
Computational Linguistics**

Bologna, Italy, March 1-3, 2021

Editors:

**Johanna Monti
Felice Dell'Orletta
Fabio Tamburini**



© 2020 by AILC - Associazione Italiana di Linguistica Computazionale
sede legale: c/o Bernardo Magnini, Via delle Cave 61, 38122 Trento
codice fiscale 96101430229
email: info@ai-lc.it

Pubblicazione resa disponibile
nei termini della licenza Creative Commons
Attribuzione – Non commerciale – Non opere derivate 4.0



Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it

isbn 979-12-80136-28-2
www.aAccademia.it/CLIC_2020

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

Table of Contents

Preface.....	1
Keynote Talks and Tutorial	
Fine-grained sentiment analysis: a piece of cake? Veronique Hoste, Ghent University, Belgium.....	3
Interaction-aware multimodal dialogue with conversational agents Stefan Kopp, Bielefeld University, Germany.....	5
Distributional Semantics: Yesterday, Today, and Tomorrow Alessandro Lenci, University of Pisa, Italy	6
Contributed Papers	
Quantitative Linguistic Investigations across Universal Dependencies Treebanks Chiara Alzetta, Felice Dell’Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, Giulia Venturi.....	9
A Machine Learning approach for Sentiment Analysis for Italian Reviews in Healthcare Luca Bacco, Andrea Cimino, Luca Paulon, Mario Merone, Felice Dell’Orletta.....	16
Investigating Proactivity in Task-Oriented Dialogues Vevake Balaraman, Bernardo Magnini	23
A Diachronic Italian Corpus based on “L’Unità” Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, Rossella Varvara	31
Domain Adaptation for Text Classification with Weird Embeddings Valerio Basile.....	37
Personal-ITY: A Novel YouTube-based Corpus for Personality Prediction in Italian Elisa Bassignana, Malvina Nissim, Viviana Patti.....	44
The “Corpus Anchise 320” and the analysis of conversations between healthcare workers and people with dementia Nicola Benvenuti, Andrea Bolioli, Alessandro Mazzei, Pietro Vigorelli, Alessio Bosca	51
Gender Bias in Italian Word Embeddings Davide Biasion, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto.....	58
Automatic Induction of FrameNet lexical units in Italian Silvia Brambilla, Danilo Croce, Fabio Tamburini, Roberto Basili.....	66
Detecting Early Onset of Depression from Social Media Text using Learned Confidence Scores Ana-Maria Bucur, Liviu Dinu	73
Building a Treebank in Universal Dependencies for Italian Sign Language Gaia Caligiore, Cristina Bosco, Alessandro Mazzei.....	79
Analysis of lexical semantic changes in corpora with the Diachronic Engine Pierluigi Cassotti, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro	85
Hate Speech Detection with Machine-Translated Data: The Role of Annotation Scheme, Class Imbalance and Undersampling Camilla Casula, Sara Tonelli	92
UDante: First Steps Towards the Universal Dependencies Treebank of Dante’s Latin Works Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, Marco Passarotti	99
“Spotto la quarantena”: per una analisi dell’italiano scritto degli studenti universitari via social network in tempo di COVID-19 Francesca Chiusaroli, Johanna Monti, Maria Laura Pierucci, Gennaro Nolano	106
Italian Counter Narrative Generation to Fight Online Hate Speech Yi-Ling Chung, Serra Sinem Tekiroğlu, Marco Guerini	115

Surviving the Legal Jungle: Text Classification of Italian Laws in extremely Noisy conditions Riccardo Coltrinari, Alessandro Antinori, Fabio Celli.....	122
Clustering verbal Objects: manual and automatic procedures compared Ilaria Colucci, Elisabetta Ježek, Vít Baisa	128
GePpeTto Carves Italian into a Language Model Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, Marco Guerini	136
Phonological Layers of Meaning: A Computational Exploration of Sound Iconicity Andrea Gregor de Varda, Carlo Strapparava	144
Natural Language Generation in Dialogue Systems for Customer Care Mirko Di Lascio, Manuela Sanguinetti, Luca Anselma, Dario Mana, Alessandro Mazzei, Viviana Patti, Rossana Simeoni.....	151
Cross-Language Transformer Adaptation for Frequently Asked Questions Luca Di Liello, Daniele Bonadiman, Alessandro Moschitti, Cristina Giannone, Andrea Favalli, Raniero Romagnoli.....	157
How good are humans at Native Language Identification? A case study on Italian L2 writings Elisa Di Nuovo, Cristina Bosco, Elisa Corino.....	164
Linguistic Features in Automatic Sarcasm Detection Martina Ducret, Lauren Kruse, Carlos Martinez, Anna Feldman, Jing Peng	171
Risorse e strumenti per le varietà storiche dell’italiano: il progetto TrAVaSI Manuel Favaro, Marco Biffi, Simonetta Montemagni.....	178
AriEmozione: Identifying Emotions in Opera Verses Francesco Fernicola, Shibingfeng Zhang, Federico Garcea, Paolo Bonora, Alberto Barrón-Cedeño...	187
The AEREST Reading Database Marcello Ferro, Sara Giulivi, Claudia Cappa	193
Græcissare: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin Greta Franzini, Federica Zampedri, Marco Passarotti, Francesco Mambrini, Giovanni Moretti	199
L’impatto emotivo della comunicazione istituzionale durante la pandemia di COVID-19: uno studio di Twitter Sentiment Analysis Gloria Gagliardi, Lorenzo Gregori, Alice Suozzi	205
On Knowledge Distillation for Direct Speech Translation Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, Marco Turchi	211
Predicting Social Exclusion: A Study of Linguistic Ostracism in Social Networks Greta Gandolfi, Carlo Strapparava	218
Grounded and ungrounded referring expressions in human dialogues: Language mirrors different grounding conditions Eleonora Gualdoni, Raffaella Bernardi, Raquel Fernández, Sandro Pezzelle.....	223
Predicting movie-elicited emotions from dialogue in screenplay text: A study on “Forrest Gump” Benedetta Iavarone, Felice Dell’Orletta.....	230
Point Break: Surfing Heterogeneous Data for Subtitle Segmentation Alina Karakanta, Matteo Negri, Marco Turchi	236
How granularity of orthography-phonology mappings affect reading development: Evidence from a computational model of English word reading and spelling Alfred Lim, Beth A. O’Brien, Luca Onnis	243
Simple Data Augmentation for Multilingual NLU in Task Oriented Dialogue Systems Samuel Louvan, Bernardo Magnini	250
The E3C Project:Collection and Annotation of a Multilingual Corpus of Clinical Cases Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, Roberto Zanolli	258
Monitoring Social Media to Identify Environmental Crimes through NLP. A preliminary study Raffaele Manna, Antonio Pascucci, Wanda Punzi Zarino, Vincenzo Simoniello, Johanna Monti	265

Does finger-tracking point to child reading strategies? Claudia Marzi, Anna Rodella, Andrea Nadalini, Loukia Taxitari, Vito Pirrelli	272
Multiword expressions we live by: a validated usage-based dataset from corpora of written Italian Francesca Masini, M. Silvia Micheli, Andrea Zaninello, Sara Castagnoli, Malvina Nissim	279
The Style of a Successful Story: a Computational Study on the Fanfiction Genre Andrea Mattei, Dominique Brunato, Felice Dell’Orletta	284
A Multimodal Dataset of Images and Text to Study Abusive Language Stefano Menini, Alessio Palmero Aprosio, Sara Tonelli	290
A Resource for Detecting Misspellings and Denoising Medical Text Data Enrico Mensa, Gian Manuel Marino, Davide Colla, Matteo Delsanto, Daniele P. Radicioni	296
Is Neural Language Model Perplexity Related to Readability? Alessio Miaschi, Chiara Alzetta, Dominique Brunato, Felice Dell’Orletta, Giulia Venturi	303
Italian Transformers Under the Linguistic Lens Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell’Orletta, Giulia Venturi	310
BERTino: an Italian DistilBERT model Matteo Muffo, Enrico Bertino	317
ItaGLAM: A corpus of Cultural Communication on Twitter during the Pandemic Gennaro Nolano, Carola Carlino, Maria Pia di Buono, Johanna Monti	323
Creativity Embedding: a vector to characterise and classify plausible triples in deep learning NLP models Isabeau Oliveri, Luca Ardito, Giuseppe Rizzo, Maurizio Morisio	330
The CREENDER Tool for Creating Multimodal Datasets of Images and Comments Alessio Palmero Aprosio, Stefano Menini, Sara Tonelli	336
(Stem and Word) Predictability in Italian verb paradigms: An Entropy-Based Study Exploiting the New Resource LeFFI Matteo Pellegrini, Alessandra Teresa Cignarella	341
A deep learning model for the analysis of medical reports in ICD-10 clinical coding task Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro	347
Exploring Attention in a Multimodal Corpus of Guided Tours Andrea Amelio Ravelli, Antonio Origlia, Felice Dell’Orletta	353
A Case Study of Natural Gender Phenomena in Translation. A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, Andy Way	359
Multifunctional ISO standard Dialogue Act tagging in Italian Gabriel Roccabruna, Alessandra Cervone, Giuseppe Riccardi	365
Tracing Metonymic Relations in T-PAS: An Annotation Exercise on a Corpus-based Resource for Italian Emma Romani, Elisabetta Ježek	373
Datasets and Models for Authorship Attribution on Italian Personal Writings Gaetana Ruggiero, Albert Gatt, Malvina Nissim	380
The Archaeo-Term Project: Multilingual Terminology in Archaeology Giulia Speranza, Raffaele Manna, Maria Pia Di Buono, Johanna Monti	387
Exploiting Distributional Semantics Models for Natural Language Context-aware Justifications for Recommender Systems Giuseppe Spillo, Cataldo Musto, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro	394
MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion Analysis for Italian Rachele Sprugnoli	402
Becoming JILDA Irene Sucameli, Alessandro Lenci, Bernardo Magnini, Maria Simi, Manuela Speranza	409

How “BERTology” Changed the State-of-the-Art also for Italian NLP Fabio Tamburini.....	415
Valutazione umana di DeepL a livello di frase per le traduzioni di testi specialistici dall’inglese verso l’italiano Mirko Tamosanis, Sirio Papa.....	422
Overprotective Training Environments Fall Short at Testing Time: Let Models Contribute to Their Own Training Alberto Testoni, Raffaella Bernardi.....	429
Topic Modelling Games Rocco Tripodi	435
Dialog-based Help Desk through Automated Question Answering and Intent Detection Antonio Uva, Pierluigi Roberti, Alessandro Moschitti	443
#andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline Pierluigi Vitale, Serena Pelosi, Mariacristina Falco	450
Polarity Imbalance in Lexicon-based Sentiment Analysis Marco Vassallo, Giuliano Gabrieli, Valerio Basile, Cristina Bosco	457
Suoidne-varra-bleahkka-mála-bihkka-senet-dielku ’hay-blood-ink-paint-tar-mustard-stain’ -Should compounds be lexicalized in NLP? Linda Wiecheteck, Chiara Argese, Tommi A Pirinen, Trond Trosterud.....	464
Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams Mehmet Can Yavuz	471



Associazione Italiana di
Linguistica Computazionale



Preface

On behalf of the Program Committee, a very warm welcome to the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020). This edition of the conference is held in Bologna and organised by the University of Bologna. The CLiC-it conference series is an initiative of the Italian Association for Computational Linguistics (AILC) which, after six years of activity, has clearly established itself as the premier national forum for research and development in the fields of Computational Linguistics and Natural Language Processing, where leading researchers and practitioners from academia and industry meet to share their research results, experiences, and challenges.

This year CLiC-it received 80 submissions against 64 submissions in 2015, 69 in 2016, 72 in 2017, 70 in 2018 and 82 in 2019 confirming the increasing trend of the past years. The Programme Committee worked very hard to ensure that every paper received at least three careful and fair reviews. This process finally led to the acceptance of 19 papers for oral presentation and 53 papers for poster presentation, with a global acceptance rate of 90% motivated by the inclusive spirit of the conference. The conference is also receiving considerable attention from the international community, with 17 (21%) submissions showing at least one author affiliated to a foreign institution. Regardless of the format of presentation, all accepted papers are allocated 5 or 6 pages plus 2 pages for references in the proceedings, available as open access publication. In line with previous editions, the conference is organised around thematic areas managed by one or two area chairs per area.

In addition to the technical programme, this year we are honoured to have as invited speakers internationally recognised researchers as Veronique Hoste (Ghent University) and Stefan Kopp (Bielefeld University). We are very grateful to Veronique and Stefan for agreeing to share with the Italian community their knowledge and expertise on key topics in Computational Linguistics.

Traditionally, around one half of the participants at CLiC-it are young postdocs, PhD students, and even undergraduate students. As in the previous edition of the conference, we organised a special track called “Research Communications”, encouraging authors of articles published in 2020 at outstanding international conferences in our field to submit short abstracts of their work. Research communications are not published in the proceedings, but will be orally presented within a dedicated session at the conference, in order to enforce dissemination of excellence in research.

Moreover, during the conference we will award the prize for the best Master Thesis (*Laurea Magistrale*) in Computational Linguistics, submitted at an Italian University between August 1st 2019 and July 31st 2020. This special prize is also endorsed by AILC. We have received 4 candidate theses, which have been evaluated by a special jury. The prize will be awarded at the conference, by a member of the jury.

As last year, we propose a tutorial at the beginning of the conference by Alessandro Lenci. We highlight the importance that this kind of opportunities have for young researchers in particular, and we are proud of having made the tutorial attendance free for all registered students.

Even if CLiC-it is a medium size conference, organizing this annual meeting requires major effort from many people. This conference will not be possible without the dedication, devotion and hard work of the members of the Local Organising Committee, who volunteered their time and energies to contribute to the success of the event. We are also extremely grateful to our Programme Committee members for producing a lot of detailed and insightful reviews, as well as to the Area Chairs who assisted the Programme Chairs in their duties. All these people are named in the following pages. We also want to acknowledge the support from endorsing organisations and institutions and from all of our sponsors, who generously provided funds and services that are crucial for the realisation of this event. Special thanks are also due to the University of Bologna for its support in the organisation of the event.

Please join us at CLiC-it 2020 to interact with experts from academia and industry on topics related to Computational Linguistics and Natural Language Processing and to experience and share new research findings, best practices, state-of-the-art systems and applications. We hope that this year's conference will be intellectually stimulating, and that you will take home many new ideas and methods that will help extend your own research.

Johanna Monti, Felice Dell'Orletta and Fabio Tamburini
CLiC-it 2020 General Chairs

Organising Committee

Conference and Programme Chairs

Johanna Monti (Università di Napoli “L’Orientale”)
Felice Dell’Orletta (Istituto di Linguistica Computazionale “A.Zampolli” – CNR)
Fabio Tamburini (Università di Bologna)

Area Chairs

Dialogue, Discourse and Natural Language Generation
Alessandro Mazzei (Università degli Studi di Torino)
Malvina Nissim (Università di Groningen)

Information Extraction, Information Retrieval and Question Answering
Elena Cabrio (Université Côte d’Azur, Inria, CNRS)
Alessandro Moschitti (Amazon)

Language Resources and Evaluation
Simonetta Montemagni (Istituto di Linguistica Computazionale “A. Zampoli”, CNR)
Viviana Patti (Università degli Studi di Torino)

Language and Cognition
Gloria Gagliardi (Università di Napoli “L’Orientale”)
Alessandro Lenci (Università di Pisa)

Linguistic Issues in CL and NLP
Francesca Chiusaroli (Università degli Studi di Macerata)
Elisabetta Jezek (Università degli Studi di Pavia)

Machine Learning for NLP
Pierpaolo Basile (Università degli Studi di Bari)
Andrea Esuli (Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, CNR)

Machine Translation and Multilinguality
Marcello Federico (Amazon)
Matteo Negri (Fondazione Bruno Kessler, Trento)

Morphology and Syntax Processing
Maria Pia di Buono (Università di Napoli “L’Orientale”)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milano)

NLP for Digital Humanities
Sara Tonelli (Fondazione Bruno Kessler, Trento)
Giulia Venturi (Istituto di Linguistica Computazionale “A. Zampoli”, CNR)

NLP for Web and Social Media
Cristina Bosco (Università degli Studi di Torino)
Paolo Rosso (Universitat Politècnica de València)

Pragmatics and Creativity
Valentina Bambini (Ne.T.S.–IUSS Center for Neurolinguistics and Theoretical Syntax, Pavia)
Marco Guerini (Fondazione Bruno Kessler, Trento)

Research and Industrial NLP Applications
Roberto Basili (Università di Roma, Tor Vergata)
Giovanni Semeraro (Università degli Studi di Bari)

Semantics and Knowledge Representation
Tommaso Caselli (Università di Groningen)
Roberto Navigli (Università di Roma, La Sapienza)

Spoken Language Processing and Automatic Speech Understanding
Barbara Gili Fivela (Università del Salento)
Antonio Origlia (Università Degli Studi di Napoli Federico II)

Vision, Robotics, Multimodal and Grounding
Raffaella Bernardi (Università degli Studi di Trento)
Francesco Cutugno (Università Degli Studi di Napoli Federico II)

Website chair

Andrea Cimino (Istituto di Linguistica Computazionale “A.Zampolli” – CNR)

Local Organisation

Claudia Roberta Combei (Università di Bologna)
Gloria Gagliardi (Università di Napoli “L’Orientale”)
Fabio Tamburini (Università di Bologna)

Reviewers

Lasha Abzianidze, Mehwish Alam, Oscar Araque, Luigi Asprino, Mattia Atzeni, Vevake Balaraman, Francesco Barbieri, Pierpaolo Basile, Valerio Basile, Roberto Basili, Núria Bel, Luisa Bentivogli, Monica Berti, Michele Bevilacqua, Luca Bischetti, Andrea Bolioli, Marianna Bolognesi, Johan Bos, Federico Boschetti, Cristina Bosco, Dominique Brunato, Christian Burgers, Davide Buscaldi, Hendrik Buschmeier, Marina Buzzoni, Elena Cabrio, Erik Cambria, Annalina Caputo, Luigi Di Caro, Giuseppe Castellucci, Flavio Massimiliano Cecchini, Giuseppe Giovanni Antonio Celano, Mauro Cettolo, Francesca Chiusaroli, Yi-Ling Chung, Alessandra Teresa Cignarella, Andrea Cimino, Fabio Ciotti, Giovanni Colavizza, Davide Colla, Claudia Roberta Combei, Sergio Consoli, Anna Corazza, Elisa Corino, Gianpaolo Coro, Danilo Croce, Hoang Cuong, Francesco Cutugno, Rossana Damiano, Marco de Gemmis, Thierry Declerck, Angelo Mario Del Grosso, Marco Del Tredici, Felice Dell’Orletta, Claudio Delli Bovi, Silvia Demartini, Danilo Dessì, Francesca Di Donato, Barbara Di Eugenio, Stefanie Dipper, Mauro Dragoni, Benamara Farah, Marion Fechino, Anna Feltracco, Alejandro Moreo Fernández, Elisabetta Fersini, Simone Filice, Franz Fischer, Greta Franzini, Diego Frassinelli, Simona Frenda, Francesca Frontini, Matteo Gabburo, Vincenzo Galatà, Lorenzo Gatti, Emiliano Giovannetti, Elisabetta Gola, Christophe Gravier, Elisa Guadagnini, Eleonora Gualdoni, Marco Guerini, Shohreh Haddadan, Christian Hardmeier, Sadid A. Hasan, Delia Irazu Hernandez Farias, Ignacio Iacobacci, Carlos A. Iglesias, Andrea Iovine, Amy Isard, Elisabetta Jezek, Roman Klinger, Kristina Kocijan, Dimitrios Kokkinakis, Sandra Kübler, Surafel Lakew, Gianluca Lebani, Luca Di Liello, Francesca Strik Lievers, Eleonora Litta, Pasquale Lops, Samuel Louvan, Simone Magnolini, Francesco Mambrini, Diego Marcheggiani, Maria Di Maro, Mirko Marras, Marco Maru, Prashant Mathur, Stefano Menini, Massimo Moneglia, Paolo Monella, Johanna Monti, Cataldo Musto, Maria Nadejde, Federico Nanni, Fedelucio Narducci, Massimo Nicosia, Nicole Novielli, Alessandro Oltramari, Antonio Origlia, Francesco Osborne, Petya Osenova, Maria Palmerini, Alessio Palmero Aprosio, Ludovica Pannitto, Alessandro Panunzi, Patrick Paroubek, Marco Passarotti, Diego Pescarini, Vito Pirrelli, Marco Polignano, Simone Paolo Ponzetto, Daniele Puccinelli, Valeria Quochi, Daniele P. Radicioni, Daniele Radicioni, Alessandro Raganato, Andrea Amelio Ravelli, Diego Reforgiato, Kevin Roitero, Marco Rospocher, Gaetano Rossiello, Irene Russo, Manuela Sanguinetti, Giorgio Satta, Loredana

Schettino, Flavia Sciolette, Marco S.G. Senaldi, Marco Senaldi, Maria Simi, Luca Soldaini, Francesca Spezzano, Damiano Spina, Rachele Sprugnoli, Carlo Strapparava, Irene Sucameli, Fabio Tamburini, Mariona Taulé, Serra Sinem Tekiroglu, Maurizio Tesconi, Francesca Tomasi, Rocco Tripodi, Emiel Vanmilttenburg, Rossella Varvara, Guido Vetere, Serena Villata, Fabio Massimo Zanzotto.

Keynote Talks and Tutorial

Fine-grained sentiment analysis: a piece of cake?



Veronique Hoste

LT3 Language and Translation Technology Team
Department of Translation, Interpreting and Communication
Ghent University
veronique.hoste@ugent.be

With the emergence of the interactive Web 2.0, the amount of opinionated online text has grown immensely, as well as the interest in exploiting that information. At the same time, digitization and globalization have profoundly changed the media ecology, with an increasing trend to consume news online, more specifically via newspaper websites or through secondary gatekeepers on social media platforms, etc.

This availability of online social and curated text sources has led to a boost in sentiment analysis research, which mainly took off in early 2000 with as initial objective the identification of semantic polarity (positive, negative, or neutral) of a given text. In the last years this primary objective has evolved into a more fine-grained paradigm. This includes identifying the entity towards which a given sentiment is aimed in aspect-based sentiment analysis (Pontiki et al., 2016), identifying emotions instead of mere polarity orientations (Mohammad et al., 2018) or modeling the implicit sentiment certain events or facts convey or evoke. In this talk, I will focus on some ongoing projects in our team in which we seek to model sentiment and emotions at this fine-grained level.

Taking the aspect-based sentiment analysis framework as a starting point, I will broaden the scope from aspects to unrestricted news events and discuss our attempts to model fine-grained news events' polarity in general and economic hard news. As factual utterances often do not contain explicitly lexicalized sentiment, but rather describe “polar facts” or real-world events or objects with implied affective information, I will mainly focus on the challenges involved in modeling implicit sentiment. This implicit or prototypical polarity modeling has also been key to our work on verbal irony and the associated SemEval-2018 shared task. As irony is frequently realized through a clash between (often) explicit opinion words and a prototypically negatively connoted activity, I will discuss how we seek to model this contrast (Van Hee et al., 2018).

Finally, in view of more refined emotion modelling in text, I will elaborate on our first steps in the domain of emotion detection, starting with our pursuit of a reliable method to label emotional properties in text (De Bruyne et al., to appear). Furthermore, I will discuss the problem of choosing an appropriate framework for building an emotion-annotated corpus and our experiments on transfer learning for emotion detection in a less-resourced scenario.

De Bruyne, L., De Clercq, O. & Hoste, V. (to appear). Annotating Affective Dimensions in User-Generated Content. Comparing the reliability of best-worst scaling, pairwise comparison and rating scales for annotating valence, arousal and dominance. *Language Resources and Evaluation*.

Mohammad, S.M., Bravo-Marquez, F., Salameh, M. & Kiritchenko, S. (2018). Semeval-2018 Task 1: Affect in tweets. *Proceedings of SemEval-2018*.

Pontiki, M., Galanis, D., et al. (2016). SemEval-2016 task 5: aspect based sentiment analysis. *Proceedings of SemEval-2016*.

Van Hee, C., Lefever, E., & Hoste, V. (2018). We usually don't like going to the dentist: using common sense to detect irony on Twitter. *Computational Linguistics*, 44(4), 793–832.

Interaction-aware multimodal dialogue with conversational agents

Stefan Kopp

Social Cognitive Systems
Faculty of Technology and CITEC
Bielefeld University
skopp@techfak.uni-bielefeld.de

Spoken language-based interfaces and dialogue systems are on the rise thanks to advances in Machine Learning. However, we are still far away from being able to develop and deploy social robots or virtual assistants that are capable of fluent, flexible and robust communication in cooperative tasks with human users. One main challenge that still remains is the modeling of face-to-face dialogue with embodied conversational agents, i.e. the orchestration of human-like understanding and generation of multi-channel, multi-functional, and multi-modal communication embedded in a coherent and fluent dialogue. I will present work that pursues this goal by bringing together data-based and model-based approaches in a framework for „interaction-aware“ dialogue processing. Issues covered will include the realtime adaptation to communication partners at different levels of dialogue processing, the use of semantic and pragmatic functions of nonverbal communication, and the perception of such conversational agents by human users in different contexts of use.

Distributional Semantics: Yesterday, Today, and Tomorrow

Alessandro Lenci

University of Pisa

`alessandro.lenci@unipi.it`

Distributional semantics is undoubtedly the mainstream approach to meaning representation in computational linguistics today. It has also become an important paradigm of semantic analysis in cognitive science, and even linguists have started looking at it with growing interest. The popularity of distributional semantics has literally boomed in the era of Deep Learning, when “word embeddings” have become the basic ingredient to “cook” any NLP task. The era of BERT & co. has brought new types of contextualized representations that have often generated hasty claims of incredible breakthroughs in the natural language understanding capability of deep learning models. Unfortunately, these claims are not always supported by the improved semantic abilities of the last generation of embeddings. Models like BERT are still rooted in the principles of distributional learning, but at the same time their goal is more ambitious than generating corpus-based representations of meaning. On the one hand, the embeddings they produce encode much more than lexical meaning, but on the other hand we are still largely uncertain about what semantic properties of natural language they actually capture. Distributional semantics has surely benefited from the successes of the deep learning, but this might even jeopardize the very essence of distributional models of meaning, by making their goals and foundations unclear.

Computational linguistics is a fast-moving field and distributional semantics makes no exception. In doing this, we always risk chasing the last hype model or using pre-trained vectors as black-box tools, without scrutinizing the relationship between distributional learning and meaning representations. The goal of this tutorial is to try to understand what distributional semantics is today, by looking also at what it was yesterday and at its grounding principles. I will present the main concepts, tools and applications of distributional semantics, to foster a critical analysis of its potentialities as well as its limits. This way, we will try to imagine what distributional semantic could and should become tomorrow.

Contributed Papers

Quantitative Linguistic Investigations across Universal Dependencies Treebanks

Chiara Alzetta* ♣, Felice Dell’Orletta*, Simonetta Montemagni*,
Petya Osenova† ♣, Kiril Simov†, Giulia Venturi*

* Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR, Pisa

♣ DIBRIS, Università degli Studi di Genova, ♣ Sofia University

† Artificial Intelligence and Language Technologies Department, IICT-BAS

chiara.alzetta@edu.unige.it, {petya, kivs}@bultreebank.org,

{felice.dellorletta, simonetta.montemagni, giulia.venturi}@ilc.cnr.it

Abstract

The paper illustrates a case study aimed at identifying cross-lingual quantitative trends in the distribution of dependency relations in treebanks for typologically different languages. Preliminary results show interesting differences rooted either in language-specific peculiarities or cross-lingual annotation inconsistencies, with a potential impact on different application scenarios.¹

1 Introduction and Motivation

The identification of cross-lingual quantitative trends in the distribution of dependency relations in “gold” treebanks is increasingly attracting the interest of the computational linguistics community for different purposes, as testified e.g. by a recently published miscellaneous book on the quantitative analysis of dependency structures (Jiang and Liu, 2018) or pilot initiatives such as the first edition of the workshop “Quantitative Syntax 2019”². Among possible applications, it is worth mentioning studies aimed at acquiring typological evidence to be integrated in multilingual NLP algorithms (see Ponti et al. (2018) for a survey and the workshop “Typology for Polyglot NLP”³), or at detecting annotation inconsistencies to improve the quality of treebanks (see (Dickinson, 2015; de Marneffe et al., 2017) to mention only a few). While the latter is a well-established research topic, although with still many open issues, automatically acquiring typological information is still at its beginning, so automatic strategies to extract such information from corpora are

needed (Cotterell and Eisner, 2017; Bjerva and Augenstein, 2018).

Multilingual resources such as the dependency treebanks developed within the Universal Dependencies (UD) project⁴, thanks to the cross-linguistically consistent syntactic annotation (Nivre, 2015), fostered the development of automatic strategies to extract cross-lingual similarities and differences in shared constructions from corpora (Murawaki, 2017; Bjerva et al., 2019). Within this line of research, the paper describes a methodology for comparing treebanks of typologically different languages with the final aim of detecting and quantifying similarities and differences in multilingual treebanks analyzed from a twofold perspective: language-specific peculiarities vs cross-lingual annotation inconsistencies. To this end, we used LISCA (*Linguistically-driven Selection of Correct Arcs*) (Dell’Orletta et al., 2013), an algorithm which has been successfully applied in different scenarios, against both the output of dependency parsers and gold treebanks. In the first case, the score returned by LISCA was meant to identify unreliable automatically produced dependency relations (Dell’Orletta et al., 2013). When used against gold annotations, LISCA was used to detect shades of syntactic markedness of syntactic constructions in manually annotated corpora from a monolingual perspective (Tusa et al., 2016), or to acquire quantitative typological evidence from a multilingual perspective (Alzetta et al., 2018b). Last but not least, it was also exploited to identify anomalous annotations (going from annotation inconsistencies to errors) from a monolingual perspective in gold treebanks (Alzetta et al., 2018a).

The methodology exploited for the present work (described in Section 2) was tested in a case study carried out on four Indo-European languages belonging to three different genera (according

¹Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.aclweb.org/anthology/W19-79.pdf>

³<https://typology-and-nlp.github.io/>

⁴<https://universaldependencies.org/>

to WALS classification, Dryer and Haspelmath (2013)): Bulgarian (Slavic, BUL), English (Germanic, ENG), Italian and Spanish (Romance, ITA and SPA). UD treebanks constitute an ideal test bed for our analysis since, sharing the same annotation scheme, allow the investigation of cross-lingual similarities and differences in shared constructions. Besides similarities connected with the UD annotation strategy aimed at maximising parallelism across languages, results in Section 4 reflect shared possibly “universal” features of languages. Differences, in turn, can either reflect typologically relevant language peculiarities or highlight inconsistencies in the application of the shared annotation scheme. The paper focuses on both aspects. Section 5 concludes the paper discussing our findings and future directions of research.

Contribution. The present contribution has two main goals: we aim to show how the methodology can be used 1) to acquire quantitative evidence of cross-linguistically shared properties, and 2) to highlight divergences due either to language idiosyncrasies or annotation inconsistencies across treebanks.

2 Method

As shown in Figure 1, our methodology for exploring multilingual treebanks is articulated in the following two steps.

I) LISCA Analysis. The LISCA algorithm operates in two steps: 1) it collects statistics about a set of linguistically motivated features extracted from an automatically dependency parsed corpus (referred to as *Reference Corpus*) to build a statistical model (SM) of the language; 2) it uses the obtained SM to assign a score to each dependency relation (DR) instance, defined as a triple $d(\text{ependent}), h(\text{ead}), t(\text{ype})$ of dependency linking d to h , in a *Target Corpus*. Borrowing a metaphor from Jakobson (1973), we can look at the SM as encoding the DNA of the language being analysed. Note, in fact, that the features considered by the LISCA algorithm to build the SM cover, for each DR instance, a wide variety of factors, both local and global. *Local features* include e.g. the distance in terms of tokens between d and h , the associative strength linking the grammatical categories involved in the relation (i.e. POS_d and POS_h), the POS of the head governor, the type of dependency connecting d to h , and the relative linear order of

d and h in the sentence. *Global features*, instead, are aimed at locating each DR within the overall sentence structure, and include e.g. the distance of d from the root of the dependency tree or from the closest or most distant leaf node, and the number of “brother” and “children” nodes of d , occurring respectively to its right or left in the linear sequence of words of the sentence. In this case study, LISCA has been used in its delexicalized version in order to abstract away from variations resulting from lexical effects, thus guaranteeing cross-lingual comparability of results. The output of LISCA consists of the list of all DRs in the Target Corpus ranked by decreasing score.

The LISCA score is a context-sensitive and frequency-based measure reflecting the degree of similarity of the “linguistic environments” in which a given DR occurs in the Reference and Target corpora: it encodes the probability to observe a DR instance occurring in a specific context on the basis of the Statistical Model constructed starting from the Reference Corpus. In more abstract terms, the LISCA score can be seen as reflecting the prototypicality degree of a specific linguistic structure: whereas higher LISCA scores identify DR instances appearing in “typical” (more frequent and likely) contexts with respect to the statistics acquired from the Reference Corpus, lower scores identify less common or even atypical DR instances of the Target Corpus. From a multilingual perspective, the comparison of the ranked DRs lists obtained from corpora of different languages can shed light on similarities and differences at linguistic and/or annotation levels. To carry out this comparative analysis, in this study the ranked list of DRs has been split into 20 intervals of equal size, henceforth “bins” (plus a further bin for the remaining ones): the first bins contain DRs presenting a high LISCA score and, conversely, the last bins contain DRs associated with low LISCA scores.

II) Ranking Exploration. We exploited CLaRK system (Simov et al., 2004) to identify and compare quantitative trends from LISCA rankings. CLaRK system work-flow is the following: firstly, each Target Corpus is converted from the CoNLL-U format⁵ into XML format, then the XPath language is used to select the nodes (sentences or tokens) with the required properties. In this way we

⁵<http://universaldependencies.org/format.html>

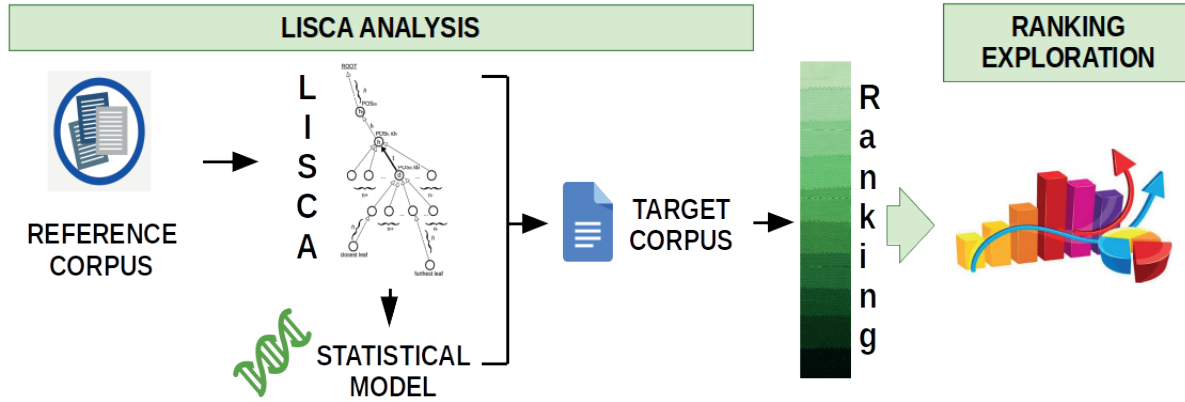


Figure 1: Method work-flow.

can define different configurations and check the distribution of the node characteristics along the DR rankings.

3 Data

For each language taken into account, two linguistically annotated corpora have been used: a large Reference Corpus and a Target Corpus.

Each *Reference Corpus* consists of a monolingual corpus of texts from the news and Wikipedia domains of around 40 million tokens, constituting a set of examples large enough to reflect the actual distribution of phenomena in the specific language. Reference corpora were morpho-syntactically annotated and dependency parsed by the UDPipe pipeline (Straka et al., 2016) trained on the Universal Dependency treebanks, version 2.2 (Nivre et al., 2017).

Target corpora correspond here to manually validated (“gold”) Universal Dependencies treebanks (v2.2). Specifically, we considered the following UD treebanks:

- i) English Web Treebank (254,830 tokens and 16,622 sentences) (Silveira et al., 2014);
- ii) Italian Stanford Dependency Treebank (278,429 tokens and 14,167 sentences) (Bosco et al., 2013);
- iii) Spanish UD treebank (547,680 tokens and 17,680 sentences) (Alonso and Zeman, 2016);
- iv) UD_Bulgarian-BTB (156,149 tokens and 11,138 sentences) (Simov et al., 2005).

4 Results

Results are analysed from a twofold perspective, focusing on the distribution across the bins of different DR types and structures.

4.1 Ranking of Dependencies

As pointed out above, higher LISCA scores are assigned to DRs that show a linguistic context highly typical for the language, whereas low scores are associated with atypical (or simply less typical) syntactic structures; (un)typicality is assessed here with respect to the statistics acquired from the Reference Corpus.

As a first step of our comparative analysis, for each language we focused on the distribution of individual DRs across the 20 LISCA bins. Figure 2 reports the median bin of occurrence for all 29 shared DRs in the ranking of each language. The median bin was selected by sorting all instances of a given DR on the basis of the associated LISCA score and by identifying the median element of the ranked list: its bin of occurrence was taken as representative of the relation. Top and bottom relations (respectively at the extreme left and right in Fig.2 graph) in language-specific rankings show interesting similarities: if on the one hand DRs involving function words (e.g. `case`, `det`, `aux(:pass)`) are associated with higher LISCA scores for all languages, on the other hand special or “loose” DRs such as `orphan` and `parataxis` or clausal subjects and adverbial clauses (`csubj(:pass)`, `advcl`) all occur in the last bins, representing relations with more variable contexts across all languages. Another cross-language parallelism concerns the relative rankings of subsets of DRs: clausal complements with obligatory control (`xcomp`) are assigned a higher score with respect to the wider class of clausal complements without it (`ccomp`); the direct object relation (`obj`) precedes in the ranking the oblique argument/modifier (`obl`); and the nominal subject (`nsubj`) always precedes its

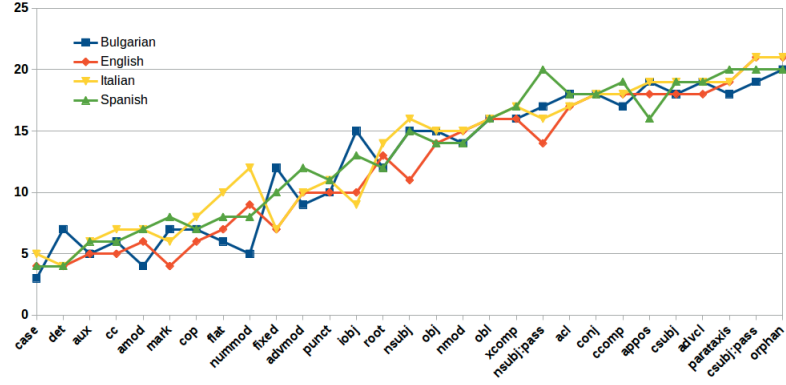


Figure 2: Median occurrence in LISCA bins of shared DRs across languages.

DR length	Bins 1 to 10				Bins 11 to 20			
	BUL	ENG	ITA	SPA	BUL	ENG	ITA	SPA
1	66.42	55.80	65.79	69.16	38.96	37.62	36.97	37.34
2	20.26	23.31	24.93	21.62	16.69	17.93	17.05	17.47
3-4	6.45	12.24	5.16	5.21	19.45	21.11	15.85	16.31
5-10	4.75	4.72	2.71	2.01	14.34	14.09	13.05	13.34
≥ 10	2.39	3.93	1.41	1.98	10.56	9.25	17.08	15.54
# DRs	64,885	110,184	141,389	248,794	22,510	42,475	44,310	96,781

Table 1: Percentage distribution by length of DRs involving leaves in the first and in last 10 LISCA bins. For each group of bins the number of all DRs involving leaves is given.

clausal counterpart (*csubj*). It is interesting to report that the frequency of a DR seems to play a minor role in determining the position of a given DR in the LISCA ranking: consider, for instance, the *punct* relation which is a highly frequent DR (covering around 11% of DRs in all four languages), but nevertheless it was placed in the middle part of the ranking for all languages. Looked at from this perspective, the LISCA ranking of relations - which is heavily influenced by the principles underlying the UD annotation schema - seems to reflect the parsing complexity of relations (Alzetta et al., 2020), where more complex to parse DRs are characterised by a higher variability in their contexts of occurrence.

Some interesting differences can also be reported, originating either in a) language-specific peculiarities or b) possibly inconsistent annotations across languages. Concerning a), ENG nominal subjects (*nsubj*, *nsubj:pass*) are ranked significantly higher with respect to the other three languages, all sharing the pro-drop and free word order properties; or determiners (*det*) show the same distribution for SPA, ENG and ITA in contrast to BUL, where the definite article is post-positioned and expressed morphologically, with the exception of some pronouns functioning as de-

terminers, e.g. demonstratives. Here are two examples for Bulgarian where the first one shows the usage of the morphologically expressed post-positioned definite article (thus no explicit (*det*) relation) while the second shows the usage of a demonstrative pronoun (marked with (*det*) relation)): (1) (*Жената влезе в стаята*) (*lit. Woman-the entered room-the*) and (2) (*Този жена влезе в стаята*) (*lit. This woman entered room-the*). The frequency of the examples type (1) in the treebank is about 10 times bigger than the frequency of the examples of type (2). Thus, the *nsubj* nodes modified by explicit determiner word is a rare case in Bulgarian treebank.

With respect to b), there are interesting examples, even among core UD DRs: this is the case of indirect objects (*iobj*), whose annotation criteria highly diverge across languages. The sources of dissimilarities might come partially from the annotation specifications per language about what a second argument (*iobj*) vs an adjunct (*obl*) is. If a closer look is taken into the data, it turns out that in ITA and ENG the *iobj* is typically expressed by a PRON(oun), as in these two examples: ITA: *‘ti (PRON) ho dato’* (*lit. ‘I gave you’*); ENG: *‘causing us (PRON) trouble’*. In ITA this represents 100% of the cases, while in ENG 84%,

whereas in SPA and BUL this relation is expressed by a pronoun in only 46.7% and 19% of the cases respectively. In Spanish, for example, the *iobj* relation is used also for NOUNs: in the Spanish example ‘*Obligarón al Gobierno* (NOUN) *a comprar créditos*’ (lit. *Forced the Government to buy credits*) the noun is annotated as indirect object of *obligaron*, whereas in Italian the construction ‘*Non ho dato soldi al presidente* (NOUN)’ (lit. *I didn’t give money to the president*) the noun is marked as *obl* relation. In Bulgarian the *iobj* relation is used not only for marking the dative pronouns, but also for marking head NOUNs in PPs. The prevalence of this relation on NOUNs is due to the following factors: (1) the existence of long dative counterparts to short dative pronouns that consist of a preposition and a noun (‘Майката даде играчка на детето’) (prep NOUN) (lit. *Mother-the gave toy to child-the-DAT*); and (2) the marking of indirect complements as indirect objects, while the *obl* relation has been reserved for adjuncts (‘Те продължават да участват в лотарията’) (non-dative prep NOUN) (lit. *They continue to participate in lottery-the*). This suggests that different annotation criteria guide the assignment of the *iobj* DR, possibly not all of them originating in peculiarities of the language.

Other interesting examples concern the annotation of multi-word expressions and proper names (*fixed* and *flat*), which are treated differently across languages. For example, in BUL all grammatically fixed multi-words, such as complex prepositions (like *с оглед на* ‘with regard to’) or conjunctions (like *за да* ‘in order to’), are treated as *fixed* while in Italian the annotation reflects the underlying syntactic structure, as in the case of, e.g., ‘*a base di*’ (lit. *made of*) and ‘*in relazione a*’ (lit. *in relation to*).

4.2 Distribution of Leaves

For each language, we investigated the distribution of DRs across the LISCA bins focusing on DRs involving leaves as dependants (henceforth *leaves*), as opposed to DRs without leave nodes (henceforth *non-leaves*). Results of this analysis are reported in Table 1. Despite minor differences, all languages share a similar trend: leaves are mostly ranked in the first 10 bins representing for Bulgarian 91.52% of the DRs occurring in them, 95.56% for English, 98.27% for Italian and

91.76% for Spanish. Interestingly, the first 6, 6, 8 and 4 bins respectively for Bulgarian, English, Italian and Spanish contain exclusively leaves. In other words, leaves are typically associated with higher LISCA scores: due to their smaller context, they are characterised by higher processing reliability. This is in line with the fact that DRs involving functional words, e.g. *case*, *det*, *aux*, etc. typically occur in the first bins (see Figure 2). On the contrary, the last 10 bins of all languages mostly contain DRs not involving leaves (68.28% BUL, 63.54% EN, 69.33% ITA, 64.54% SP). For what concerns the leaves in the second half of the bins, they turned out to be typically involved in particularly complex syntactic contexts, such as long distance dependencies or occurring in constructions that are not typical for that relation.

5 Conclusion

In this paper we presented method for studying the distribution of DRs in gold treebanks which was tested in a case study carried out on four languages belonging to three different genera. The cross-lingual comparison of the LISCA-based ranking of UD relations across the bins shows: on the one hand, shared (possibly universal) trends, concerning e.g. the similar distribution of dependencies involving leaves or of long distance dependencies, which are respectively concentrated at the top and at the bottom of the LISCA ranking for each language; on the other hand, recorded differences in the ranking of relations can be explained in terms of either language peculiarities (e.g. the pro-drop property of BUL-ITA-SPA vs ENG, or the surface realisation of definite determiners in BUL vs ENG-ITA-SPA) or potential inconsistencies in the application of the UD annotation scheme (see the case of the indirect object relation). Both types of results play a potentially key role in different scenarios, going from typology-driven multilingual NLP to the improvement of the cross-lingual consistency of treebanks.

Acknowledgements

Thanks to the anonymous reviewers for their helpful comments. This work was partially supported by the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage*, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant num-

References

- H. M. Alonso and D. Zeman. 2016. Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.
- C. Alzetta, F. Dell’Orletta, S. Montemagni, and G. Venturi. 2018a. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 201–210, Prague, Czech Republic, January.
- C. Alzetta, F. Dell’Orletta, S. Montemagni, and G. Venturi. 2018b. Universal dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the 11th Edition of International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4540–4549. Association for Computational Linguistics.
- C. Alzetta, F. Dell’Orletta, S. Montemagni, and G. Venturi. 2020. Uncovering typological context-sensitive features. In *Proceedings of the Second Workshop on Typology for Polyglot Natural Language Processing*.
- Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.
- Johannes Bjerva, Yova Kementchedjheva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- C. Bosco, S. Montemagni, and M. Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August.
- Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192.
- M.C. de Marneffe, M. Grioni, J. Kanerva, and F. Ginter. 2017. Assessing the Annotation Consistency of the Universal Dependencies Corpora. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2007)*, pages 108–115, Pisa, Italy, September.
- F. Dell’Orletta, G. Venturi, and S. Montemagni. 2013. Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas*, 2:125–136.
- M. Dickinson. 2015. Detection of Annotation Errors in Corpora. *Language and Linguistics Compass*, 9(3):119–138.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Roman Jakobson. 1973. *Essais de linguistique générale t. 2: rapports internes et externes du langage*. Les éditions de Minuit.
- J. Jiang and H. Liu. 2018. *Quantitative Analysis of Dependency Structures*. De Gruyter Mouton, Berlin, Boston.
- Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461.
- J. Nivre, A. Željko, A. Lars, and et alii. 2017. Universal dependencies 2.0. In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University*.
- J. Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing - Proceedings of the 16th International Conference, CICLing 2015, Part I*, pages 3–16, Cairo, Egypt, April.
- E.M. Ponti, H. O’Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.
- N. Silveira, T. Dozat, M.C. de Marneffe, S. Bowman, M. Connor, J. Bauer, and C.D. Manning. 2014. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.
- K. Simov, A. Simov, H. Ganey, K. Ivanova, and I. Grigorov. 2004. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238.
- K. Simov, P. Osenova, A. Simov, and M. Kouylekov. 2005. Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation. Special Issue*, pages 495–522.
- M. Straka, J. Hajic, and J. Strakova. 2016. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of*

the Tenth International Conference on Language Resources and Evaluation (LREC).

- E. Tusa, F. Dell’Orletta, S. Montemagni, and G. Venturi. 2016. Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, pages 3–16, Napoli, Italy, December.

A Machine Learning approach for Sentiment Analysis for Italian Reviews in Healthcare

Luca Bacco^{1,2,3}, Andrea Cimino², Luca Paulon³, Mario Merone¹, Felice Dell’Orletta²

¹Università Campus Bio-Medico (UCBM)

² Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC–CNR), ItaliaNLP Lab

³Webmonks s.r.l.

{l.bacco, m.merone}@unicampus.it

paulon@webmonks.it

{andrea.cimino, felice.dellorletta}@ilc.cnr.it

Abstract

In this paper, we present our approach to the task of binary sentiment classification for Italian reviews in healthcare domain. We first collected a new dataset for such domain. Then, we compared the results obtained by two different systems, one including a Support Vector Machine and one with BERT. For the first one, we linguistic pre-processed the dataset to extract hand-crafted features exploited by the classifier. For the second one, we over-sampled the dataset to achieve better results. Our results show that the SVM-based system, without the worry of having to oversample, has better performance than the BERT-based one, achieving an F1-score of 91.21%.

1 Introduction

Nowadays, when people want to buy a product or service, they often rely on online reviews of other buyers/users (think of online sales giants like Amazon). Likewise, patients are increasingly relying on reviews on social media, blogs and forums to choose a hospital where to be cured. This behaviour is occurring not only abroad (Greaves et al., 2012; Gao et al., 2012), but also in Italy. This is also demonstrated by the increasing amount of reviews in QSalute¹, one of the most popular Italian ranking websites in healthcare. These reviews are often ignored by hospital companies, which do not exploit the potential of such data to understand patients’ experiences and consequently improve their services. Due to the large amount of data, there is a need for automatic

analysis techniques. To meet these needs, we decided to introduce a sentiment analysis system based on machine learning techniques, in order to classify whether a review has positive or negative sentiment. Since such systems require annotated data, the first step was to build a brand-new dataset. We present it in the next section. Then, we developed two systems based on two different classifiers described in Section 3 together with the features extracted from the text. In Sections 4 and 5 we show the experiments conducted during this study, the obtained results and their discussion. Finally, the last section provides concluding remarks and some possible future developments. While there exist several works on affective computing in several domains for the Italian language (Basile et al., 2018; Cignarella et al., 2018; Barbieri et al., 2016), at the time we are writing there are no references in literature that address this particular domain in Italian. Thus, for the best of our knowledge, this is the first work of sentiment analysis on Italian reviews in healthcare.

2 Dataset

QSalute is an Italian portal where users share their experiences about hospitals, nursing homes and doctors. We have collected a total of 47,224 documents (i.e. reviews). Each document consists of the free text of the review and other metadata such as the document id, the disease area to which the document belongs and the title. In addition, among the provided metadata there is the average grade, i.e. the mean over the votes in four categories: Competence, Assistance, Cleaning and Services.

In this work, documents with an average grade less than or equal to 2 were assigned to the negative class (-1), while documents with an average grade greater than or equal to 4 were assigned to the positive class (1). The remaining documents were labelled with the neutral class (0). The

Corresponding author: Mario Merone

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹www.qsalute.it

dataset is strongly unbalanced towards the positive class: 40641 reviews for the positive class, 3898 for the neutral class and 2685 for the negative class. However, in this work, neutral reviews were discarded thus resulting in a dataset composed by 43326 reviews. The following analyses are then referred to this subset: in Table 1 we report some features of the dataset for each site (i.e. the disease area), while the distribution of tokens over their length is reported in Figure 1.

Site	Positive / Total	Lexicon	Overlap(%)
Nervous System	9984 / 10595	34827	69.93
Hearth	5297 / 5491	22677	79.27
Haematology	353 / 377	5336	93.91
Endocrinology	630 / 699	7417	92.40
Endoscopy	1342 / 1484	12046	88.31
Facial	757 / 791	7686	92.13
Genital	2365 / 2552	15605	85.33
Gynaecology	2115 / 2293	14438	90.57
Infections	187 / 220	4001	94.98
Ophthalmology	2167 / 2339	13449	85.43
Oncology	5732 / 6033	25178	79.70
Otorhinology	1156 / 1227	9738	89.91
Skin	763 / 883	8442	90.43
Plastic Surgery	766 / 795	8026	92.04
Pneumology	824 / 982	9454	90.09
Rheumatology	528 / 598	7239	92.14
Senology	3644 / 3783	17497	87.99
Thoracic Surgery	1131 / 1225	10214	90.59
Vascular Surgery	900 / 959	9157	90.05

Table 1: Dataset features for each site. In the first column are reported the name of sites, in the second column are reported the number of positive reviews whit respect to the total numbers of reviews, while in the third one are reported the lexicon values in terms of the number of unique words. Furthermore, in the last column are reported the lexicon overlap (in percentage) of each site with respect to all the others.

The dataset is released on: www.github.com/lbacco/Italian-Healthcare-Reviews-4-Sentiment-Analysis.

3 Methods

We developed two systems based on two state-of-the-art classifiers from the state-of-the-art for sentiment analysis, Support Vector Machine and BERT. In this Section, we present the implemented classifiers.

3.1 System 1 based on Support Vector Machine (SVM)

In order to build the first system, we followed the approach proposed by (Mohammad et al., 2013) for the sentiment analysis of English tweets and we adapted it for Italian reviews in healthcare. More precisely, we implemented a Support Vec-

tor Machine (SVM) classifier with linear kernel, in terms of liblinear (Fan et al., 2008) rather than libsvm in order to scale better to large numbers of samples, as also reported in the documentation² of the LinearSVC model.

Firstly, all documents pass through a pre-processing pipeline, consisting of a sentence splitter, a tokenizer and a Part-Of-Speech (POS) tagger (all of these tools have been previously developed by the *ItaliaNLP*³ laboratory). Then, documents pass through a step of feature extraction, illustrated in the next section.

3.1.1 Feature Extraction

All features were chosen due to their effectiveness shown in several tasks for sentiment classification for Italian (Cimino and Dell’Orletta, 2016). We refer to these features under the name of *hand-crafted* features and *embedding* features.

Raw and Lexical Text Features

- **(Uncased) Word n -grams:** presence or absence of contiguous sequences of n tokens in the document text, with $n=\{1, 2, 3\}$.
- **Lemma n -grams:** presence or absence of contiguous sequences of n lemmas occurring in the document text, with $n=\{1, 2, 3\}$.
- **Character n -grams:** presence or absence of contiguous sequences of n characters occurring in the document text, with $n=\{2, 3, 4, 5\}$.
- **Number of tokens:** total number of tokens of the document.
- **Number of sentences:** total number of sentences of the document.

Morpho-syntactic Features

- **Coarse-grained Part-Of-Speech n -grams:** presence or absence of contiguous sequences of n grammatical categories, with $n=\{1, 2, 3\}$.
- **Fine-grained Part-Of-Speech n -grams:** presence or absence of contiguous sequences of n (fine-grained) grammatical categories, with $n=\{1, 2, 3\}$.

²www.scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

³www.italianlp.it

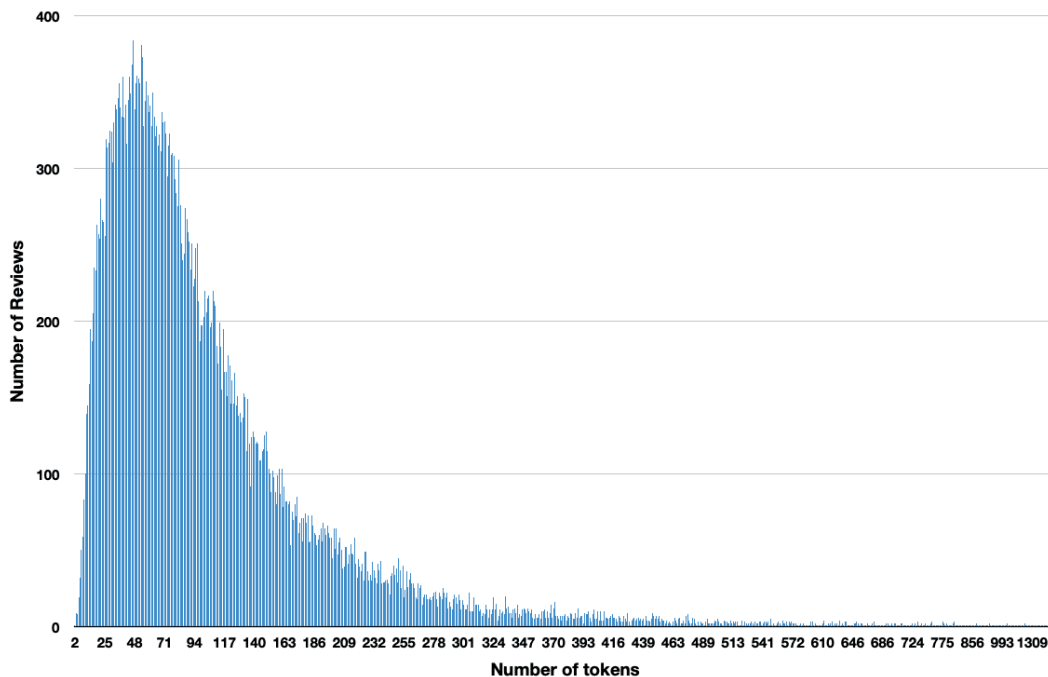


Figure 1: Distribution of documents according to their length, in terms of the number of tokens. The shortest document has only two tokens, while the longest one has 3571 tokens. On average, the reviews are 106.41 tokens long, with a standard deviation of 102.18 tokens.

Word Embeddings Combination: this feature is composed of three vectors. Each vector was calculated by the mean over word embeddings belonging to a specific fine-grained grammatical category: adjectives (excluding possessive adjectives), nouns (excluding abbreviations), and verbs (excluding modal and auxiliary verbs). Word embeddings used in this work are vectors of 128 dimensions, and they were extracted from a corpus of more than 46 million tweets. Such embeddings were already used in (Cimino et al., 2018) and they are available for download at the website of *ItaliaNLP*⁴. Furthermore, three features have been added to indicate the absence of word embeddings belonging to such categories, for a total of 387 ($128 * 3 + 3$) features.

3.2 System 2 based on BERT

We also implemented Bidirectional Encoder Representations from Transformers, or as better known, BERT, to classify the sentiment of the reviews. BERT is a pre-trained language model developed by (Devlin et al., 2018) at Google AI Language. Pre-trained BERT (available at its GitHub

⁴www.italianlp.it/resources/italian-word-embeddings

page⁵) may be fine-tuned on a specific NLP task in a specific domain, such as the sentiment analysis for reviews in the healthcare domain. To do that, the original text must be tokenized with its own tokenizer.

4 Experiments

We conducted two types of experiments. In the first one, we wanted to evaluate which of the systems was the best. For each configuration, we have trained and tested the system using a stratified k -fold cross-validation (with $k = 5$). In the second part, we wanted to evaluate the robustness of the best system in a context out-domain, dividing the folders by disease sites. The software has been entirely developed in Python.

4.1 System 1

We tested three different configurations of our SVM-based system, depending on the sets of features used in the experiment: only hand-crafted features (more than 626 thousands features), only embeddings (387 features), and a combination of both. For such experiments, the features that have shown to not bring improvements to the performance (numbers of tokens and sentences), or even

⁵www.github.com/google-research/bert

to lower it (fg-POS n -grams, Lemmas n -grams with $n=\{2, 3\}$) during a preliminary experimental phase were excluded from the hand-crafted features set. Thus, it turns out that such set is composed only of Uncased Word and cg-POS n -grams with $n=\{1, 2, 3\}$, and Lemmas. In order to reduce the dimensionality of the set, but also to improve the performance of our system, the features pass through a step of filtering after being extracted for the training set. Each feature that appears less than a certain threshold th within the training set can be assumed to be not so relevant and is therefore discarded. Such threshold has been set equal to 1 ($th=1$) after a search of the optimal value during the preliminary experimental phase.

4.2 System 2

The experiments with BERT were conducted using the same partition into the 5 folds used during the experiments with the SVM-based classifier. This division allowed us to compare the results achieved by the two classifiers. The BERT model used in our experiments is the multilingual cased pre-trained one.

We tested two different approaches. These experiments have followed two pipelines. In the first one, the model was fine-tuned with folds from the original dataset described in section 2. In the second one, each fold was obtained by oversampling the minority class (i.e. the negative one) in the original fold. The oversampling was obtained by multiplying each negative sample in the fold by 4. These results in the ratio of negative to positive samples being increased from about 1:16 to about 1:4. Other experiments were conducted further increasing the ratio to about 1:2, but this has not led to significant improvements in performance at the expense of computational time. For both the approaches, the model was fine-tuned for 5 epochs on a 12 GB *NVIDIA* GPU with *Cuda 9.0* with the following hyperparameters:

- maximum sequence length of 128 tokens (it seems reasonable since this number is very close to the average length of the documents in the dataset, as reported in Figure1),
- batch size of 24 samples,
- and a learning rate of $5 * 10^{-5}$.

	F1₍₁₎ (%)	F1₍₋₁₎ (%)	F1 (%)
SVM			
Hand-crafted	98.90 \pm 0.07	82.73 \pm 1.03	90.81 \pm 0.55
aEmbeddings	96.16 \pm 0.15	62.37 \pm 0.74	79.27 \pm 0.44
Both	98.94 \pm 0.04	83.47 \pm 0.72	91.21 \pm 0.48
BERT			
w/o oversampling	/	/	/
w/ oversampling	98.60 \pm 0.04	77.56 \pm 0.81	88.08 \pm 0.42
Baseline	96.80	0.00	48.40

Table 2: Results of the experiments in the stratified 5-fold cross-validation. Performances are reported in terms of F1-score (%) on each class and the (macro) average between the two. The best results are shown in bold.

5 Results and Discussion

Table 2 resumes the results of the experiments in stratified 5-fold cross-validation. The performances are reported in terms of the macro average of *F1-score*.

After analyzing these results, we took the best model and we used it in the leave-one-site-out cross-validation context to test the reliability of the system in an out-domain (site) problem. These results are resumed in Table 3.

First of all, we can notice that such performances are much higher of the baseline system, i.e. the performance achieved by a hypothetical model that classifies all the samples as belonging to the majority class (that is, the positive class).

Due to the strong dataset imbalance and the low batch size, training BERT without oversampling the dataset leads the system to classify all samples as belonging to the majority class, i.e. the positive class. This leads to often obtain very bad performance, i.e. the baseline performance. Anyway, when this problem does not come up, the classifier shows the lowest value of the F1-score. These results clearly show the difficulties of BERT to deal with unbalanced datasets. Oversampling the minority class has shown to partially cope with such problems, leading to an improvement in terms of repeatability and performance.

For what concerns the experiments with the SVM-based system, they have shown that hand-crafted features have greater relevance for the task than the embedding features. This suggests that the (Italian) healthcare reviews domain may be particularly lexical. Thus, sets of lexical features show better performance than those similarity-based features. However, the resulting best model is the one with both sets of features, outperforming the BERT-based system best configuration by about three percentage points.

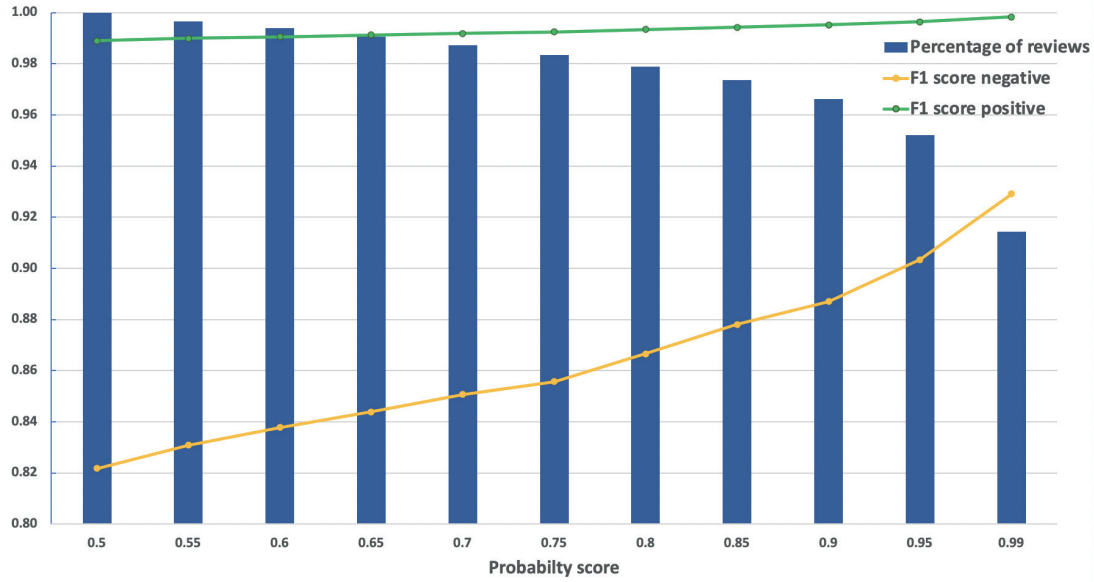


Figure 2: Results in terms of percentage of classified reviews and F1-score over threshold values on the probabilistic score $p \in [0, 1]$ returned by the Platt scaling method applied on top of the SVM-based system. All the results are referred to the k -fold cross-validation (with $k = 5$) fashion. Note that for $threshold = 0.5$, even if the percentage of classified documents is 100%, the value of the macro average of F1-score is lower to the one reported in Table 2. This is due to the inherent inconsistency between the probabilities calculated through the Platt scaling method p and the decision score of the SVM model (i.e. the distance of the sample from the trained boundary, $d \in (-\infty, +\infty)$).

Site	F1 (%)	F1 _{Baseline} (%)
Nervous System	89.91	48.52
Hearth	90.20	49.10
Haematology	91.10	48.36
Endocrinology	87.79	47.40
Endoscopy	94.34	47.49
Facial	88.31	48.90
Genital	92.12	48.10
Gynaecology	93.64	47.98
Infections	91.09	45.95
Ophthalmology	90.74	48.09
Oncology	90.85	48.72
Otorhinology	89.56	48.51
Skin	93.86	46.35
Plastic Surgery	93.63	49.07
Pneumology	92.76	45.63
Rheumatology	90.75	46.90
Senology	91.29	49.06
Thoracic Surgery	92.62	48.01
Vascular Surgery	90.01	48.41
Average	91.24	47.92

Table 3: Results of the experiments in leave-one-site-out cross-validation. The first column shows the site used for testing, while the next two columns are the values of performance and baseline in terms of the (macro) average of F1-score of each test set.

Furthermore, the leave-one-site-out experiments with this model result in a very good performance, showing the system to be reliable also in an out-domain (site) context. This last result can be due to two factors: 1) the high degree of overlap of the lexicon found in one domain on the lexicon of all other domains; 2) a larger size of the set used for training.

In addition to the two main phases of experiments, we further investigated the confidence of the best model developed in making decisions. The motivation behind this study is that it may have application in real-world cases, where an automated system is required to filter the documents on which it is highly confident (i.e., above a certain threshold) and then passes the most complex documents to a human operator. To do so, we applied the Platt scaling (Platt, 1999) method on top of the trained SVM model. This step is needed to convert the output of the model from a decision score $d \in (-\infty, +\infty)$, i.e. the distance of the test sample from the trained boundary, to a probabilistic score $p \in [0, 1]$, representative of the system confidence in making the decision. Figure 2 resumes the results of this analysis. As expected, the number of documents on which the system makes a decision falls as the confidence threshold required of the system increases. However, this trend does not have such a negative slope and still classify more

than 91% of the documents with 99% confidence. At the same time, the performance advantage is clear, leading to an increase of F1-score on negative samples by more than ten percentage points.

6 Conclusion

In this paper, we have introduced a novel system for sentiment analysis for Italian reviews in Healthcare. For the best of our knowledge, this is the first work of this kind in such domain. To do so, we have collected the first dataset for this domain from the web. Then, we have implemented and compared two types of classifiers of the state of the art for such task, the SVM and BERT. Despite the strong dataset imbalance, we have obtained very good results, especially with the SVM-based system, which outperformed the BERT-based one, while maintaining a low computational burden during training. However, there is a chance that increasing the maximum sequence length of BERT it may outperform our best-developed system. Also, recent work (Nozza et al., 2020) has analyzed the contribution of language-specific models, showing in general improvements over BERT multilingual for a wide variety of NLP tasks. For this reason, it might be worth including in future works the use of specific models for Italian, such as GilBERTo⁶, UmBERTo⁷, and AlBERTo⁸. The latter was already used for a sentiment classification task (Polignano et al., 2019). Future works on this dataset may also tackle the task of sentiment classification including the neutral class or sentiment regression of the average scores. Moreover, future research may tackle the task of cataloguing reviews to the area of disease they belong, maybe including other features from metadata such as titles.

Acknowledgements

Our work is possible thanks to a general R&D agreement between the National Research Council of Italy (CNR) and Confindustria (the main association representing manufacturing and service companies in Italy) and a specific R&D agreement between Webmonks s.r.l., CNR and the Campus Bio-Medico University of Rome (UCBM).

⁶www.github.com/idb-ita/GilBERTo

⁷www.github.com/musixmatchresearch/umberto

⁸www.github.com/marcopoli/AlBERTo-it

References

- [Barbieri et al.2016] Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. 2016. *Overview of the Evalita 2016 SENTiment POLarity Classification Task*. Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016
- [Basile et al.2018] Basile, P., Croce, D., Basile, V., & Polignano, M. 2018. *Overview of the EVALITA 2018 Aspect-based Sentiment Analysis Task (ABSITA)*. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.
- [Cignarella et al.2018] Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., & Rosso, P. 2018. *Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)*. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.
- [Cimino et al.2018] Cimino A., De Mattei L. & Dell’Orletta F. 2018. *Multi-task Learning in Deep Neural Networks at EVALITA 2018*. Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA’18), 86-95.
- [Cimino and Dell’Orletta2016] Cimino, A., & Dell’Orletta, F. 2016. *Tandem LSTM-SVM approach for sentiment analysis*. In of the Final Workshop 7 December 2016, Naples (p. 172).
- [Devlin et al.2018] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.
- [Fan et al.2008] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. 2008. *LIBLINEAR: A library for large linear classification*. Journal of machine learning research 9.Aug (2008): 1871-1874.
- [Gao et al.2012] Gao, G. G., McCullough, J. S., Agarwal, R., & Jha, A. K. 2012. *A changing landscape of physician quality reporting: analysis of patients’ online ratings of their physicians over a 5-year period*. Journal of medical Internet research, 14(1), e38.
- [Greaves et al.2012] Greaves, F., & Millett, C. 2012. *Consistently increasing numbers of online ratings of healthcare in England*. J Med Internet Res, 14(3), e94.

- [Mohammad et al.2013] Mohammad S.M., Kiritchenko S., and Zhu X. 2013. *NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets*. In Proceedings of the Seventh international workshop on Semantic Evaluation Exercises, SemEval-2013. 321-327, Atlanta, Georgia, USA
- [Nozza et al.2020] Nozza, D., Bianchi, F., & Hovy, D. 2020. *What the [mask]? making sense of language-specific BERT models*. arXiv preprint arXiv:2003.02912.
- [Platt1999] Platt, J. 1999. *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*. Advances in large margin classifiers, 10(3), 61-74.
- [Polignano et al.2019] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. 2019. *ALBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets*. In CLiC-it.

Investigating Proactivity in Task-Oriented Dialogues

Vevake Balaraman^{1,2}, Bernardo Magnini¹

¹ Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento — Italy

² ICT Doctoral School, University of Trento — Italy.
{balaraman, magnini}@fbk.eu

Abstract

Proactivity (i.e., the capacity to provide useful information even when not explicitly required) is a fundamental characteristic of human dialogues. Although current task-oriented dialogue systems are good at providing information explicitly requested by the user, they are poor in exhibiting proactivity, which is typical in human-human interactions. In this study, we investigate the presence of proactive behaviours in several available dialogue collections, both human-human and human-machine and show how the data acquisition decision affects the proactive behaviour present in the dataset. We adopt a two-step approach to semi-automatically detect proactive situations in the datasets, where proactivity is not annotated, and show that the dialogues collected with approaches that provide more freedom to the agent/user, exhibit high proactivity.

1 Introduction

Proactivity is the collaborative attitude of humans to offer information in a dialogue even when such information was not explicitly requested. As an example, a travel operator may suggest points of interest and attractions in a certain area, even if the customer did not explicitly requested for them. The following portion of dialogue, extracted from the Nespole dataset (Mana et al., 2003), shows proactive contributions of the travel agent (displayed in italics).

Client: good morning; could you suggest any village in the Val di Fiemme to me; where it's possible to skate for example; that is does any skating rink exist in the Val di Fiemme;
Agent: yes; in the whole of Val di Fiemme there are some outdoor skating rinks; *where you can skate usually in the afternoon; in some rinks even in the morning;* and then right in Cavalese there's a skating rink an ice rink; *where even some courses are organized; where they also hold hockey or skating shows; and it's indoors.*

In this dialogue situation the travel agent provides indications both about the opening time of skating rinks and about skating courses, which were not requested by the customer. We may think proactivity as a guess of the agent with respect to the customer needs, with the purpose of anticipating expected requests, this way facilitating the achievements of the dialogue goals.

Proactivity is a crucial characteristics of human-human dialogues. It is related to the so called *principles of cooperative dialogue*, which have been summarized in the popular Grice's maxims (Grice, 1975). In particular, proactivity follows the maxim of quantity, where one tries to be as informative as one possibly can, and gives as much information as it is needed, and no more. Under this maxim, proactivity has to find a trade-off between providing useful not requested information and limiting excessive not needed information. For instance, in the context of our dialogue about skating in Val di Fiemme, an agent suggesting a good pizzeria would probably be perceived as a violation of the quantity maxim, as this information seems not enough needed in that context.

Despite the large use of proactivity that we note in everyday human-human dialogues, proactive behaviours are poorly represented in most of the models at the core of the last generation of task-oriented dialogue systems. Overall, we notice a general lack of cooperative phenomena (e.g., clar-

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ification questions, explanatory dialogues, proactivity, etc.), that characterize, and somehow make efficient, task-oriented human-human dialogues. A notable exception are recommendation systems (Thompson et al., 2004; Sun and Zhang, 2018; Yoshino and Kawahara, 2015), where, however, the focus is on influencing the user towards a specific goal (e.g., buy a certain product). Instead, we intend proactivity to be a general collaborative strategy aiming at improving the quality and effectiveness of the conversation. As an example, proactivity can be used to anticipate future requests of the user (e.g., providing the telephone number of a certain restaurant), or to recover from failure situations (e.g., offering possible alternatives when there are no restaurants satisfying the user desires).

The main purpose of the paper is to conduct an empirical analysis over several existing task-oriented dialogue datasets, used to train dialogue models, in order to verify the presence of proactive behaviours. More specifically, we consider a human-human dialogue corpus collected with a role-taking methodology, i.e., Nespole, and compare it with other task-oriented dialogues collected either with Wizard of Oz or with bootstrapping methods. To conduct such a comparison, the major obstacle is that in both cases, proactivity is not annotated in any way, and we had to figure out methods (addressed in Section 2) to semi-automatically detect proactive situations. Results confirm that dialogues collected through role-taking methodology show a much richer presence of proactivity, which is a challenge for future dialogue systems.

2 Methodology

In this Section, first we define proactivity behaviours in the context of task-oriented dialogues, and then we describe the methodology we use to detect proactivity in available dialogue corpora.

2.1 Defining Proactivity

Our starting point is the work on proactivity presented in (Balaraman and Magnini, 2020), where a pro-active behaviour is defined as any information that: (i) is introduced by the system; (ii) was not previously introduced in the dialogue by the user; and (iii) is assumed to be relevant to achieve the user needs. According to this definition, system turns like the following are all proactive:

- System: *We have good reviews for restaurant X.*
- System: *There are no Eritrean restaurants in the city center, but there are several of them in the south of the city.*
- System: *In case it might be useful, the telephone of the restaurant is X, after a certain restaurant has been chosen by the user.*
- System: *There is a metro station close to the restaurant you have chosen.*

As the examples show, proactive information is strictly related to domain knowledge (e.g., knowledge about restaurants in a city). Moreover, the system may decide to be proactive only in certain dialogue situations, where there is need to help the user to positively conclude a dialogue. In our second example, for instance, the user needs do not match any instance in a domain Knowledge Base (i.e., there are no Eritrean restaurants in the city center), and the system informs the user that there are Eritrean restaurants in the south of the city, this way avoiding a longer follow up interaction.

2.2 Detecting Proactivity

Unfortunately, proactivity was not a designing feature of any of the datasets considered in this study. This means that proactivity is not marked, and that we need to figure out how it can be detected at a reasonable cost. The approach taken in this paper detects proactivity occurring in *intermediate failure situations*, when the system tries to recover from a dialogue failure. There are two reasons for this choice: (i) failure situations are easy to be detected through simple patterns (e.g., *I am sorry...*, *We do not have...*); (ii) as the capacity to recover from failure situations is crucial to maximize the final success of the dialogue, we assume that the attitude of a system to be proactive is particularly revealed in failure situations. In other words, we look at failure situations as typical situations where proactivity should be applied by a system. Given a dialogue collection, we can consider the proportion of proactivity within intermediate failures as a sort of upper bound of proactivity in the whole collection.

Under this assumption, we adopted a failure-based, two-step methodology for detecting proactivity. At the first step we detect as much as possible turns where the system inform the user that

his/her request cannot be satisfied. This step is implemented through either pattern-based search of typical linguistic expressions indicating failure (e.g., *I am sorry...*, *We do not have...*, *There are no...*, etc.) or patterns in dialogue acts of the system. At the second step, we focus on system failure responses, and check whether the response contains any proactive information (see Section 2.1: if any proactive information is present, we mark the system turn as proactive, otherwise as non-proactive. This second step is either performed manually or by finding patterns in the dialogue acts of the system response.

3 Experimental Data

In this section we describe the different data acquisition approaches used for the collection of task-oriented dialogue datasets, and provide details about them.

3.1 Data Acquisition Approaches

We consider three data acquisition approaches that are widely used for dialogue collection.

Wizard of Oz (WoZ) is the most popular approach to collect task-oriented dialogues, possibly using crowd workers (Fraser and Gilbert, 1991; Kelley, 1984). This involves a pair of crowd workers who are provided with respective dialogue goals and are asked to communicate in natural language to achieve the goal. Each crowd worker, acting either as the wizard or the user, is provided with the instructions to achieve the dialogue goal.

The following is an example of a dialogue script provided to the crowd worker in the MultiWoZ (Budzianowski et al., 2018) dataset.

1. You are looking for a *place to stay*. The hotel should be in the *cheap* price range and should be in the type of *hotel*
2. The hotel should *include free parking* and should *include free wifi*
3. Once you find the *hotel* you want to book it for *6 people* and *3 nights* starting from *tuesday*
4. If the booking fails how about *2 nights*
5. Make sure you get the *reference number*

The dialogue script is typically filled in using placeholders in a template (shown in *italics* in our

example). We notice the amount of details present in the dialogue description, which could influence the crowd worker utterance for a given turn, and induce to follow a structure similar to the dialogue script.

Bootstrapping, also referred to as Machines talking to Machines (M2M), is a simulation-based approach for generating *outlines* for a number of dialogues via self-play (Shah et al., 2018), a methodology that takes advantage of a task-specific information input provided by the developer. The task-specification defines the schema of intents, the slot names and the slot values for a certain domain. Based on the task-specification, the framework first generates a set of dialogue outlines containing natural language utterances and their corresponding annotations. The obtained dialogues are then paraphrased using crowd workers in order to obtain linguistic variations. This approach reduces the resources required to collect a large dialogue dataset and enables the developer to control for the diversity both in the dialogue flow and in the user behaviors. Table 1 shows an example of a dialogue outline generated through a bootstrapping approach, which is then paraphrased using crowd workers.

Role-Taking. This methodology involves people playing two roles, typically with minimum training, interacting in order to achieve a given goal (e.g., a travel agent and a customer with the goal of organizing a trip; an applicant and a job operator with the goal of finding a job opportunity). For both the participants responses are unscripted and are supposed to be natural as expected in a real-world conversation. This is similar to the MAP task approach (Anderson et al., 1991; Meena et al., 2013), which allows to collect unscripted dialogues with specific communication goals. Both the participants can be trained workers, acting respectively as the user and the expert, and are provided with a dialogue goal and information they can use (e.g. an applicant with a CV and a job operator with job offers). Table 2 shows a sample interaction for this approach.

3.2 Datasets

We have analysed proactivity in five available collections of task-oriented dialogue datasets in English, all of them used to train dialogue models. In addition, we have compared them with Nespole (Mana et al., 2003), a human-human dia-

Dialogue Outline		Paraphrase
Annotation	Template utterances	NL utterances
S: greeting()	Greeting.	Hi, how can I help you?
U: inform(intent=book_movie, name=Inside Out, date=tomorrow, num_tickets=2)	Book movie with name is Inside Out and date is tomorrow and num tickets is 2.	I want to buy 2 tickets for Inside Out for tomorrow.
S: ack() request(time)	OK. Provide time.	Alright. What time would you like to see the movie?
U: inform(time=evening)	Time is evening.	Anytime during the evening works for me.
S: offer(theatre=Cinemark 16, time=6pm)	Offer theatre is Cinemark 16 and time is 6pm.	How about the 6pm show at Cinemark 16?
U: affirm()	Agree.	That sounds good.
S: notify_success()	Reservation confirmed.	Your tickets have been booked!

Table 1: A sample dialogue collected through the *bootstrapping* approach (Shah et al., 2018).

logue dataset which was collected to study real-world human-human interactions. Table 3 reports the main characteristics of the six datasets, including the method of data acquisition.

WOZ2.0 includes textual conversations for restaurant booking in Cambridge and was collected using Wizard of Oz by pairing users in Amazon Mechanical Turk. The user and the wizard contribute a single turn to each dialogue (Wen et al., 2017). (Mrkšić et al., 2017) expanded the original WoZ dataset producing the WoZ2.0 dataset, consisting of 1200 dialogues.

MultiWOZ2.1 includes dialogues in multiple domains collected via Wizard of Oz. The developers explicitly encouraged goal changes, in order to model realistic conversations (Budzianowski et al., 2018). Different versions of the dataset have been published recently, addressing annotation errors occurring in the original dataset (Ramadan et al., 2018; Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020). We use the MultiWoZ2.1 dataset, containing 10438 dialogues.

Schema-Guided Dataset (SGD) consists of 22825 dialogues in multiple domains collected using the Machine Talking to Machine (Bootstrapping) approach (Rastogi et al., 2019). Dialogues generated via simulation are then paraphrased by the crowd workers for language variability. SGD promotes research towards dialogue systems that can handle dynamic schemas.

Microsoft Dialogue dataset (Li et al., 2016; Li et al., 2018) consists of dialogues collected via

Amazon Mechanical Turk using a bootstrapping approach for three different domains *Movie-Ticket Booking*, *Restaurant Reservation* and *Taxi Ordering* with 2890, 4103 and 3094 dialogues, respectively.

Maluuba Frames dataset (El Asri et al., 2017) consists of 1369 dialogues collected via Wizard of Oz using a Slack bot for travel vacation domain. Users were assigned a task using a template where placeholder values are filled by drawing values from a database. If the task is successful, the user either ended the dialogue or received an alternate task. In case of no match, suggestions were sometimes provided to the wizards, who then decided whether to use or not the suggestion for the user.

Nespole (Mana et al., 2003; ?) is a VoIP (Voice over Internet Protocol) corpus consisting of spoken interactions between a professional agent and a recruited worker acting as a user or client. We use the DB-1 part of the Nespole dataset, consisting of 39 dialogues (in the transcribed version of the dataset 3 client side dialogues were missing, leaving 36 dialogues for a total of 1549 turns). Dialogues are about vacation planning in the Trentino region and, unlike other datasets, they do not have a fixed user-side goal, but rather a collaborative goal. Specifically, the user and the agent collaborate via a spoken conversation to achieve a goal that satisfies the user.

Speaker	Utterance
System	Could you help me to find my way to the bus stop?
User	start from the department store
System	yeah
User	and eh
System	Should I start by going west?
User	yeah do that
User	then you will get to a meadow and when you get to the meadow
System	Eh, could you repeat that?
User	you go straight and you see a meadow on your right side
System	A green field?
User	ehm yeah a field
System	mhm
User	pass the meadow and turn right so you are going north
System	okay
...	...
User	at the junction go south and then you will get to the bus stop
System	okay, thanks a lot.

Table 2: A sample dialogue collected through the Role-Taking approach (Meena et al., 2013).

4 Results and Discussion

We have applied the methodology described in Section 2 to detect proactivity in the six datasets. First we detect the number of failure turns in each dataset and then, among failures, we identify the turns that exhibit proactivity.

Table 4 reports the number of failure turns we were able to detect for each dataset, and the proportion of them that exhibit a proactive behaviour, according to our definition in Section 2.1. We can notice that the datasets collected via Wizard of Oz (WoZ) typically exhibit very low proactivity. This could be due to the fact that in the WoZ approach users are provided with a task description detailing how to proceed with the dialogue. This indirectly influences the users to use certain formats as defined in the description. The MultiWoZ2.1 dataset shows the highest proactivity among the datasets collected via WoZ approach: this is due to the explicit encouragement of goal changes in task-descriptions. As for the SGD and Microsoft dialogue datasets, collected via a *bootstrapping* approach, we can notice that over 50% of the fail-

ure turns exhibiting proactivity. This is because of the choice of the developers to specifically include such failure and recovery scenarios in the dialogue flow.

Datasets collected via WoZ and bootstrapping have different approaches in adopting proactivity. Since WoZ is collected by pairing humans, proactive turns often contain information that would lead to a dialogue success. However, in the bootstrapping approach, as it is based on a script, the proactive turns contain information that are possible for the user to request but may not lead to dialogue success. An example in MultiWoZ2.1 is the following: *"There are no hotels that fit your criteria in the South, but there are two Guesthouses. Would you like to book one of those?"*. Here the crowd-worker acting as a wizard has already looked the availability of two Guesthouses and is providing this information to another crowd-worker who is acting as the user. If the user chooses the guesthouse, the dialogue would be a success. A similar example in Microsoft Dialogue dataset is the following: *"I'm sorry The Other Side of The Door is not playing in your area on Tuesday. I am able to find show times for The Witch and Triple 9"*. Here, the system-agent is providing information that the user-agent can choose as alternatives, but the alternatives may not always directly lead to dialogue success. When the user-agent responds *"The Witch will be fine"*, the system-agent searches the knowledge-base and responds *"I'm sorry they are only showing The Witch at 4:40 pm. Would that be acceptable for you?"* which again is a proactive response.

The analysis for proactivity in Nespole differs from the other datasets, as Nespole is not modeled to find an exact match for the user needs, and, as a consequence, there are no clear failure situations. In addition, while the other datasets were collected focused towards using them for training dialogue systems, Nespole was collected to analyze linguistic features in real-world dialogues. However, we manually analysed the 36 dialogues (1549 turns) of vacation planning and identified the turns where the agent exhibits proactivity. We found that 49 turns in 26 dialogues are proactive responses, where the agent provides information not explicitly requested by the user (see the example in the Introduction). Since Nespole is a VoIP dataset, the number of turns are not comparable to the other datasets as they contain frequent in-

Dataset	Data Acquisition	#Dialogues	#Turns	Avg. Turn length
WoZ2.0	Wizard of Oz	1,200	8,824	11.27
MultiWoZ2.1	Wizard of Oz	10,438	143,048	13.18
Maluuba Frames	Wizard of Oz	1,369	19,986	12.60
Schema-Guided Dataset	Bootstrapping	22,825	463,284	9.86
Microsoft Dialogue	Bootstrapping	2,890	21,656	10.96
- Movie-Ticket Booking		4,103	29,719	11.45
- Restaurant Reservation		3,094	23,311	11.04
- Taxi Ordering				
Nespole	Role-Taking	36	1,549	18.48

Table 3: Statistics about the datasets used for the proactivity analysis.

Dataset	#Failure	#Proactive	%
WoZ2.0	414	26	5.9
MultiWoZ2.1	2,127	325	15.3
Maluuba Frames	1,214	77	6.3
Schema-Guided	3,362	1,737	51.7
Microsoft			
- Movie	318	161	50.6
- Restaurant	775	323	41.7
- Taxi	104	38	36.5
Nespole	–	49	–

Table 4: Number of failure situations (turns) and corresponding proactivity, for each dataset.

terruptions and fillers. An example of proactive turn in Nespole is the following: *"no; there's no entertainment for the kids; entertainment for the kids would be at the Olimpionic Hotel; but it's a 3 star one already"*. We can see that the agent provides information for a scenario that was requested by the user with a piece of proactive information (*entertainment for the kids would be at the Olimpionic Hotel; but it's a 3 star one*). We notice that proactive turns in Nespole exhibit much richer information compared to the other datasets, which could be attributed to the freedom of expression provided to the agent, unlike to the other approaches considered.

We now discuss a few research questions that arise from our study on proactivity in dialogue collections.

Does our failure-based methodology provide reasonable coverage about proactivity in our datasets? We assume that task-oriented dialogue systems should maximize their success rate (i.e., matching the user needs), and that recovering from intermediate failure situations potentially

increases their success rate. Under this assumption failure situations act as an upper bound for the situations in which the systems is expected to be proactive. As an example, having found that 5.9% of intermediate failures in WoZ2.0 are proactive, we infer that the amount of proactivity in the whole WoZ2.0 will not be higher than 5.9%.

Does proactivity correlate with the method of collection of the dataset? As seen in Table 4, the Wizard of Oz approach consistently has very low proactivity, while the *bootstrapping* approach exhibits high proactivity. While the amount of proactivity in each dataset depends on the developer choice about the dialogue goals and on the instructions provided to users, we can conclude that the WoZ approach indirectly influences the user to deviate from a collaborative approach and to follow a scripted dialogue.

5 Conclusion

Task-oriented dialogue systems have shown to be effective in providing services to users with a high success rate. However, the interaction still lacks an effective proactive approach, which is typical in human-human conversations. In this study, we compare proactive behaviours in several available dialogue datasets, and show that the dialogues collected through Wizard of Oz contain a small proportion of system proactive responses, while dialogues collected through simulation-based and role-taking methodologies contain higher degree of proactivity. To sum up, we suggest that data collection strategies should be better aware that their designing principles have strong influence on the quality of the dialogues. Particularly, we recommend higher attention to proactive behaviours, and, in general, to collaborative phenomena.

References

- Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The hrc map task corpus. *Language and Speech*, 34(4):351–366.
- Vevake Balaraman and Bernardo Magnini. 2020. Pro-active systems and influenceable users: Simulating pro-activity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey, July. SEMDIAL.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August. Association for Computational Linguistics.
- Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.
- Norman M. Fraser and G.Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech Language*, 5(1):81 – 99.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, January.
- Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.
- Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. *arXiv preprint arXiv:1807.11125*.
- N. Mana, S. Burger, Roldano Cattoni, L. Besacier, V. MacLaren, J. McDonough, and F. Metze. 2003. The nespole! voip multilingual corpora in tourism and medical domains. In *INTERSPEECH*.
- Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2013. The map task dialogue system: A test-bed for modelling human-like dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 366–368, Metz, France, August. Association for Computational Linguistics.
- Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada, July. Association for Computational Linguistics.
- Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855*.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June. Association for Computational Linguistics.
- Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, SIGIR ’18*, page 235–244, New York, NY, USA. Association for Computing Machinery.
- Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A personalized system for conversational recommendations. *J. Artif. Int. Res.*, 21(1):393–428, March.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449. Association for Computational Linguistics.

- Koichiro Yoshino and Tatsuya Kawahara, 2015. *News Navigation System Based on Proactive Dialogue Strategy*, pages 15–25. Springer International Publishing, Cham.
- Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

A Diachronic Italian Corpus based on “L’Unità”

Pierpaolo Basile

Dept. of Computer Science

University of Bari, Italy

pierpaolo.basile@uniba.it

Annalina Caputo

ADAPT Centre

School of Computing, Dublin City University

annalina.caputo@dcu.ie

Tommaso Caselli

CLCG

University of Groningen, Netherlands

t.caselli@rug.nl

Pierluigi Cassotti

Dept. of Computer Science

University of Bari, Italy

pierluigi.cassotti@uniba.it

Rossella Varvara

DILEF

University of Florence, Italy

rossella.varvara@unifi.it

Abstract

English. In this paper, we describe the creation of a diachronic corpus for Italian by exploiting the digital archive of the newspaper “L’Unità”. We automatically clean and annotate the corpus with PoS tags, lemmas, named entities and syntactic dependencies. Moreover, we compute frequency-based time series for tokens, lemmas and entities. We show some interesting corpus statistics taking into account the temporal dimension and describe some examples of usage of time series.

1 Motivation and Background

Diachronic linguistics is one of the two major temporal dimensions of language study proposed by de Saussure in his *Cours de linguistique générale* and has a long tradition in Linguistics. Recently, the increasing availability of diachronic corpora as well as the development of new NLP techniques for representing word meanings has boosted the application of computational models to investigate historical language data (Hamilton et al., 2016; Tahmasebi et al., 2018; Tang, 2018). This culminated in SemEval-2020 Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020), the first attempt to systematically evaluate automatic methods for language change detection.

Italian is a Romance language which has undergone lots of changes in its history. Its official

adoption as a national language occurred only after the Unification of Italy (1861), having previously been a literary language. Diachronic corpora of Italian are currently available and accessible to the public (e.g., DiaCORIS and MIDIA). Unfortunately, restricted access/distribution of these resources limits their utilisation. This actually prevents the investigation of more recent NLP methods to the diachronic dimensions.

To obviate this limit, we collect and make freely available¹ a new corpus based on the newspaper “L’Unità”. Founded by Antonio Gramsci on February, 12th 1924, “L’Unità” was the official newspaper of the Italian Communist Party (PCI², henceforth). The newspaper had a troubled history: with the dissolution of PCI in 1991, the newspaper continued to live as the official newspaper of the new Democratic Party of the Left (PDS/DS) until July, 31th 2014. After that date, it ceased its publication until June, 30th 2015, and it was definitely closed on June, 3rd 2017.

Since 2017, the historical archive of “L’Unità” has been made again visible and available on the Web.³ One of the main issues of this resource is the lack of information about who owns the rights of the original archive. To our knowledge, the online version of the archive was legally obtained by downloading the original archive before the closure of the newspaper. The current archive, available online, does not contain the local editions of the newspaper and the photographic archive.

The main contribution of this work lies in the

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://github.com/swapUniba/unita/>

²It is the acronym of *Partito Comunista Italiano*.

³<https://archivio.unita.news/>

resource itself and its accessibility to the research community at large. The corpus is distributed in two formats: raw text and pre-processed. The validity of the corpus for the automatic study of language change is currently tested as part of the DIACR-Ita task ⁴ at EVALITA 2020. However, we illustrate some further potential applications of the use of the corpus.

2 Italian diachronic corpora

Various Italian diachronic corpora are currently available and accessible to the public. DiaCORIS ⁵ (Onelli et al., 2006) comprises written Italian texts produced between 1861 and 1945, for a total of 100 million words, while MIDIA ⁶ (Gaeta et al., 2013) covers written documents in Italian between the beginning of the XIII century and the first half of the XX century, for a total of 7,5 million words over 800 texts belonging to different genres. The Corpus OVI dell’Italiano antico⁷ consists of 1948 texts from the XII to the XIV centuries, for a total of 536.000 words. The LIZ⁸ database comprehends 1,000 literary texts from the XIII to the XX century. Lastly, the *Corpus of Alcide de Gasperi’s* public documents (Tonelli et al., 2019) includes 1,762 documents (newspaper articles, propaganda documents, official letters, parliamentary speeches, for a total of 3.000.000 tokens) written from the Italian politician Alcide De Gasperi and published between 1901 and 1954.

These existing resources differ from each other and from the present corpus in different ways. First, the span of time the texts come from. The OVI Corpus considers texts from the early stages of the Italian language, with a time span of three centuries. The MIDIA corpus and the LIZ database cover 7 centuries, from the XIII to the first half of the XX century. DiaCORIS, the De Gasperi’s corpus and L’Unità corpus contain texts from a shorter and more recent period of time. However, the time span considered in L’Unità corpus is interesting for the study of the Italian language because of the deep changes that occurred

in that period. Indeed, the second half of the XX century has seen a wider spread and use of Italian among all the social classes.

Second, these corpora differ for the genres represented. The DiaCORIS and MIDIA corpora have been designed as representative and balanced samples of written Italian (considering, among other genres, academic prose, fiction, press, legal texts, etc). The OVI corpus and the LIZ database comprehend only literary texts. The De Gasperi’s corpus is representative of political text from a single author. L’Unità corpus is representative only of press language, but this restriction may be an advantage in the study of diachronic lexical change. Indeed, observed semantic changes cannot be attributed to attestation from different genres in different periods, but can be interpreted as true semantic shifts.

Lastly, even if most of the corpora can be queried online (with the exception of the LIZ database), only the De Gasperi’s corpus can be freely downloaded. This restriction affects the usability of these resources for the NLP community. With L’Unità corpus we aim at releasing a new diachronic resource that is freely available and that can be used in the theoretical and computational study of language change.

3 Corpus Creation

The corpus creation consists of several steps:

Downloading All PDF files are downloaded from the source site and stored into a folder structure that mimics the publication year of each article.

Text extraction The text is extracted from the PDF files by using the Apache Tika library.⁹ First, the library tries to extract the embedded text if present in the PDF; otherwise the internal OCR is exploited. It is important to notice that during this step several OCR errors occur. In particular, during the processing of the early years, the newspaper has an unconventional format where a few large pages contain many articles split into several columns. Due to this format, the OCR is not able to correctly identify the column boundaries.

Cleaning In this step, we try to fix some text extraction issues. The previous step leaves an empty

⁴<https://diacr-ita.github.io/DIACR-Ita/>

⁵<http://corpora.dslo.unibo.it/DiaCORIS/>

⁶www.corpusmidia.unito.it

⁷<http://gattoweb.oivi.cnr.it>

⁸<https://www.zanichelli.it/ricerca/prodotti/liz-4-0-letteratura-italiana-zanichelli>

⁹<https://tika.apache.org/>

1	Ehud	Ehud	PROPN	SP	nsubj	3	B-PER	False	False	False	Xxxx
2	Barak	Barak	PROPN	SP	flat:name	1	I-PER	False	False	False	XXXXX
3	scende	scendere	VERB	V	ROOT	0	O	False	False	False	xxxx
4	direttamente	direttamente	ADV	B	advmod	3	O	False	False	False	xxxx
5	in	in	ADP	E	case	6	O	False	False	True	xx
6	campo	campire	NOUN	S	obl	3	O	False	False	False	xxxx
7	per	per	ADP	E	mark	8	O	False	False	True	xxx
8	ufficializzare	ufficializzare	VERB	V	advcl	3	O	False	False	False	xxxx
9	la	la	DET	RD	det	10	O	False	False	True	xx
10	candidatura	candidatura	NOUN	S	obj	8	O	False	False	False	xxxx
11	dell'	dell'	DET	DD	det	13	O	False	False	False	xxxx'
12	ex	ex	ADJ	A	amod	13	O	False	False	True	xx
13	premier	premier	NOUN	S	obj	8	O	False	False	False	xxxx
14	laburista	laburista	PROPN	SP	amod	13	O	False	False	False	xxxx

Table 1: An example of generated token features for the sentence: “*Ehud Barak scende direttamente in campo per ufficializzare la candidatura dell’ex premier laburista.*” [Ehud Barak takes the field to announce the candidacy of the former labour leader.]

line when the end of a paragraph is reached. However, a paragraph can be composed of multiple lines which sometimes contain a word break at the end of the line. We manage word breaks in order to obtain a paragraph on a single text line; we still retain the empty line for delimiting paragraphs. Moreover, we remove noisy text by adopting two heuristics: (1) paragraphs must contain at least five tokens composed by only letter characters; (2) 60% of the paragraph must contain words that belong to a dictionary. The dictionary is built by extracting words that occur into the Paisà corpus (Lyding et al., 2014) taking into account only words composed by letters. The output of this process is a plain text file for each year where each paragraph is separated by an empty line.

Processing All plain text files produced by the cleaning step are processed by a Python script that splits each paragraph into sentences and analyses each sentence by performing several natural language processing tasks. We rely on the spaCy¹⁰ Python library for performing: tokenization, PoS-tagging, lemmatization, named entity recognition and dependency parsing. The spaCy library provides performance comparable to the state-of-the-art approaches with a good processing speed when compared to other NLP tools.¹¹ We also provide the plain text in order to allow the processing with other tools. Each plain text file is analysed and transformed in vertical format adding two tags: `<p>...</p>` for the begin and the end of a paragraph, and `<s>...</s>` for delimiting sentences. The vertical format is compliant to the CONLL representation standard and the tag-set for the Italian¹² is automatically mapped to the

Universal Dependencies scheme¹³.

Feature	Description
Position	The token position in the sentence starting from 1
Token	The token
Lemma	The lemma
PoS-tag	The PoS tag
Tag	Additional tags, such as morphological tags
Dependency	Dependency type
Head position	Head position of the dependency
IOB2 NE	IOB2 tag of the named entity
Punctuation	Boolean indicating if punctuation
Space	Boolean indicating if space character
Stop word	Boolean indicating if stop word
Shape	The word shape – capitalisation, punctuation, digits

Table 2: Description of token features.

The corpus spans 67 years from 1948 to 2014. For each year, we provide two files: (1) the plain text file containing the cleaned text extracted from PDF where each paragraph is delimited by an empty line; (2) a vertical file. In the vertical file format, exemplified in Table 1, each paragraph is split in sentences and tokens occurring in each sentence are annotated with 12 features, whose symbols and descriptions are reported in Table 2.

4 Corpus Statistics

In this section, we report some corpus statistics. Table 3 illustrates the total number of occurrences and the dictionary size for each feature (token, lemma, and named entity, respectively).

	dict. size	occurrences
token	4,177,128	425,833,098
lemma	4,053,561	425,833,098
named entity	5,429,470	26,330,273

Table 3: Dictionary size and total number of occurrences.

¹⁰<https://spacy.io/>

¹¹<https://spacy.io/usage/facts-figures>

¹²<https://spacy.io/api/annotation>

¹³<http://universaldependencies.org/u/pos/>

The corpus contains more than 400 million occurrences and more than 25 million named entities occurrences. The most frequent entities are *Italia*, *Roma* and *PCI*. This result is expected since “L’Unità” was the newspaper of the Italian Communist Party.

Figure 1 shows the PoS-tags¹⁴ frequency over time for open-class tags: NOUN, VERB, ADJective, ADVerb and PROPer Noun. The most frequent tag is NOUN followed by VERB, PROPn, ADJ and ADV. We observe that the frequency of PoS-tags is almost constant over time (excluding PROPn) underlying a stable language style that is typical for the news domain. We observe a variable usage of proper nouns, that may be related to the different types of events narrated over time that do not depend on a particular language style. Moreover, after the 1976, we observe a complementary trend between the adjectives and adverbs frequencies: the former slightly increase over time, while the latter decrease. This may denote a change in the language style that has varied to prefer the usage of adjectives over adverbs in more contemporary writing.

An interesting analysis concerns the tokens occurrences per year, whose result is plotted in Figure 2. We observe a low number of occurrences in the period (1948-1970), probably due to two factors: (1) the first period contains many OCR errors and noise removed during the cleaning step; (2) the number of pages of the newspaper increases over time. The latter may also explain the lower number of tokens for some of the years, such as 1981, 1995, 2000, 2007-2008, 2014. In particular, the latest years are characterised by management issues (e.g. the newspaper liquidation in July 2000) that were reflected in the newspaper format.

We also compute the time series of normalised occurrences (frequency) for each token, lemma, and named entity. All the aforementioned statistics are distributed in separate files together with the corpus.

As an illustrative example of the potential use of the corpus, in Figure 3 we plot the time series for two keywords. The first, *comunismo* [communism], is assumed to be pivotal to this corpus due to the specific role played by the newspaper in relation to the PCI. The second keyword, *antipolitica* [anti-politics], is particularly interesting as it is

a term used to describe the current state of the political life in Italy, characterised by a high level of distrusts in parties and, more generally, in politics. The lifespan of *comunismo* [communism] appears to be extremely influenced and characterised by history. We observe two big spikes in the time series. The first is around 1962, one of the harshest year of the Cold War, witnessing the Cuban missile crisis. The second spike is between 1989 and 1991, corresponding to the beginning of the worldwide crisis of the communist movement and the dissolution of PCI. After 1991, the frequency of the term constantly decreases. Interestingly, the frequency for *comunismo* [communism] is low between 1968 and 1988, a period of time that witnessed a cultural hegemony of leftist movements and strong criticism against the U.S.S.R. On the other hand, we observe that *antipolitica* [anti-politics] is a recent term whose first appearance dates back to 1977. The word frequency starts to increase slowly from 1999 and it reaches its peak in 2012 with the unexpected electoral success of the populist 5 Star Movement at the local elections in May.

Using the same approach, we plot the time series for two named entities: *PCI* and *Berlusconi*. We notice that the frequency of *PCI* start dropping in 1986, few years before its dissolution in 1991, while the name *Berlusconi* has a substantial increase in 1994 when he became the Italian Prime Minister.

Finally, we investigate how the vocabulary changes between two periods: $T_1 = [1948 - 1958]$ and $T_2 = [2004 - 2014]$. For each period we build the vocabulary V_i taking into account only words that occur at least 10 times. Then, we compute the differences between the two dictionaries, $V_1 \setminus V_2$ and $V_2 \setminus V_1$, and sort the words in descending order by occurrences. We observe that the words *agrari*, *imperialisti*, *mezzadri*, *monarchici*¹⁵ appear frequently in T_1 and never appear in T_2 , conversely the words *euro*, *centrosinistra*, *centrodestra*, *immigrati*¹⁶ appear only in T_2 . A similar analysis was executed on named entities¹⁷ and shows that *Scelba*, *D.C.*, *PSI*, *U.R.S.S.* are specific to T_1 , while *Berlusconi*, *PD*, *Bush*, *Obama* to T_2 , revealing differences in topics and people covered

¹⁵In English: *agrarians*, *imperialists*, *sharecroppers*, *monarchists*.

¹⁶In English: *euro*, *centre-left politics*, *centre-right politics*, *immigrants*.

¹⁷In this case we consider only entities that appear at least 5 times.

¹⁴The used tag-set is described here <https://universaldependencies.org/u/pos/>

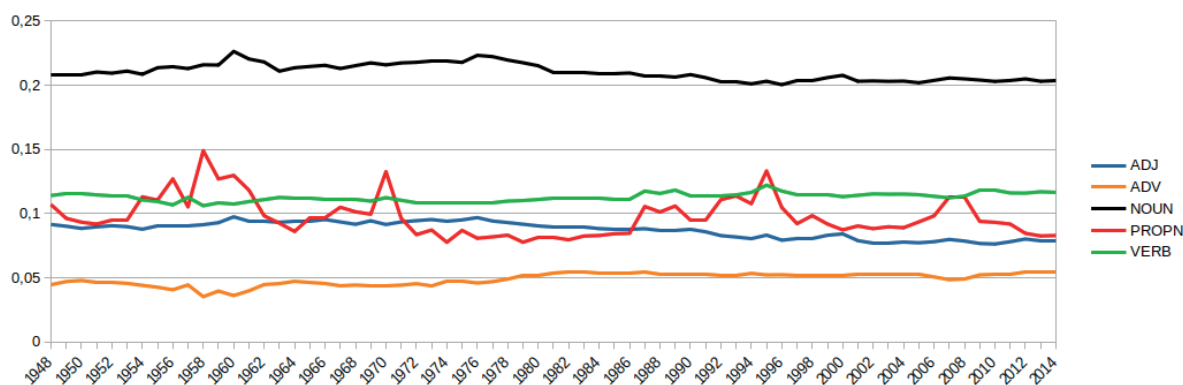


Figure 1: PoS tags frequency over time for: NOUN, VERB, ADjective, ADVerb

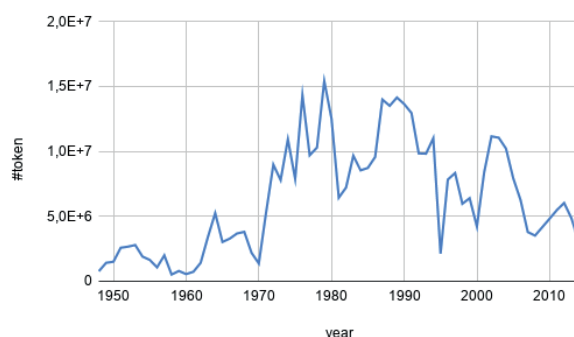


Figure 2: The plot of token occurrences per year.

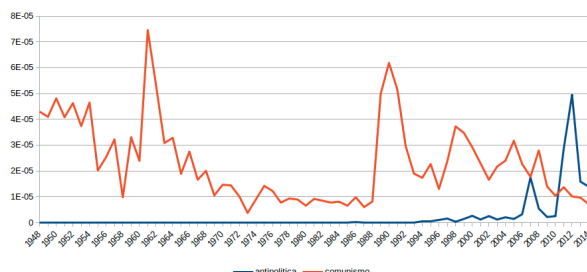


Figure 3: Plot of the time series for the words *comunismo* [communism] and *antipolitica* [anti-politics].

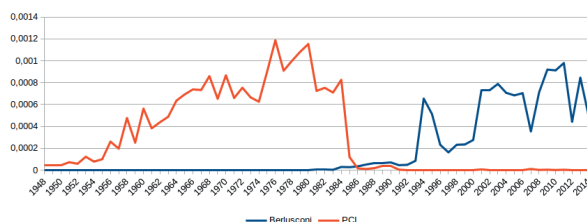


Figure 4: Plot of the time series for the entities *PCI* and *Berlusconi*.

by the newspaper.

5 Conclusions

In this paper, we describe an Italian diachronic corpus based on the newspaper “L’Unità”. The corpus spans 67 years (1948-2014) and is provided

both in plain text and in an annotated format that includes PoS-tags, lemmas, named entities, and syntactic dependencies. We compute some statistics and time series for each token, lemma and named entity. We think that the corpus and the pre-computed data are a valuable source of information both for linguists and researchers interested in diachronic analysis of the Italian language, and for historians, political scientists, and journalists as a digital resource enriched with automatic text analysis technologies.

However, the corpus has some issues that we plan to fix in the future, such as OCR errors and logical document structure recognition. We also plan to process the corpus by exploiting other Italian NLP pipelines in order to understand the differences between the output of different tools. Finally, we are working on generating and making available temporal word embeddings for each year.

References

- Livio Gaeta, Iacobini Claudio, Ricca Davide, Angster Marco, De Rosa Aurelio, and Schirato Giovanna. 2013. Midia: a balanced diachronic corpus of italian. In *21st International Conference on Historical Linguistics*.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1489–1501, may.
- Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell’Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa’corpus of italian web texts. In *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, pages 36–43. EACL (European chapter of the Association for Computational Linguistics).

- Corinna Onelli, Domenico Proietti, Corrado Seidenari, and Fabio Tamburini. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Lexical Semantic Change. *1st International Workshop on Computational Approaches to Historical Language Change 2019*.
- Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676, sep.
- Sara Tonelli, Rachele Sprugnoli, Giovanni Moretti, and Fondazione Bruno Kessler. 2019. Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain. In *CLiC-it*.

Domain Adaptation for Text Classification with Weird Embeddings

Valerio Basile

University of Turin

valerio.basile@unito.it

Abstract

Pre-trained word embeddings are often used to initialize deep learning models for text classification, as a way to inject pre-computed lexical knowledge and boost the learning process. However, such embeddings are usually trained on generic corpora, while text classification tasks are often domain-specific. We propose a fully automated method to adapt pre-trained word embeddings to any given classification task, that needs no additional resource other than the original training set. The method is based on the concept of word *weirdness*, extended to score the words in the training set according to how characteristic they are with respect to the labels of a text classification dataset. The *polarized weirdness* scores are then used to update the word embeddings to reflect task-specific semantic shifts. Our experiments show that this method is beneficial to the performance of several text classification tasks in different languages.

1 Introduction

In recent years, the Natural Language Processing community has directed a great deal of effort towards text classification, in different declinations. The list of shared tasks proposed at the recent editions (2016–2019) of the International Workshop on Semantic Evaluation (SemEval) shows an increasing number of tasks that can be cast as text classification problems: *given a text and a set of labels, choose the correct label to associate with the text*. If the cardinality of the set of labels is two, we speak of *binary* classification, as opposed

to *multiclass* classification. Furthermore, not all binary classification tasks are the same. When the labels indicate the presence or absence of a given phenomenon, we speak of a *detection* task.

Classification tasks are mainly approached in a supervised fashion, where a labeled dataset is employed to train a classifier to map certain features of the input text to the probability of a certain label. Arguably, the most useful features in a NLP problem are the words that compose the text. However, in order to be processed by a machine learning algorithm, words need to be represented in a dense and machine readable format. *Word embeddings* solve this issue by providing vectorial representations of words where vectors that are close in the geometric space represent words that occur often in the same contexts. Among their applications, pre-trained word embeddings are a powerful source of knowledge to boost the performance of supervised models that aim at learning from textual instances.

Several deep learning models compute word embeddings at training time. However, they can be initialized with *pre-trained* word embeddings, typically computed on the basis of concordances in large corpora. This kind of initialization not only boosts the training of the model, but it also represents a way of injecting precomputed world knowledge into a model otherwise trained on a (sometimes very specific) data set.

An issue with word embedding models, including recent contextual embeddings such as Peters et al. (2018), is that they are typically trained on general-purpose corpora. Therefore, they may fail to capture semantic shifts that occur in specific domains. For instance, in a dataset of online hate speech, negatively charged words such as insults often co-occur with words that would normally be considered neutral, but carry instead a negative signal in that particular context. More concretely, in a dataset of hate speech towards immigrant in

the post-Trump U.S., a word that otherwise would be considered neutral such as *wall* carries a definite negative connotation.

In this work, we try to capture this intuition computationally, and model this phenomenon in a word embedding space. We employ an automatic measure to score words in a labeled corpus according to their association with a given label (Section 3.1) and use this score in a fully automated method to adapt generic pre-trained word embeddings (Section 3.2). We test our method on existing benchmarks of hate speech detection (Section 4.1) and gender prediction (Section 4.2), reporting improvements in precision and recall.

2 Related Work

Kameswara Sarma et al. (2018) propose a method to adapt *generic* word embeddings by computing *domain specific* word embeddings on a corpus of text from the target domain and aligning the two vector spaces, obtaining a performance boost on sentiment classification. Another recent approach is based on projecting the vector representations from two domain-specific spaces into a joint word embedding model (Barnes et al., 2018b), building on a similar method applied to cross-lingual word embedding projection (Barnes et al., 2018a). With respect to these works, the approach proposed in this paper is significantly more lightweight, acting directly on a generic word embedding model without the need to train a domain specific one.

The word-level measure introduced in the next section is reminiscent of similar metrics from Information Theory, e.g., Information Content (Pedersen, 2010), and measures of frequency distribution similarity such as Kullback-Leibler divergence (Kullback and Leibler, 1951). However, in this paper we aimed at keeping the complexity of such computation low, in order to manually explore its effect on the word embeddings.

In the domain of hate speech, several approaches mix word embeddings and supervised learning with domain-specific lexicons (e.g., dictionaries of hateful terms), as highlighted by the description of participant systems to recent evaluation campaigns (Fersini et al., 2018; Bosco et al., 2018). These methods are computationally inexpensive, but require curated resources that are not always available for less represented languages.

3 Weirdness-based Embedding Adaptation

In this section, we present our method for automatic domain adaptation of pre-trained word embeddings. The input of the procedure is a set of pre-trained word embeddings and a corpus of texts paired with labels.

3.1 Polarized Weirdness

The Weirdness index was introduced by Ahmad et al. (1999) as an automatic metric to retrieve words characteristic of a *special language* with respect to their typical usage. According to this metric, a word is highly *weird* in a specific collection of documents if it occurs significantly more often in that context than in a general corpus. In practice, given a *specialist* text corpus and a *general* text corpus, the weirdness index of a word is the ratio of its relative frequencies in the respective corpora. Calling w_s the frequency of the word w in the specialist language corpus, w_g the frequency of the word w in the general language corpus, and t_s and t_g the total count of words the specialist and general language corpora respectively, the weirdness index of w is computed as:

$$Weirdness(w) = \frac{w_s/t_s}{w_g/t_g}$$

The weirdness index is used to retrieve words that are highly typical of a particular domain. For instance, in Ahmad et al. (1999), the words *dollar*, *government* and *market* are extracted from the TREC-8 corpus, a collection of governmental and financial domain, by comparing their frequencies to the general domain British National Corpus.

In this work, we propose a new application of the weirdness index to the task of text classification. Rather than comparing the frequencies of words from corpora of different domains, we compute the weirdness index based on the frequency of words occurring in labeled datasets. The mechanism is straightforward: instead of comparing the relative frequencies of a word in a special language corpus against a general language corpus, we compare the relative frequencies of a word as it occurs in the subset of a labeled dataset identified by one value of the label against its complement. Consider a labeled corpus $C = \{(e_1, l_1), (e_2, l_2), \dots\}$ where $e_i = \{w_1, w_2, \dots\}$ is an instance of text (e.g., an online comment), and

l_i is the label associated with e_i , belonging to a fixed set L (e.g., $\{positive, negative\}$).

The *polarized weirdness* (Florio et al., 2020) of w with respect to a specific label $l^* \in L$ is the ratio of the relative frequency of w in the subset $\{e_i \in C : l_i = l^*\}$ over the relative frequency of w in the subset $\{e_i \in C : l_i \neq l^*\}$

Here is an example of how polarized weirdness is computed. Consider a corpus of 100 instances, 50 of which labeled *positive* and 50 labeled *negative*. The total number of words in instances labeled *positive* is 3,000, while the total number of words in instances labeled *negative* is 2,000. The word *good* occurs 50 times in *positive* instances and 5 times in *negative* instances. Therefore its polarized weirdness with respect to the positive label is:

$$PW_{positive}(good) = \frac{50/3,000}{5/2,000} = 6.66$$

However, the polarized weirdness of *good* with respect to the negative label is:

$$PW_{negative}(good) = \frac{5/2,000}{50/3,000} = 0.15$$

indicating that *good* is much more indicative of *positiveness* than *negativeness*.

Polarized weirdness can be computed at a low computational cost on any dataset labeled with categorical values, with just tokenization for pre-processing. The outcome of the calculation of the polarized weirdness index is a set of rankings, one for each label, over the vocabulary, where the top words in the ranking relative to a given label l are the most characteristic for that label.

3.2 Word Embedding Adaptation

In Section 3.1, we introduced an automatic metric that allows us to compute how much a word is characteristic to a certain label. We use this information to transpose the vector representing words highly typical of a label closer to each other in the vector space. Formally, once a label has been decided and the polarized weirdness is computed with respect to it, *for each pair of vectors* \vec{v}_1, \vec{v}_2 in a word embedding model, representing words with polarized weirdness pw_1 and pw_2 respectively, we compute new representations:

$$\vec{v}_1 = ((1 - \alpha \cdot pw_1)\vec{v}_1) + ((\alpha pw_2)\vec{v}_2)$$

$$\vec{v}_2 = ((1 - \alpha \cdot pw_2)\vec{v}_2) + ((\alpha pw_1)\vec{v}_1)$$

where α is a parameter controlling the extent of the adaptation. The result of the application of this algorithm is a new word embedding model over the same vocabulary as the original model, where pairs of word vectors are closer in the space to an extent proportional to their respective polarized weirdness score.

4 Experimental Evaluation

We test the word embedding adaptation introduced in Section 3 by adapting pre-trained multilingual word embeddings to three different tasks. For each task, the polarized weirdness index is computed on the labeled training sets as described in Section 3.1, and the generic word embeddings are adapted to the particular task domain applying the algorithm described in Section 3.2.

Our baseline model is a convolutional neural network (CNN) with a 64x8 hidden layer and Rectified Linear Units activation (ReLU), followed by a 4-size max pooling layer. We use the implementation from the Keras Python library¹, with ADAM optimization (Kingma and Ba, 2014), leaving the hyperparameters at their default value, except for optimization of learning rate (set between 10^{-2} and 10^{-3} depending on the dataset) and number of epochs (between 10 and 25).

We use the multilingual word embeddings provided by Polyglot (Al-Rfou et al., 2013). These are distributed word representations for over 100 languages trained on Wikipedia. The vector representations of words in Polyglot are 64-dimensional. The choice of this model is motivated by the need to have word embedding models for different languages that were created with the same method, to be able to measure improvements introduced merely by our adaptation method. In these experiments, we set $\alpha = 0.5$.

4.1 Experiment 1: Multilingual Hate Speech Detection

In the first experiment, the generic word embeddings are adapted to provide a better representation for words used in online messages containing hate speech towards women and immigrants. We use the dataset provided by the SemEval Task 5 (HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter), a public challenge where participants are invited to submit the predictions of systems for hate speech

¹<https://keras.io/>

detection (Basile et al., 2019). In particular, we employ the data of the subtask A, where the prediction is binary (*hateful* vs. *not hateful*). The shared task website² provides datasets in Spanish and English, already divided into training, development and test sets. The topics of the messages are mainly two, namely women and immigrants, in a fairly balanced proportion. In fact, the dataset has been created by querying the Twitter API with a set of keywords crafted to capture these two topics. The English dataset comprises 13,000 tweets (10,000 for training and 3,000 for testing), with about 42% of the messages labeled as hateful. The Spanish dataset is smaller (6,600 tweets in total, 5,000 for training and 1,600 for testing), and it follows a similar distribution of topics and labels as the English set. Following are two examples of tweets from the English HatEval data, with their Hate Speech label:

I'd say electrify the water but that would kill wildlife. #SendThemBack
label: yes

Polish Prime Minister Mateusz Morawiecki insisted that Poland would push against any discussion on refugee relocations as part of the EU's migration politics.
label: no

Similarly, two examples of tweets from the Spanish HatEval data, with translation and label:

@rubenssambueza eres una basura de persona, lo cual no me sorprende porque eres SUDACA, y así son los tercermundistas
@rubenssambueza you are garbage, which does not surprise me because you are a SUDACA, and so are third-worlders
label: yes

Yo creía que ese jueguito solo existía para los árabes, jajaja.
I thought that this little game was only for arabs, ahahah.
label: no

The polarized weirdness of the words in the HatEval datasets (English and Spanish) is computed on the respective training sets as the ratio of their relative frequency in hateful messages over their relative frequency in non hateful messages. A modified version of the Polyglot embeddings is then

²<https://competitions.codalab.org/competitions/19935>

Table 1: Results of the English and Spanish Hate Speech Detection, for the negative (*no-HS*) and positive class (*HS*) and their macro-F1.

Model	Acc.	no-HS			HS			Avg. F1
		Pr.	R.	F1	Pr.	R.	F1	
English								
CNN	.468	.567	.401	.470	.398	.564	.466	.468
CNN+W	.482	.588	.394	.472	.413	.608	.492	.482
Spanish								
CNN	.528	.592	.595	.594	.437	.434	.436	.515
CNN+W	.527	.614	.497	.549	.450	.568	.502	.527

computed³ and the performance of the CNN using the adapted embeddings for initialization is compared with the performance obtained by initializing the CNN with the generic embeddings.

The results on the English dataset, presented in Table 1, show a clear improvement in the detection of hateful messages, leading to a +1.2% performance gain in macro-average F1-score. Recall is particularly impacted by the adapted embeddings, indicating that the modified model successfully helps in correcting false negatives.

The results on the Spanish HatEval task dataset, presented in Table 1 are even better than on English, with improvements in precision and recall for both the positive and the negative class, and a total gain of almost 2% macro-averaged F1-score. Similarly to English, the largest improvement is measured on the recall.

One of the advantages of the proposed method is that it is transparent with respect to the semantic shift computed on the pre-trained embeddings. Firstly, the words with the highest polarized weirdness index can be extracted, to gain insights into the specificity of the datasets. The top twenty weird words in the hateful English HatEval set are the following: nodaca, enddaca, kag, womensuck, @hillaryclinton, americafirst, trump2020, taxpayers, buildthewallnow, illegals, @senatemajldr, dreamer, buildthewall, they, @potus, walkawayfromdemocrat, votedem-sout, wethepeople, illegalalien, backtheblue. The top twenty weird words in the hateful Spanish HatEval set with English translations are the following: mantero (*street vendor*), turista (*tourist*), negratos (*nigger*), caloría (*calory*), sanidad (*health-care*), drogar (*to drug*), países (*countries*), emigrante (*immigrant*), Hija (*daughter*), ZORRA (*bitch*), impuesto (*tax*), zorro (*bitch (masculine)*),

³To speed up to computation without major loss of information, we consider only the top 2,000 items from the weirdness ranking.

Table 2: Examples of words from the HatEval datasets, showing how their vector representation moves to reflect the semantic shift. Particular words that are generally neutral get closer to offensive words in the hate speech context.

Word embeddings	Generic word	Offensive word	Semantic shift	Cosine distance
Polyglot EN	wall	fuck	yes	1.224
Polyglot EN + P.W.	wall	fuck	yes	0.444
Polyglot EN	car	fuck	no	1.279
Polyglot EN + P.W.	car	fuck	no	1.413
Polyglot ES	directora (<i>director (F)</i>)	puta (<i>whore</i>)	yes	1.271
Polyglot ES + P.W.	directora (<i>director (F)</i>)	puta (<i>whore</i>)	yes	1.222
Polyglot ES	director (<i>director (M)</i>)	puta (<i>whore</i>)	no	1.366
Polyglot ES + P.W.	director (<i>director (M)</i>)	puta (<i>whore</i>)	no	1.411

totalmente (*totally*), lleno (*full*), invasor (*invader*), costumbre (*custom*), barrio (*neighborhood*), PAIS (*country*), Oye (*hey*), Españoles (*Spaniards*).

Secondly, one can extract the word embeddings after the polarized weirdness adaptation is applied, and qualitatively inspect their respective position in the vector space. Table 2 shows how certain pairs of words become more related in the adapted space, while others are untouched by the process. The example in Spanish is particularly interesting (and worrying), where a misogynistic derogatory word (*puta*) becomes closer to the feminine inflection of “director” but not to the masculine inflection.

4.2 Experiment 2: Gender Prediction

In the second experiment, we test our word embedding adaptation method in a different scenario, that is, the prediction of the gender of the author of messages. The assumption is that the most typical words used by each gender will cluster in the vector representation, thus helping the model discriminate them better.

We use the dataset distributed for the Cross-Genre Gender Prediction in Italian (GxG) shared task of the 2018 edition of EVALITA, the evaluation campaign of language technologies for Italian (Dell’Orletta and Nissim, 2018). The participants to the shared task are invited to submit the prediction of their system on a set of short and medium-length texts in Italian from different sources, including social media, news articles and personal diaries, on the gender of the author. The task is therefore a binary classification, evaluated by means of accuracy. We downloaded the data from the task website⁴, comprising 22,874 in-

⁴<https://sites.google.com/view/gxg2018/>

Table 3: Results of the Gender Prediction.

Model	Acc.	Female			Male			Avg. F1
		Pr.	R.	F1	Pr.	R.	F1	
CNN	.511	.507	.879	.643	.543	.143	.227	.435
CNN+W	.513	.508	.851	.636	.539	.174	.263	.450

stances divided into training set (11,000) and test set (10,874). The labels of the GxG are perfectly balanced between M (male) and F (female).

Following are two examples of instances from the GxG dataset with their label and translation:

@ElfoBruno no la barba la devo tenere lunga per sembrare folta perchè in realtà è rada...

@ElfoBruno no I have to keep the beard long to make it look thick because it really is patchy...

label: M

Sabato prossimo sono davvero curiosa di scoprire cosa farà @Valerio_Scanu a #BallandoConLeStelle

Next Saturday I am very curious to find out what @Valerio_Scanu will do at #DancingWithTheStars

label: F

Since this is a *classification* rather than a *detection* task, the process is slightly different from the previous experiment, to account for the symmetry between the labels. First, the polarized weirdness is computed on the training set twice, once on the texts written by males (against the women’s texts) and once on the texts written by females (against the men’s texts). Then the general Polyglot embeddings are adapted by applying the algorithm in Section 3.2 twice, in both directions, using the respective weirdness rankings. The adapted embeddings are used to initialize the CNN, resulting

in the classification performance presented in Table 3. The overall performance improves when the adapted embeddings are included in the model. However, the classification of the *male* label improves while the classification of *female* does not, due to the difference in recall.

Qualitative analysis reveals interesting patterns, confirming that strong bias is present in some pre-trained word embedding models. The twenty top weird words in the Male GxG set are: *costituzionale* (*constitutional*), *socialisto* (*socialist*), *Lecce* (*name of a city and a football club*), *DALLA* (*name of a singer*), *utente* (*user*), *Samp* (*name of a football team*), *Sampdoria* (*same of a football team*), *Nera* (*black*), *allenatore* (*coach*), *Orlando* (*proper name*), *Bp* (*acronym*), *ni* (*yes and no*), *maresciallo* (*marshall*), *garanzia* (*guarantee*), *cerare* (*to wax*), *voluto* (*willing*), *pilotare* (*to pilot*), *disco* (*disco*), *caserma* (*barracks*), *From* (*proper name*).

The top twenty weird words in the Female GxG set are instead the following: *qualcuna* (*someone (feminine)*), *HEART EMOJI*, *Qualcuna* (*someone (feminine)*), *KISS EMOJI*, *83* (*number*), *essi* (*them*), *leonessa* (*lioness*), *Sarah* (*proper name*), *06* (*number*), *HEART-EYED EMOJI*, *nervoso* (*nervous*), *James* (*proper name*), *Dante* (*proper name*), *coreografia* (*choreography*), *Strada* (*street*), *Fra* (*proper name*), *Chiama* (*call*), *en* (*en*), *bravissimi* (*very good (plural)*), *Moratti* (*proper name*). Arguably, a stronger topic bias (football) is present in the male subset, possibly explaining the better performance induced by the adaptation.

5 Conclusion and Future Work

In this work, we adapted an extension of the weirdness index to score the words in a labeled corpus according to how much they are typical of a given label. The polarized weirdness score is used to automatically adapt an existing word embedding space to better reflect target-specific semantic associations of words. We measured a performance boost on tasks of hate speech detection in English and Spanish, and gender prediction in Italian.

On detection tasks, the improvement from our method is remarkable in terms of recall, indicating the potential of weirdness-adapted word embeddings to correct false negatives. This result is in line with the original motivation for this approach, i.e., to account for semantic shift oc-

curing in domain-specific corpora of opinionated content. For instance, in the hate speech domain, the adapted embeddings are able to capture that certain neutral words (e.g., “wall”) assume a polarized connotation (e.g., negatively charged).

The results from this study are promising, and encourage us to extend the method to richer representations (e.g., “weird” ngrams), languages other than European, and its integration into more sophisticated deep neural models. Recent Transformer models, in particular, compute contextualized embeddings, therefore including transformations similar to the present method. Although such models are less transparent with respect to such transformation, an experimental comparison is among the next steps planned in this research.

References

- Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, Maryland, November.
- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018a. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493. Association for Computational Linguistics.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018b. Projecting embeddings for domain adaption: Joint modeling of sentiment analysis in diverse domains. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 818–830. Association for Computational Linguistics.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.
- Cristina Bosco, Felice Dell’Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018.

- Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*.
- Felice Dell’Orletta and Malvina Nissim. 2018. Overview of the EVALITA 2018 cross-genre gender prediction (gxg) task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*.
- Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.
- Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).
- Prathusha Kameswara Sarma, Yingyu Liang, and Bill Sethares. 2018. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 51–59. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 329–332, Los Angeles, California, June. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Personal-ITY: A Novel YouTube-based Corpus for Personality Prediction in Italian

Elisa Bassignana

Dipartimento di Informatica
University of Turin

elisa.bassignana@edu.unito.it

Malvina Nissim

CLCG
University of Groningen

m.nissim@rug.nl

Viviana Patti

Dipartimento di Informatica
University of Turin

viviana.patti@unito.it

Abstract

We present a novel corpus for personality prediction in Italian, containing a larger number of authors and a different genre compared to previously available resources. The corpus is built exploiting Distant Supervision, assigning *Myers-Briggs Type Indicator (MBTI)* labels to YouTube comments, and can lend itself to a variety of experiments. We report on preliminary experiments on Personal-ITY, which can serve as a baseline for future work, showing that some types are easier to predict than others, and discussing the perks of cross-dataset prediction.

1 Introduction

When faced with the same situation, different humans behave differently. This is, of course, due to different backgrounds, education paths, and life experiences, but according to psychologists there is another important aspect: personality (Snyder, 1983; Parks and Guay, 2009).

Human Personality is a psychological construct aimed at explaining the wide variety of human behaviours in terms of a few, stable and measurable individual characteristics (Vinciarelli and Mohammedi, 2014).

Such characteristics are formalised in *Trait Models*, and there are currently two of these models that are widely adopted: *Big Five* (John and Srivastava, 1999) and *Myers-Briggs Type Indicator (MBTI)* (Myers and Myers, 1995). The first examines five dimensions (OPENNESS TO EXPERIENCE, CONSCIENTIOUSNESS, EXTROVERSION, AGREEABLENESS and NEUROTICISM) and for each of them assigns a score in a range. The

second one, instead, considers 16 fixed personality types, coming from the combination of the opposite poles of 4 main dimensions (EXTRAVERT-INTROVERT, INTUITIVE-SENSING, FEELING-THINKING, PERCEIVING-JUDGING). Examples of full personality types are therefore four letter labels such as ENTJ or ISFP.

The tests used to detect prevalence of traits include human judgements regarding semantic similarity and relations between adjectives that people use to describe themselves and others. This is because language is believed to be a prime carrier of personality traits (Schwartz et al., 2013). This aspect, together with the progressive increase of available user-generated data on social media, has prompted the task of *Personality Detection*, i.e., the automatic prediction of personality from written texts (Youyou et al., 2015; Argamon et al., 2009; Litvinova et al., 2016; Whelan and Davies, 2006).

Personality detection can be useful in predicting life outcomes such as substance use, political attitudes and physical health. Other fields of application are marketing, politics and psychological and social assessment.

As a contribution to personality detection in Italian, we present Personal-ITY, a new corpus of YouTube comments annotated with MBTI personality traits, and some preliminary experiments to highlight its characteristics and test its potential. The corpus is made available to the community¹.

2 Related Work

There exist a few datasets annotated for personality traits. For the shared tasks organised within the *Workshop on Computational Personality Recognition* (Celli et al., 2013), two datasets annotated with the *Big Five* traits have been released in 2013

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://github.com/elisabassignana/Personal-ITY>

Corpus	Model	# user	Avg.
PAN2015	Big Five	38	1258
TwISTY	MBTI	490	21.343
Personal-ITY	MBTI	1048	10.585

Table 1: Summary of Italian corpora with personality labels. Avg.: average tokens per user.

(Essays (Pennebaker and King, 2000) and myPersonality²) and two in 2014 (YouTube Personality Dataset (Biel and Gatica-Perez, 2013) and Mobile Phones interactions (Staiano et al., 2012)).

For the 2015 PAN Author Profiling Shared Task (Pardo et al., 2015), personality was added to gender and age in the profiling task, with tweets in English, Spanish, Italian and Dutch. These are also annotated according to the *Big Five* model.

Still in the Big Five landscape, Schwartz et al. (2013) collected a dataset of FaceBook comments (700 millions words) written by 136.000 users who shared their status updates. Interesting correlations were observed between word usage and personality traits.

If looking at data labelled with the MBTI traits, we find a corpus of 1.2M English tweets annotated with personality and gender (Plank and Hovy, 2015), and the multilingual TwISTY (Verhoeven et al., 2016). The latter is a corpus of data collected from Twitter annotated with MBTI personality labels and gender for six languages (Dutch, German, French, Italian, Portuguese and Spanish) and a total of 18,168 authors. We are interested in the Italian portion of TwISTY.

Table 1 contains an overview of the available Italian corpora labelled with personality traits. We include our own, which is described in Section 3.

Regarding detection approaches, Mairesse et al. (2007) tested the usefulness of different sets of textual features making use of mostly SVMs.

At the PAN 2015 challenge (see above) a variety of algorithms were tested (such as Random Forests, decision trees, logistic regression for classification, and also various regression models), but overall most successful participants used SVMs. Regarding features, participants approached the task with combinations of style-based and content-based features, as well as their combination in n -gram models (Pardo et al., 2015).

Experiments on TwISTY were performed by

²<http://mypersonality.org>

the corpus creators themselves using a LinearSVM with word (1-2) and character (3-4) n -grams. Their results (reported in Table 2 for the Italian portion of the dataset) are obtained through 10-fold cross-validation; the model is compared to a weighted random baseline (WRB) and a majority baseline (MAJ).

Trait	WRB	MAJ	f-score
EI	65.54	77.88	77.78
NS	75.60	85.78	79.21
FT	50.31	53.95	52.13
PJ	50.19	53.05	47.01
Avg	60.41	67.67	64.06

Table 2: TwISTY scores from the original paper. Note that all results are reported as *micro-average* F-score.

3 Personal-ITY

First, we explain two major choices that we made in creating Personal-ITY, namely the source of the data and the trait model. Second, we describe in detail the procedure we followed to construct the corpus. Lastly, we provide a description of the resulting dataset.

Data YouTube is the source of data for our corpus. The decision is grounded on the fact that compared to the more commonly collected tweets, YouTube comments can be longer, so that users are freer to express themselves without limitations. Additionally, there is a substantial amount of available data on the YouTube platform, which is easy to access thanks to the free YouTube APIs.

Trait Model Our model of choice is the MBTI. The first benefit of this decision is that this model is easy to use in association with a Distant Supervision approach (just checking if a message contains one of the 16 personality types; see Section 3.1). Another benefit is related to the existence of TwISTY. Since both TwISTY and Personal-ITY implement the MBTI model, analyses and experiments over personality detection can be carried out also in a cross-domain setting.

Ethics Statement

Personality profiling must be carefully evaluated from an ethical point of view. In particular, often, personality detection involves ethical dilem-

mas regarding appropriate utilization and interpretations of the prediction outcomes (Weiner and Greene, 2017). Concerns have been raised regarding the inappropriate use of these tests with respect to invasion of privacy, cultural bias and confidentiality (Mehta et al., 2019).

The data included in the Personal-ITY dataset were publicly available on the YouTube platform at the time of the collection. As we will explain in detail in this Section, the information collected are comments published under public videos on the YouTube platform by authors themselves. For a major protection of user identities, in the released corpus only the YouTube usernames of the authors are mentioned which are not unique identifiers. The YouTube IDs of the corresponding channels, which are the real identifiers in the platform, allowing to trace the identity of the authors, are not released. Note also that the corpus was created for academic purposes and is not intended to be used for commercial deployment or applications.

3.1 Corpus Creation

The fact that users often self-disclose information about themselves on social media makes it possible to adopt *Distant Supervision* (DS) for the acquisition of training data. DS is a semi-supervised method that has been abundantly and successfully used in affective computing and profiling to assign silver labels to data on the basis of indicative proxies (Go et al., 2009; Pool and Nissim, 2016; Emery et al., 2017).

Users left comments to some videos on the MBTI theory in which they were stating their own personality type (e.g. *Sono ENTJ...chi altro?* [en: "I'm ENTJ...anyone else?"]). We exploited such comments to create Personal-ITY with the following procedure.

First, we searched for as many Italian YouTube videos about MBTI as possible, ending up with a selection of ten with a conspicuous number of comments as the ones above³.

Second, we retrieved all the comments to these videos using an AJAX request, and built a list of authors and their associated MBTI label. A label

³Links to the 10 YouTube videos:

<https://www.youtube.com/watch?v=VCo9R1DRpz0>
<https://www.youtube.com/watch?v=N4kC8iqUNyk>
<https://www.youtube.com/watch?v=Z8S8PgW8t2U>
<https://www.youtube.com/watch?v=wH2OG8k7nSw>
https://www.youtube.com/watch?v=1O2z3_DINgs
https://www.youtube.com/watch?v=NaKPl_y1JXg
<https://www.youtube.com/watch?v=814o4VBX1GY>
<https://www.youtube.com/watch?v=GK5J6PLj218>
<https://www.youtube.com/watch?v=9P95dkVLmps>
<https://www.youtube.com/watch?v=g0ZIFNgUmoE>

Comment	User - MBTI label
<i>Io sono ENFJ!!!</i>	User1 - ENFJ
<i>Ho sempre saputo di essere connessa con Lady Gaga! ISFP!</i>	User2 - ISFP

Table 3: Examples of automatic associations *user* - *MBTI personality type*.

was associated to a user if they included an MBTI combination in one of their comments. Table 3 shows some examples of such associations. The association process is an approximation typical of DS approaches. To assess its validity, we manually checked 300 random comments to see whether the mention of an MBTI label was indeed referred to the author's own personality. We found that in 19 cases (6.3%) our method led to a wrong or unsure classification of the user's personality (e.g. *O tutti gli INTJ del mondo stanno commentando questo video oppure le statistiche sono sbagliate :-)*). We can assume that our dataset might therefore contain about 6-7% of noisy labels.

Using the acquired list of authors, we meant to obtain as many comments as possible written by them. The YouTube API, however, does not allow to retrieve all comments by one user on the platform. In order to get around this problem we relied on video similarities, and tried to expand as much as possible our video collection. Therefore, as a third step, we retrieved the list of channels that feature our initial 10 videos, and then all of the videos within those channels.

Fourth, through a second AJAX request, we downloaded all comments appearing below all videos retrieved through the previous step.

Lastly, we filtered all comments retaining those written by authors included in our original list. This does not obviously cover all comments by a relevant user, but it provided us with additional data per author.

3.2 Final Corpus Statistics

For the final dataset, we decided to keep only the authors with a sufficient amount of data. More specifically, we retained only users with at least five comments, each at least five token long.

Personal-ITY includes 96,815 comments by 1048 users, each annotated with an MBTI label. The average number of comments per user is 92

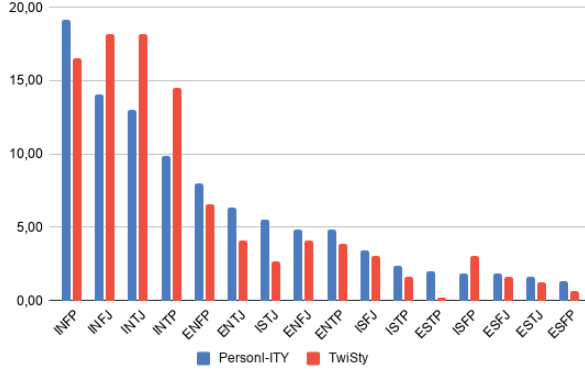


Figure 1: Distribution of the 16 personality types in the YouTube corpus and in the Italian section of TWISTY.

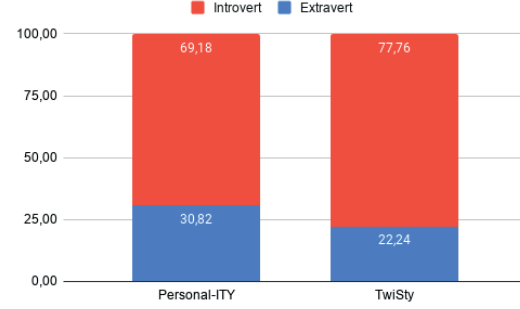
and each message has on average 115 tokens.

The amount of the 16 personality types in the corpus is not uniform. Figure 1 shows such distribution and also compares it with the one in TWISTY. The unbalanced distribution can be due to personality types not being uniformly distributed in the population, and to the fact that different personality types can make different choices about their online presence. Goby (2006) for example, observed that there is a significant correlation between online–offline choices and the MBTI dimension of EXTRAVERT-INTROVERT: extroverts are more likely to opt for offline modes of communication, while online communication is presumably easier for introverts. In Figure 1, we also see that the four most frequent types are introverts in both datasets. The conclusion is that, despite the different biases, collecting linguistic data in this way has the advantage that it reflects actual language use and allows large-scale analysis (Plank and Hovy, 2015).

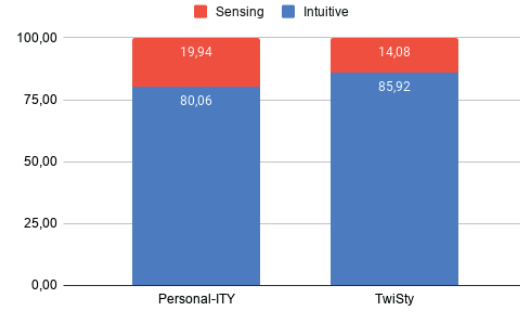
Figure 2 shows more in detail, trait by trait, the distribution of the opposite poles through the users in Personal-ITY and in TWISTY. As we might have expected, in line with what is observed in Figure 1, the two datasets present very similar trends. Such similarities between Personal-ITY and TWISTY are these similarities are a further confirmation of the reliability of the data we collected.

4 Preliminary Experiments

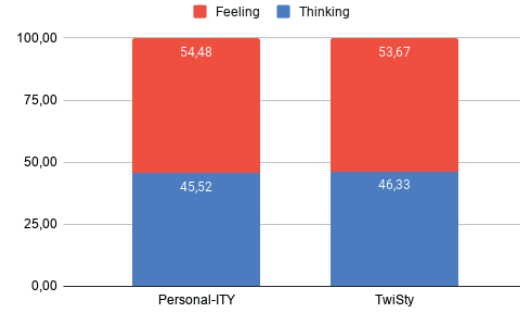
We ran a series of preliminary experiments on Personal-ITY which can also serve as a baseline for future work on this dataset. We pre-processed texts by replacing hashtags, urls, usernames and



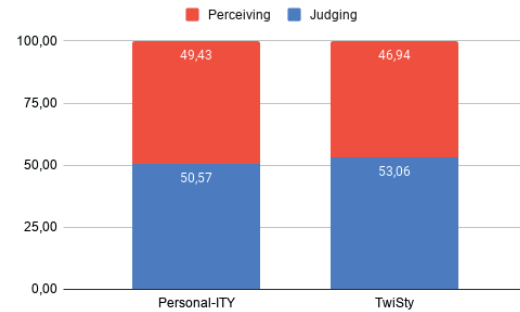
(a) *Extravert - Introvert*



(b) *Sensing - Intuitive*



(c) *Thinking - Feeling*



(d) *Judging - Perceiving*

Figure 2: Comparison of the distributions of the four MBTI traits between Personal-ITY and the Italian part of TWISTY.

emojis with four corresponding placeholders. We adopted the `sklearn` (Pedregosa et al., 2011) implementation of a linear SVM (LinearSVM), with standard parameters. We tested three types of features. At the lexical level, we experimented with word (1-2) and character (3-4) n -grams, both as raw counts as well as tf-idf weighted. Character n -grams were tested also with a word-boundary option. At a more stylistically level, we considered the use of emojis, hashtags, pronouns, punctuation and capitalisation. Lastly, we also experimented with embeddings-based representations, by using, on the one hand, YouTube-specific (Nieuwenhuis and Nissim, 2019) pre-trained models, on the other hand, more generic embeddings, such as the Italian version of GloVe (Pennington et al., 2014), which is trained on the Italian Wikipedia⁴. We looked for all the available embeddings of the words written by each author, and used the average as feature. If a word appeared more than once in the string of comments, we considered it multiple times in the final average.

We used 10-fold cross-validation, and assessed the models using macro f-score. Note that the original TWISTY paper uses micro f-score. Thus, for the sake of comparison, we include also micro-F in Table 5 for the MAJ baseline and our lexical n -gram model. Table 4 shows the results of our experiments with different feature types.⁵ Overall, lexical features (n -grams) perform best. Combining different feature types did not lead to any improvement. Classification was performed with four separate binary classifiers (one per dimension), and with one single classifier predicting four classes, i.e. the whole MBTI labels at once. In the latter case, we observe that the results are quite high considering the increased difficulty of the task. Table 5 reports the scores of our models on TWISTY. As for Personal-ITY, best results were achieved using lexical features (tf-idf n -grams); stylistic features and embeddings are just above the baseline. Our model outperforms the one in (Verhoeven et al., 2016) for all traits (micro-F).

To test compatibility of resources and to assess model portability, we also ran cross-domain experiments on Personal-ITY and TWISTY. In the first setting, we tested the effect of merging the

⁴<https://hlt.isti.cnr.it/wordembeddings>

⁵In Tables 4–5, we report the highest scores based on averages of the four traits. Considering the dimensions individually, better results can be obtained by using specific models.

Trait	MAJ	Lex	Sty	Emb	FL
EI	40.55	51.85	40.46	40.55	51.65
NS	44.34	51.92	44.34	44.34	49.04
FT	35.01	50.67	36.27	35.01	50.86
PJ	29.49	50.53	51.04	47.06	51.03
Avg	37.35	51.24	43.03	41.74	50.65

Table 4: Results of the experiments on Personal-ITY. FL: prediction of the full MBTI label at once, with a character n -gram model.

	micro F		macro F			
Trait	MAJ	Lex	MAJ	Lex	Sty	Emb
EI	77.75	79.18	43.69	55.23	43.69	43.69
NS	85.92	85.92	46.15	46.15	46.15	46.15
FT	53.67	55.31	34.79	52.98	35.34	34.70
PJ	53.06	54.08	34.56	53.01	35.20	34.90
Avg	67.6	68.62	39.80	51.84	40.09	39.86

Table 5: Results of our experiments on TWISTY.

two datasets on the performance of models for personality detection, maintaining the 10-fold cross-validation setting and by using the model performing better on average for YouTube and Twitter data (a character n -grams model). Table 6 contains the result of such experiments⁶. Scores are almost always lower compared to the in-domain experiments (excepts for NS as regards Twitter scores reported in Table 5: 46.15 \rightarrow 48.31), but quite increased compared to the majority baseline.

Trait	MAJ	Lex
EI	41.64	50.57
NS	44.93	48.31
FT	35.04	51.31
PJ	30.66	48.24
Avg	38.07	49.61

Table 6: Merging Personal-ITY with TWISTY.

In the second setting, instead, we divided both corpora in fixed training and test sets with a proportion of 80/20 and ran the models using lexical features, in order to run a cross-domain experiment. For direct comparison, we run the model in-domain again using this split. Results are shown

⁶Prediction of the full label at once.

Train	Personal-ITY			TWiSTY		
Test	IN		CROSS	IN		CROSS
	Pers	MAJ	Twi	Twi	MAJ	Pers
EI	58.94	44.94	49.33	55.66	44.59	44.59
NS	52.88	47.87	47.31	47.87	45.31	45.31
FT	49.20	37.58	47.09	65.26	39.13	51.04
PJ	54.43	32.41	32.50	56.87	36.56	38.54
Avg	53.86	40.70	44.06	56.42	41.40	44.87

Table 7: Results of the cross-domain experiments. MAJ = baseline on the cross-domain testset.

in Table 7. Cross-domain scores are obtained with the best in-domain model.⁷ They drop substantially compared to in-domain, but are always above the baseline.

5 Conclusions

The experiments show that there is no single best model for personality prediction, as the feature contribution depends on the dimension considered, and on the dataset. Lexical features perform best, but they tend to be strictly related to the context in which the model is trained and so to overfit.

The inherent difficulty of the task itself is confirmed and deserves further investigations, as assigning a definite personality is an extremely subjective and complex task, even for humans.

Personal-ITY is made available to further investigate the above and other issues related to personality detection in Italian. The corpus can lend itself to a psychological analysis of the linguistic cues for the MBTI personality traits. On this line, it is interesting to investigate the presence of evidences linking linguistic features with psychological theories about the four considered dimensions (EXTRAVERT-INTROVERT, INTUITIVE-SENSING, FEELING-THINKING, PERCEIVING-JUDGING). First results in this direction are presented in (Bassignana et al., 2020).

Acknowledgments

The work of Elisa Bassignana was partially carried out at the University of Groningen within the framework of the Erasmus+ program 2019/20.

⁷Better results can be obtained with other specific models.

References

- Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, February.
- Elisa Bassignana, Malvina Nissim, and Viviana Patti. 2020. Personal-ITY: a YouTube Comments Corpus for Personality Profiling in Italian Social Media. In Viviana Patti, Malvina Nissim, and Barbara Plank, editors, *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, (PEOPLES@COLING 2020)*. Association for Computational Linguistics.
- Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55.
- Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition: Shared task. In *Seventh International AAAI Conference on Weblogs and Social Media*.
- Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on Twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.
- Valerie Goby. 2006. Personality and Online/Offline Choices: MBTI Profiles and Favored Communication Modes in a Singapore Study. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 9:5–13, 03.
- Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality: Theory and research*, page 102–138. Guilford Press.
- Tatiana Litvinova, P. Seredin, Olga Litvinova, and Olga Zagorovskaya. 2016. Profiling a set of personality traits of text author: What our words reveal about us. *Research in Language*, 14, 12.
- François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, sep.
- Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

- I.B. Myers and P.B. Myers. 1995. *Gifts Differing: Understanding Personality Type*. Mobius.
- Moniek Nieuwenhuis and Malvina Nissim. 2019. The Contribution of Embeddings to Sentiment Analysis on YouTube. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Francisco M. Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Laura Parks and Russell P Guay. 2009. Personality, values, and motivation. *Personality and individual differences*, 47(7):675–684.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- James Pennebaker and Laura King. 2000. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77:1296–312, 01.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal, September. Association for Computational Linguistics.
- Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. The COLING 2016 Organizing Committee.
- H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.
- Mark Snyder. 1983. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of personality*, 51(3):497–516.
- Jacopo Staiano, Bruno Lepri, Nadav Aharoni, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. Friends don’t lie - inferring personality traits from social network structure. In *UbiComp’12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 321–330, 09.
- Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1632–1637, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.
- Irving B. Weiner and Roger L. Greene, 2017. *Ethical Considerations In Personality Assessment*, chapter 4, pages 59–74. Wiley.
- Susan Whelan and Gary Davies. 2006. Profiling consumers of own brands and national brands using human personality. *Journal of Retailing and Consumer Services*, 13(6):393–402.
- Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

The “Corpus Anchise 320” and the analysis of conversations between healthcare workers and people with dementia

Nicola Benvenuti
Università di
Torino

Andrea Bolioli
CELI

Alessio Bosca
CELI

Alessandro Mazzei
Università di
Torino

Pietro Vigorelli
Gruppo Anchise

Abstract

The aim of this research was to create the first Italian corpus of free conversations between healthcare professionals and people with dementia, in order to investigate specific linguistic phenomena from a computational point of view. Most of the previous researches on speech disorders of people with dementia have been based on qualitative analysis, or on the study of a few dozen cases executed in laboratory conditions, and not in spontaneous speech (in particular for the Italian language). The creation of the Corpus Anchise 320 aims to investigate Dementia language by providing a broader number of dialogues collected in ecological conditions. Automatic linguistic analysis can help healthcare professionals to understand some characteristics of the language used by patients and to implement effective dialogue strategies.¹

Introduction

In this paper we will present the construction of the first annotated corpus of conversations between healthcare workers and people with dementia for Italian, called “Corpus Anchise 320”, and the quantitative linguistic analysis we carried out. The aim of the project is twofold. On the one hand, we created a dataset of spoken dialogue transcriptions that is useful for research on the language of people with dementia. On the other hand, techniques typical of computational linguistics are applied to help doctors in assessing the state of the disease and implement effective dialogue strategies. Focusing attention on verbal exchanges between speakers is one of the

cornerstones of the approach developed by the Anchise Group to support people with dementia and their caregivers, i.e. the “Enabling Approach” (Vigorelli 2018).

The paper is divided in 4 sections. Section 1 introduces the topic of Alzheimer’s language. Section 2 presents the recent researches and related works. In Section 3, the creation of the Corpus Anchise 320 will be discussed, which collects the transcripts and annotations of a set of dialogues between healthcare professionals and dementia patients carried out by the Anchise Group from 2007 until today, in Italian language. Section 4 will report the results of the computational linguistic analysis with the StanfordNLP library for Italian. The results obtained will be discussed to outline some of the peculiarities of the Dementia language. Section 5 concludes this paper with some final considerations.

1 The Alzheimer’s language

Dementia refers to a series of symptoms that manifest in “*difficulties with memory, language, problem solving and other cognitive skills that affect a persons ability to perform everyday activities.*” (Alzheimer’s-Association 2018, 368). These symptoms change over time and reflect the degree of neuronal damage in different parts of the brain. Alzheimer’s disease (AD), a neurodegenerative brain disease, is the most common form of dementia. One of the most popular neuropsychological tests for assessing a patient’s neurocognitive and functional status is still the *Mini-Mental Test*, designed by Folstein et al. (1975).

The first symptoms are memory loss or a state of frequent confusion. Alzheimer’s disease, semantic dementia, aphasia and amnesia all share

¹ Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

a close link with lexical memory and therefore declarative memory, while they would leave grammar and procedural memory intact. Language would thus move between a structural component, formed by grammatical rules that are stabilized over the course of life and are preserved longer as a crystallized function; and a semantic component that would collapse more quickly because it requires a mnemonic and contextualized effort that makes the cognitive activity of the individual more complex. This dissociation is confirmed by studies on Alzheimer's language (Almor 1999), (Kempler 2008), (Bucks 2000) in which it has been amply demonstrated that one of the first symptoms is anomia, or the difficulty in finding the lexical target; as opposed to a good ability to construct the sentence up to the advanced stages of the disease. These deficits would then be compensated through linguistic strategies, such as the high use of pronouns, circumlocutions and passepartout words present in the speech of Alzheimer's patients: *“empty words (“things”, “do”, “he”, “it”, etc.) are successfully and relatively easily activated precisely because they are high in frequency and allow the patients to produce fluent and grammatical sentences in the presence of debilitating semantic deficits”* (Almor 1999, 205). In the more advanced stages of the disease, communication becomes increasingly problematic as Alzheimer's patients experience difficulties in understanding and constructing a coherent discourse: *“their narratives are often repetitive with topic changes, unclear references, and lack of coherence and informativeness”* (Kempler 2008, 76).

2 Related works

The recent workshop on the creation of medical dialogues corpora (Bhatia et al. 2020) is a consequence of an increasing interest on this specific application field. The main reason of this interest is on the possibility of design and realize software applications which can assist professionals in medicine in their daily work in order to avoid errors: *“It is imperative to find a solution to minimize causes of such errors, via better tooling and visualization or by providing automated decision support assistants to medical practitioners.”*. With this final aim, the creation of medical dialogues corpora can be seen as a first step toward the creation of a virtual medical assistant that can assist, speed-up, improve the capacities of medical practitioners.

As stated in (de la Fuente Garcia 2020) *“datasets containing both clinical information and spontaneous speech suitable for statistical learning are relatively scarce. In addition, speech data are often collected under different conditions, such as monologue and dialogue recording protocols.”* A notable example is the Carolina Conversations Collection (CCC), that is amongst the few spontaneous dialogue datasets available in English in the context of AD research. It is hosted and distributed by the Medical University of South Carolina (Pope 2011).

The study of AD language with computational methods is fairly recent, but a number of work showed the applicability of symbolic and statistic algorithms for the prediction of dementia and similar diseases (Karlekar et al. 2018, Mirheidari et al. 2019, Kong et al. 2019).

In (Karlekar et al. 2018) neural networks have been used on the publicly available DementiaBank dataset in order to predict Alzheimer's dementia of a patient starting from the language produced and annotated with the POS feature. They reached precision result between 80-85%. Interestingly, they showed that there is no significative difference between the prediction results by considering the gender.

In (Mirheidari et al. 2019) an automatic dementia detection system was presented, including a diarisation unit, an automatic speech recogniser, conversation analysis (CA) based acoustic and lexical feature extraction module and a machine learning classifier, in order to facilitate and improve screening procedures for dementia. They showed that using these features, they can obtain a high value of precision in detecting dementia for both a neurologist-patient and VirtualAgent-patient conversations.

In (Kong et al. 2019) neural networks on DementiaBank dataset have been used too. They reached precision results close to the state of art (80-85%), but they pointed out on the scalability of their neural methods that need less data. Moreover, they showed that *“the attention mechanism of the model manages to capture similar key concepts as the information unit features specified by human experts.”*

As for Italian language, in (Beltrami 2018) the participants (both healthy and cognitively impaired) were asked to answer to three specific tasks, i.e. the description of a drawing, details of a last dream and the description of a working day. The researchers investigated whether the analysis performed by Natural Language Processing

techniques could reveal alterations of the language performance in early cognitive decline.

3 The Corpus Anchise 320

Corpus Anchise 320 collects the transcripts of dialogues between healthcare professionals and patients carried out over the period from 2007 until today by the Anchise Group, an association of experts (doctors, psychologists, nurses, trainers) for the research, training and care of the elderly with dementia. The corpus consists of an unselected series of people diagnosed with dementia and not only those with an established diagnosis with specific criteria for Alzheimer's were included. For probabilistic reasons, most patients are affected by AD. The corpus contains 320 individual conversations resulting from transcription of about 15 minutes of dialogue for each patient in which the patient can speak freely with the health worker. This peculiarity is of considerable importance in a field of investigation that was mainly based on "*formal medical-psychological situations of the anamnestic investigation and the collection of tests*" (Lai 2000).

The corpus contains 20,588 turns of conversation, consisting of 10,193 turns of patients with dementia and 10,381 turns of health workers. The total number of tokens is 222,856 and the total number of types (different words) is 14,513. In the table below we present a small portion of one conversation.

7	P	<i>Eh ma mia figlia... è dura, è dura.</i> [Eh but my daughter... it's hard, it's hard.]
8	O	<i>Lo sa Marta che da quando abbiamo iniziato a vederci lei parla molto, molto meglio?</i> [You know, Marta, that since we started seeing each other you talk much, much better?]
9	P	<i>Ma a casa mia no! Loro non capiscono. Han detto che non capiscono niente...</i> [But not at home! They don't understand. They said they don't understand anything...]
10	O	<i>Forse sono loro che non capiscono.</i> [Maybe it is they who do not understand.]

Table 1: An excerpt from a conversation between a patient (P) and an health worker (O), with English translation (turn 7 to 10).

The corpus has been created in two phases. In the first phase, health professionals of Anchise Group created the audio recordings, transcribed portions of dialogues and annotated each transcription with a series of metadata with the aim of investigating the relationship between the language, age, sex and stage of dementia (MMSE² score). In the second phase, we collected the 320 transcriptions, we removed pragmalinguistic comments of health professionals, such as "[Touch the recorder]", "[Silence]", "[Laughs]", etc., and we analyzed and annotated the corpus as described in the following sections.

Corpus Anchise 320 has been built and archived according to EU General Data Protection Regulation (GDPR). Audio recording and transcriptions were made with the consent of the speaker, as far as possible, of the family member and of the head of the facility or department. Personal data have been anonymized. The dataset is not publicly available but it can be requested to the authors for research purposes.

4 Computational linguistic analysis of Corpus Anchise 320

In this Section we will discuss the results of the lexical analysis (3.1) and of the morphosyntactic analysis (3.2) carried out on the Corpus Anchise 320.

3.1 Lexical analysis

The Corpus Anchise 320 contains 222,856 tokens and 14,513 types. The relationship between types and tokens constitutes the *Types-Token Ratio* (TTR), which represents a type of index to calculate the lexical richness of a text (Torruella 2013). The number of tokens and types were subsequently calculated for the patient's total turns and the health worker's total turns. The results are shown in Table 2.

	<i>Token</i>	<i>Types</i>	<i>TTR</i>
Corpus Anchise	222.856	14.513	0,07
Patients	144.405	8.499	0,06
Health workers	78.451	6.014	0,08

Table 2: Token/types data.

² Mini-Mental State Examination.

TTR is low for both speakers. As for the patient, this trend is closely linked to Alzheimer's disease, in which *“the production of high-frequency words is relatively preserved while the production of low-frequency is impaired”* (Almor 1999, 204). As for the health professional, this trend reflects the Grice Principle of Cooperation between speakers in which it is necessary to conform the conversational contribution to what is required, when it occurs, by the accepted common intent or by the direction of the verbal exchange. Finally, if look at the number of tokens, we get that the patients speak more but with a poorer vocabulary relative to the lower lexical richness index than the sample of the health workers.

A frequency list was then created on the corpus sample of patients with dementia. The table 3 is the result of a pre-processing phase where 4 types of function words have been removed from the frequency list, i.e. adpositions, determiners, conjunctions and auxiliaries.

From the analysis of the data it emerges that the first 50 words in order of frequency cover 32% of the entire Corpus Anchise 320 and 49.4% of the patients' speech; the first 100 words cover 40.0% of the entire corpus and 61.8% of that of patients with dementia; the first 200 words cover 46.7% of the entire corpus and 72.1% of the words used by patients. This means that, on an expressive level, patients with the use of 200 words cover almost three quarters of all the vocabulary used in these conversations.

	Words	Frequency		Words	Frequency
1	non	3.954	10	adesso	687
2	si	3.123	11	li	666
3	mi	2.272	12	mia	639
4	io	2.063	13	qui	639
5	no	1.546	14	casa	637
6	eh	1.343	15	me	612
7	anche	1032	16	fare	577
8	bene	1007	17	lei	552
9	cosa	768	18	so	535

Table 3: Words frequency of patients with dementia.

The analysis of the words most used by patients diagnosed with dementia present in the Corpus Anchises 320 shows a high percentage of deictics, such as “io” (“me”), “qui” (“here”), “lì” (“there”), and the presence of semantically empty words, such as “cosa” (“thing”) and “cose”

(“things”). This attitude confirms the scientific research carried out so far on Alzheimer's language regarding word finding deficits: *“the earliest language deficits observed in DAT is anomia. (...) Semantically empty words are scattered throughout the DAT patient's utterances in place of content words, thereby maintaining fluency and sacrificing informational content.”* (Kempner 1991, 98). From the analysis of the first 100 words used, we note the presence of the words “casa” (“home”) with 637 occurrences, “mamma” (“mother”) with 394 occurrences, “marito” (“husband”) with 190 occurrences, “figli” (“children”) with 162 occurrences. As the corpus contains spontaneous speech, we can note that the most common topic is the patient's family.

3.2 Morphosyntactic analysis

The Corpus Anchise 320 was analyzed morpho-syntactically by means of the StanfordNLP library in Python language (Qi 2018). The default pre-trained neural model for Italian was used. Specifically, tokenization, lemmatization, POS tagging and Dependency parsing were carried out. These annotations, i.e. ID, Form, Lemma, POS, FEATS, HEAD, DEPREL, were organized according to the CoNLL-U format (Zeman 2018, Bosco 2014). A linguist reviewed the automatic annotations.

ID	TOKEN	XPOS	LEMMA	FEATS	HEAD	DEPREL
3	che	PRON	che	...	4	nsubj
4	fa	VERB	fare	...	2	acl:relcl
5	un	DET	uno	...	6	det
6	po'	ADV	poco	...	7	advmod
7	fatica	NOUN	fatica	...	4	obj
8	a	ADP	a	...	9	mark
9	parlare	VERB	parlare	...	4	xcomp

Table 4: An excerpt from the annotated corpus. Features are not shown due to space constraints.

The analysis of the linguistic data of patients suffering from dementia was made using both the LIP³ corpus (De Mauro 1993, 155) and the speech corpus of healthcare professionals as a reference.

³ *Lessico di frequenza dell'italiano parlato*

The analysis of the percentages of occurrence of the parts of speech, in the patient corpus sample, reveals a superior use of pronouns and adverbs both with respect to LIP and with respect to the corpus of health workers. With reference to the LIP, the use of pronouns records 10.9% of occurrence, while the use of adverbs 10.1%. If we compare these data with the rates of occurrence in the patients' speech (Table 5), 13.9% frequency for pronouns and 14.2% for adverbs respectively, we notice a notable difference. Furthermore, these two indices, when added together, are 1.7 percentage points higher than the health workers' speech (ADV 13,2%, PRON 13,2 %). This trend would confirm what was said in the analysis relating to word frequency, i.e. the difficulty for patients to access the lexicon and therefore to compensate for this deficit with the use of deictics, closely linked to the context. If we cross these data with the rate of names used by patients (1.6 percentage points lower than the corpus of the health workers, NOUN 13,2%), we can deduce that the patient implements a real compensatory strategy linked to the impairment of access to semantic memory. A significant difference is also present with the LIP, which records a rate of names of 15.7% against 11.6% of the corpus relating to patients with dementia.

	<i>Patients</i>	<i>%</i>
ADJ	4.843	3,3
ADP	11.177	7,7
ADV	20.560	14,2
AUX	10.586	7,3
CCONJ	5.562	3,8
DET	14.200	9,8
INTJ	7.383	5,1
NOUN	16.787	11,6
NUM	1.145	0,7
PRON	20.118	13,9
PROP	1.799	1,2
SCONJ	5.078	3,5
VERB	24.749	17,1
X	418	0,3
TOT.	144.405	

Table 5: Percentages of occurrence of the parts of speech.

At the morphosyntactic level, it is known in the literature that Alzheimer's patients do not suffer from serious deficits in the construction of the sentence: *"sentence production in DAT is characterized by intact morphosyntactic structure*

(i.e., subject verb agreement, well formed plural and tense markings)", (Kempler 2008, 75). However, some linguistic phenomena that emerged from the analysis of the occurrences of the verbal system could be linked to a spatial-temporal disorientation characteristic of Alzheimer's disease (Macri 2016). This disorientation is reflected in the massive use of the indicative mode present in 95.9% of cases (Table 6). The use of the subjunctive and conditional modes appears to be almost minimal with percentages that are around 1%. This tendency could be paraphrased in terms of cognitive work, since the two verbal modes require both the ability to imagine possible worlds and - at the level of sentence construction - of conjugation and temporal concordance.

Verb form	Fin	Inf	Part	Ger
	25.771	4.655	4.751	158
	(72,9%)	(13,2%)	(13,4%)	(0,4%)
Mood	Ind	Sub	Imp	Cnd
	24.702	452	326	285
	(95,9%)	(1,8%)	(1,2%)	(1,1%)
Tense	Pres	Past	Imp	Fut
	21.958	4.800	3.459	304
	(71,9%)	(15,7%)	(11,33%)	(0,9%)

Table 6:: Percentages of occurrence of the verbal system.

5 Conclusion and further research

In this paper we presented the first Italian corpus of conversations between healthcare professionals and people with dementia, called "Corpus Anchise 320". The study of this corpus with computational linguistic analysis confirmed some characteristics of the language of people with dementia, such as the reduction in the rate of names and the increase in deictics. Corpus Anchise 320 has been built and archived according to GDPR. It is not publicly available but it can be requested to the authors for research purposes.

The large number of the sample (320 conversations) and the use of computational analysis will make it possible to identify indicators of pathological language to be used in the preclinical phase, to trace the change in the linguistic abilities of people with dementia as the disease progresses, to put in relation the characteristics of the pathological language with a

series of metalinguistic data such as age, sex and degree of dementia. The corpus will be increased in the coming months with the addition and annotation of other transcripts of dialogues of people with dementia.

References

Alzheimer's-Association. "2018 Alzheimer's disease facts and figures." *Alzheimer's and Dementia*, 2018: 367-429.

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., & Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and language*, 67(3), 202-227.

Associazione Gruppo Anchise.
<http://www.formalzheimert.it/>.

P. Bhatia, S. Lin, R. Gangadharaiah, B. Wallace, I. Shafran, C. Shivade, N. Du, and M. Diab, editors. Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, Online, July 2020. Association for Computational Linguistics

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?. *Frontiers in aging neuroscience*, 10, 369.

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1), 71-91.

Bosco, C., Montemagni, S., Simi, M. (2013). Converting Italian Treebanks: Towards an Italian Stanford Treebanks.

Bosco, C., Dell'Orletta, F., Montemagni, S., Sanguinetti, M., & Simi, M. (2014). The EVALITA 2014 dependency parsing task. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian* (pp. 1-8). Pisa University Press.

de la Fuente Garcia, S., Haider, F., & Luz, S. (2020). Cross-corpus Feature Learning between Spontaneous Monologue and Dialogue for Automatic Classification of Alzheimer's Dementia Speech. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 5851-5855). IEEE.

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: Etaslibri.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3), 189-198.

S. Karlekar, T. Niu, and M. Bansal. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 701–707, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Kempler, D. (1991). Language Changes in Dementia of Alzheimer Type. In *Dementia and Communication*, by Rosemary Lubinsky, 98-114. Philadelphia: B.C. Decker, Inc.

Kempler, D., & Goral, M. (2008). Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28, 73.

W. Kong, H. Jang, G. Carenini, and T. Field. A neural model for predicting dementia from language. volume 106 of Proceedings of Machine Learning Research, pages 270–286, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

Lai, G. (2000). Conversazioni con l'Alzheimer. *Prospettive sociali e sanitarie*, 18, 2-5.

Macri, A. (2016). La lingua della demenza di Alzheimer. Analisi linguistica del parlato spontaneo. In *Le lingue della malattia*, 329-424. Milano: Mimesis Edizioni.

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., & Christensen, H. (2019). Dementia

detection using automatic analysis of conversations. *Computer Speech & Language*, 53, 65-79.

Pope, C., & Davis, B. H. (2011). Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1), 143-161.

Qi, P., Dozat, T., Yuhao Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160-170.

Torruella, J.; Capsada, R. (2013). "Lexical Statistics and Tipological Structures: A Measure of Lexical Richness." *Procedia - Social and Behavioral Sciences*, pp. 447-454.

Vigorelli, P. (2018). *Alzheimer, come parlare e comunicare nella vita quotidiana nonostante la malattia*. Milano: Franco Angeli Editore.

Vigorelli, P. (2004). *La conversazione possibile con il malato Alzheimer*. Milano: Franco Angeli Editore.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., ... & Petrov, S. (2018, October). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies* (pp. 1-21).

Gender Bias in Italian Word Embeddings

Davide Biasion, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto

Università degli Studi di Padova

{biasiondav, fabrisal, silvello, sustogia}@dei.unipd.it

Abstract

In this work we study gender bias in Italian word embeddings (WEs), evaluating whether they encode gender stereotypes studied in social psychology or present in the labor market. We find strong associations with gender in job-related WEs. Weaker gender stereotypes are present in other domains where grammatical gender plays a significant role.

1 Introduction

In the literature, the study of gender bias in word embeddings (WEs) is of interest for two main reasons: (i) WEs, as components of automatic decision systems (e.g. job search tools), may contribute to harm some user groups (De-Arteaga et al., 2019); (ii) WEs can be employed as a tool to measure the biases of text corpora (Garg et al., 2018) and systems for automatic text classification or information retrieval (Fabris et al., 2020). In both applications, it is important to isolate the gender-related information in a subspace (Bolukbasi et al., 2016) and subsequently (i) eliminate it via orthogonal projection or (ii) exploit it as a lens to study association of concepts with gender.

A common taxonomy of bias in algorithms concentrates on the types of harm that they may cause (Barocas et al., 2017). **Allocational harms** happen when a limited resource (e.g. jobs) is assigned unfairly to subgroups of a population (e.g. women and men). **Representational harms** arise when groups or individuals are unable to determine their image, which is presented unfavourably or neglected. Autocomplete suggestions in search engines (Noble, 2018; Olteanu et al., 2020) are a

clear example of this situation. Query completion suggestions for “why are italian ...” associate diverse concepts to the country and its inhabitants. Italians contribute very little to these results as they are unlikely to search information about themselves in English.

Italian WEs have been developed (Berardi et al., 2015; Bojanowski et al., 2017) and analyzed (Tripodi and Li Pira, 2017), following seminal work in English; analysis of gender bias has unfortunately lagged behind. Our main contribution is to close this gap, by undertaking a systematic study of gender stereotypes in Italian WEs, adapting established approaches that assess gender bias in English WEs.

2 Related work

Gender stereotypes are representational harms which influence the lives of women and men both descriptively and prescriptively, shaping the qualities, priorities and needs that members of each gender are expected to possess (Ellemers, 2018). In seminal work, Bolukbasi et al. (2016) uncover problematic associations with gender in English WEs. Their approach to identify gender information is adapted to Italian in Section 3.1.1. Caliskan et al. (2017) study the stereotypical association of gender with dichotomies such as career and family, science and arts, following the Implicit Association Test (IAT - Greenwald et al. (1998)). We recall their approach in Section 3.1.2. WEs of jobs have also been analyzed extensively due to their potential for allocational harms in resume search engines (De-Arteaga et al., 2019; Prost et al., 2019) and representational harms (Caliskan et al., 2017), e.g. in general purpose search engines (Kay et al., 2015). Grammatical gender has been found to interact strongly with semantic gender in Spanish and German (McCurdy and Serbetci, 2020), showing that the study of bias in gendered languages poses an additional challenge. We adapt

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

these experiments to the Italian language, detailing our approach in Sections 3-5.

3 Gender in Italian WEs

3.1 Identifying gender information

3.1.1 Gender score

To identify a vectorial subspace which encodes information about gender, we follow Bolukbasi et al. (2016) by building a list of gender definitional pairs: [*lui* (*he*), *lei* (*she*)], [*uomo* (*man*), *donna* (*woman*)], [*padre* (*father*), *madre* (*mother*)], [*marito* (*husband*), *moglie* (*wife*)], [*fratello* (*brother*), *sorella* (*sister*)], [*maschio* (*male*), *femmina* (*female*)].

These pairs are built so that the second word denotes a female entity and the first word is, semantically, its male counterpart. Moreover, given we are interested in capturing *semantic* information about gender, while avoiding entanglement with *grammatical* gender, we ensure that the words in a pair do not derive from the same root via inflection. An example of pair discarded due to this criterion is [*figlio* (*son*), *figlia* (*daughter*)].

Principal Component Analysis. We perform a Principal Component Analysis (PCA) on the six vector differences resulting from each gender definitional pair. The first eigenvalue dominates the remaining ones, with the first PC explaining 57% of variance. We normalize the first PC and consider it the main *gender direction*, denoted by \mathbf{g}_{PCA} .

This is an established procedure to isolate the direction that captures most of the information about gender (Bolukbasi et al., 2016; Ethayarajh et al., 2019). In other words, by finding the direction that best fits the six vector differences ($\vec{lui} - \vec{lei}$, $\vec{uomo} - \vec{donna}$, ...), we aim to obtain a direction that summarizes them.

Vector differences. To evaluate the robustness of this approach and highlight potential anomalies, we also consider each vector difference on its own, defining six unit length *gender directions* $\mathbf{g}_{\text{diff}_i}$:

$$\begin{aligned} \mathbf{g}_{\text{diff}_0} &= \vec{lui} - \vec{lei} & \mathbf{g}_{\text{diff}_3} &= \vec{marito} - \vec{moglie} \\ \mathbf{g}_{\text{diff}_1} &= \vec{uomo} - \vec{donna} & \mathbf{g}_{\text{diff}_4} &= \vec{fratello} - \vec{sorella} \\ \mathbf{g}_{\text{diff}_2} &= \vec{padre} - \vec{madre} & \mathbf{g}_{\text{diff}_5} &= \vec{maschio} - \vec{femmina} \end{aligned}$$

Gender score computation. Given a word w , let us indicate with \mathbf{w} its corresponding word vector. Let us consider any of the gender directions \mathbf{g} defined above. We call *gender score* the normalized

projection of \mathbf{w} onto the direction \mathbf{g} , defined as

$$s_{\mathbf{g}}(w) = \mathbf{w} \cdot \mathbf{g} / (|\mathbf{w}| |\mathbf{g}|). \quad (1)$$

This scalar captures associations of w along gendered lines. Informally, a highly positive value means that \mathbf{w} is closer to the male terms of the pairs than to the female ones, while a strongly negative value entails the opposite.

3.1.2 WEAT

The Implicit Association Test (IAT - Greenwald et al. (1998)) is an assessment developed in cognitive psychology to measure subconscious associations between categories and concepts. It is commonly employed to assess implicit stereotypes in people. The *Word Embedding Association Test* (WEAT - Caliskan et al. (2017)) is a technique inspired by the IAT to measure associations between concepts in WEs. Let X and Y be two equal-sized sets of target words and A and B two sets of attribute words, e.g., $X = \{\text{programmer}, \text{engineer}\}$, $Y = \{\text{nurse}, \text{teacher}\}$, $A = \{\text{man}, \text{male}\}$, $B = \{\text{woman}, \text{female}\}$. Let $\cos(\mathbf{a}, \mathbf{b})$ be the cosine similarity between the word vectors \mathbf{a} and \mathbf{b} . The differential association of a word w (taken from X or Y) with the attribute sets A and B is measured as

$$c(w, A, B) = \text{mean}_{a \in A} \cos(\mathbf{w}, \mathbf{a}) - \text{mean}_{b \in B} \cos(\mathbf{w}, \mathbf{b}). \quad (2)$$

The normalized differential association between targets and attributes is defined as

$$d = \frac{\text{mean}_{x \in X} c(x, A, B) - \text{mean}_{y \in Y} c(y, A, B)}{\text{std-dev}_{w \in X \cup Y} c(w, A, B)}. \quad (3)$$

This is called effect size in statistics, and summarizes how different the quantity $c(w, A, B)$ is, when evaluated on elements of target set X as opposed to target set Y . It is computed as a difference of means within each set, divided by overall standard deviation.

Gender score and WEAT. It is worth noting that, when $|A| = |B| = 1$, WEAT is almost equivalent to the gender score defined in Section 3.1.1. Let $A = \{a_0\}$ and $B = \{b_0\}$ be the sets of attribute words. Since we are using normalized vectors and the distributive property holds for the dot product, then

$$\begin{aligned} c(w, A, B) &= \cos(\mathbf{w}, \mathbf{a}_0) - \cos(\mathbf{w}, \mathbf{b}_0) \\ &= \mathbf{w} \cdot (\mathbf{a}_0 - \mathbf{b}_0) = \mathbf{w} \cdot \mathbf{g} = s_{\mathbf{g}}(w). \end{aligned} \quad (4)$$

3.2 Handling grammatical gender

Italian is a gendered language, wherein grammatical gender is assigned to all nouns. Within a sentence, each word is surrounded by other words of agreeing grammatical gender. This phenomenon, called *grammatical gender agreement*, in conjunction with the *distributional hypothesis* (Harris, 1954), plays an important role when training WEs. Due to these properties, words that share the same grammatical gender to have similar vector representations. Accordingly, grammatical and semantic gender become entangled in WEs (McCurdy and Serbetci, 2020; Gonen et al., 2019). As a consequence, when computing the gender score, we tend to obtain positive values for (grammatically) masculine terms and a negative score for feminine ones, making stereotypical association more noisy and harder to study.

Mean gender score. To compute the gender score (Equation 1) for gendered words that have both a feminine and a masculine version, we propose the following approach. Let us indicate with w_f and w_m the feminine and masculine version of a gendered word w . We define their gender score as

$$s_{\text{mean}_g}(w) = (s_g(w_f) + s_g(w_m))/2. \quad (5)$$

Averaging the masculine and feminine version with equal weights corresponds to giving both versions of the word the same importance. Different approaches, based for instance on word frequency, may be applicable in other contexts.

Orthogonal projection. Some nouns cannot be inflected into the opposite grammatical gender, making the above approach impractical. An example is *ufficio* (*office*). In this context, we propose to mitigate the effect of grammatical gender by re-embedding every word through an orthogonal projection. We build a list of 138 inflected word pairs. Each pair consists of the feminine and masculine inflections of the same root, such as *cara* and *caro* (*dear*), which only differ in grammatical gender. We take the embedding of both words in a pair and compute their difference.

We perform PCA on these vector differences. The resulting PCs span a subspace U that contains most of the variance due to grammatical gender. To reduce the influence of grammatical gender, we re-embed vectors by projecting them on the orthogonal complement of U . In other words, given a word embedding \mathbf{w} , let us call $\text{proj}_U \mathbf{w}$ its orthogonal projection onto the “grammatical

gender subspace” U . We propose re-embedding every word vector \mathbf{w} to

$$\mathbf{w}' = \frac{\mathbf{w} - \text{proj}_U \mathbf{w}}{\|\mathbf{w} - \text{proj}_U \mathbf{w}\|}. \quad (6)$$

By means of this procedure, we obtain a new set of WEs. By construction, in this new embedding space, grammatical gender should have a lower influence on the geometry of word vectors.

4 Datasets and embeddings

To study gender bias we use WEs trained on two different datasets for the Italian language, both made available by FastText (Bojanowski et al., 2017; Grave et al., 2018). The first group of vector representations, which we refer to as *wiki*, consists of word vectors trained on a 2016 Wikipedia dump (Bojanowski et al., 2017).¹ The second group of word vectors (labeled *wiki-cc*) was trained on the May 2017 Common Crawl² and the Wikipedia dump from September 11, 2017 (Grave et al., 2018).

We compare our results from the analyses on Italian WEs with results on their English counterpart. To this end, we also download two sets of FastText WEs trained on the English version of the same corpora, i.e. the English counterparts of *wiki* and *wiki-cc*. Given Wikipedia is a more curated source, we expect to find weaker stereotypes in *wiki* than in *wiki-cc* for both languages. As a pre-processing step we normalize every word vector to unit length.

Census data about the labor market is required to analyse the correlation between the gender gap in professions and the gender score of the respective WEs. The statistics on the American occupation and gender representation are readily available (Census Bureau, 2019). For their Italian counterpart, we retrieve statistics about occupation participation from several institutions, including professional chambers (Comitato Unitario Permanente degli Ordini e Collegi Professionali, Confprofessioni) and academic databases (AlmaLaurea).³

Finally, in order to perform the Word Embedding Association Tests (WEAT), we need sets of

¹The authors provide no information about which Wikipedia dump they use.

²Common Crawl is a corpus of web pages, aimed at representing “a copy of the internet” at a given time. The authors train WEs on pages written in Italian, exploiting language identification as preliminary step for their pipeline.

³The detailed list of sources is available upon request.

target and attribute words in Italian. The sets of target words for the gender-science WEAT (Section 5.2) are derived from the Italian version of IAT;⁴ those for the gender-career WEAT (Section 5.3) were unavailable and have been translated by the authors of this work from the original IAT (Greenwald et al., 1998).

5 Experiments

5.1 Occupations

This experiment investigates gender representation for different jobs in Italy and their association with gender-related information in WEs, following studies on the English language (De-Arteaga et al., 2019; Garg et al., 2018; Prost et al., 2019). For each occupation, we compute its gender score using the different gender directions defined in Section 3.1.1, namely \mathbf{g}_{PCA} and $\mathbf{g}_{\text{diff}_i}$, $i \in \{0 \dots 5\}$. We calculate the plain gender score for the ungendered occupations (Equation 1) and the mean gender score for occupations characterized by grammatical gender (Equation 5).

We compute Pearson’s correlation r between the gender scores and the percentage of women employed in each profession. The same analyses are carried out on English WEs, restricting them to the same set of occupations considered in Italian. Results are summarized in Table 1 and Figure 1, showing that Italian WEs consistently capture information about different gender representation in jobs. Informally, this means that ordering jobs by percentage of women and by projection on a gender direction yields similar results. The right pane of Figure 1 demonstrates the significant effect of grammatical gender.

5.2 Science and Arts

In this WEAT, the sets of target words for Science and Arts, taken from the Italian version of the IAT, are: $X = \{\text{biologia (biology), fisica (physics), chimica (chemistry), matematica (mathematics), geologia (geology), astronomia (astronomy), ingegneria (engineering)}\}$, $Y = \{\text{filosofia (philosophy), umanesimo (humanism), arte (arts), letteratura (literature), italiano (italian), musica (music), storia (history)}\}$. The sets of male and female attributes are taken from the gender definitional pairs (Section 3.1.1): $A = \{\text{lui, uomo, padre, marito, fratello, maschio}\}$, $B = \{\text{lei, donna, madre,$

⁴<https://implicit.harvard.edu/implicit/italy/takeatest.html>

	wiki-cc	wiki
IT	$r(p)$	$r(p)$
\mathbf{g}_{PCA}	-0.634 (1.3×10^{-4})***	-0.589 (4.9×10^{-4})***
$\mathbf{g}_{\text{diff}_0}$	-0.664 (4.7×10^{-5})***	-0.490 (5.1×10^{-3})***
$\mathbf{g}_{\text{diff}_1}$	-0.594 (4.3×10^{-4})***	-0.528 (2.3×10^{-3})***
$\mathbf{g}_{\text{diff}_2}$	-0.575 (7.1×10^{-4})***	-0.537 (1.8×10^{-3})***
$\mathbf{g}_{\text{diff}_3}$	-0.401 (2.5×10^{-2})**	-0.160 (3.9×10^{-1})
$\mathbf{g}_{\text{diff}_4}$	-0.658 (5.7×10^{-5})***	-0.599 (3.8×10^{-4})***
$\mathbf{g}_{\text{diff}_5}$	-0.358 (4.8×10^{-2})**	-0.205 (2.7×10^{-1})
EN	$r(p)$	$r(p)$
\mathbf{g}_{PCA}	-0.830 (2.0×10^{-6})***	-0.707 (2.3×10^{-4})***

Table 1: Results of the Occupation analysis. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

moglie, sorella, femmina}. We compute the effect size d and the p -value using the whole attribute sets A and B , and label this analysis “all”. Moreover, we also perform the WEAT test over single word pairs, e.g. $A = \{\text{lui}\}$, $B = \{\text{lei}\}$. Results are reported in Table 2. We find no stereotypical association in the expected direction. We hypothesize that this is due to the feminine grammatical gender of all science-related target words, deferring a more detailed analysis to Section 5.4.

	wiki-cc	wiki
IT	$d(p)$	$d(p)$
all	-0.172 (6.3×10^{-1})	-0.140 (5.9×10^{-1})
$\mathbf{g}_{\text{diff}_0}$	-0.464 (7.9×10^{-1})	-0.396 (7.5×10^{-1})
$\mathbf{g}_{\text{diff}_1}$	-0.016 (5.1×10^{-1})	-0.064 (5.4×10^{-1})
$\mathbf{g}_{\text{diff}_2}$	-0.408 (7.5×10^{-1})	-0.152 (6.1×10^{-1})
$\mathbf{g}_{\text{diff}_3}$	-0.002 (5.0×10^{-1})	0.271 (3.3×10^{-1})
$\mathbf{g}_{\text{diff}_4}$	-0.127 (6.0×10^{-1})	-0.174 (6.2×10^{-1})
$\mathbf{g}_{\text{diff}_5}$	-0.144 (6.1×10^{-1})	-0.195 (6.3×10^{-1})
EN	$d(p)$	$d(p)$
all	1.420 (1.5×10^{-3})***	1.304 (3.2×10^{-3})***

Table 2: Results of the Science and Arts WEAT. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

5.3 Career and Family

In essence, the Career and Family WEAT is very similar to the Science and Arts WEAT; the only difference is in the sets of target words. The target sets are translated into Italian from the original English IAT as follows: $X = \{\text{esecutivo (executive), management (management), professionale (professional), azienda (corporation), stipendio (salary), ufficio (office)}\}$, $Y = \{\text{casa (home),$

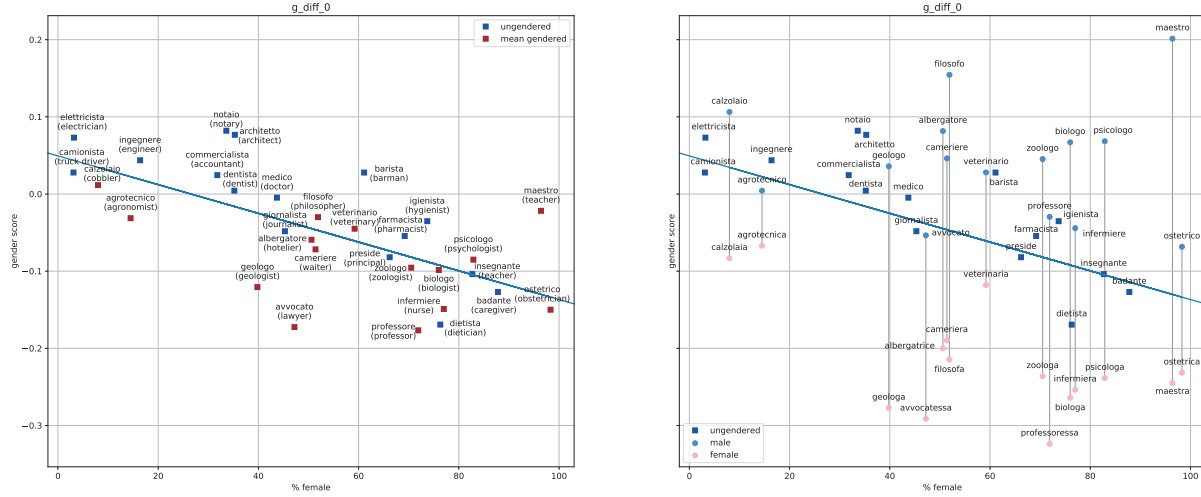


Figure 1: (Left) Gender score of occupations g_{diff_0} , on the y axis, vs percentage of women in that occupation, represented on the x axis. (Right) Same variables, including both feminine and masculine versions of gendered jobs. Translations, omitted for readability, can be found in the left pane.

genitori (parents), *bambini* (children), *famiglia* (family), *cugini* (cousins), *matrimonio* (marriage), *nozze* (wedding), *parenti* (relatives)}. Results are summarized in Table 3. Stereotypical associations for *wiki-cc* WEs are present but weak, whereas they are more significant for *wiki*.

	wiki-cc	wiki
IT	$d(p)$	$d(p)$
all	0.838 (5.3×10^{-2})*	1.351 (2.7×10^{-3} ***)
g_{diff_0}	0.457 (2.0×10^{-1})	1.172 (8.7×10^{-3} ***)
g_{diff_1}	1.265 (4.7×10^{-3} ***)	1.512 (5.4×10^{-4} ***)
g_{diff_2}	0.614 (1.2×10^{-1})	1.181 (8.3×10^{-3} ***)
g_{diff_3}	0.299 (2.9×10^{-1})	0.876 (4.4×10^{-2} **)
g_{diff_4}	0.952 (3.2×10^{-2} **)	0.898 (4.2×10^{-2} **)
g_{diff_5}	0.713 (8.7×10^{-2} *)	-0.566 (8.6×10^{-1})
EN	$d(p)$	$d(p)$
all	1.879 (0.0×10^0 ***)	1.568 (2.3×10^{-4} ***)

Table 3: Results of the Career and Family WEAT. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

5.4 Mitigating the effect of grammatical gender

In this section we quantify the extent to which the semantic gender information (Section 3.1) is influenced by grammatical gender, and test one approach designed to mitigate its influence (Section 3.2).

The dataset used in the experiment about job-related WEs (Section 5.1) is suitable for this analysis, as it consists of words which have (i) a *se-*

mantic association with gender, as measured objectively by the percentage of women in each profession and (ii) a *grammatical* association with gender, as half of those words admit a feminine and a masculine version.

We measure the relative strength of semantic and grammatical associations in the proposed gender directions as follows. Let us denote by S_g , the set of job-related words which admit a feminine and a masculine version and by $\Delta_w = s_g(w_m) - s_g(w_f)$ the difference in their gender scores.⁵ We compute the average influence of grammatical gender on direction g (based on set S_g) as

$$\Delta_g = \frac{1}{|S_g|} \sum_{w \in S_g} \Delta_w. \quad (7)$$

Visually, this corresponds to the average (signed) length of the vertical lines, in the right pane of Figure 1, connecting the feminine and masculine version of a job-related word.

Moreover, let us denote by S the complete set of job-related words w_j and by x_j the percentage of women in job w_j . Let us indicate with $s_g(w_j)$ the respective gender score, computed according to Equation 1 or 5, depending on whether w_j admits different masculine and feminine inflections. We define $\max_S(x)$ ($\min_S(x)$) as the maximum (minimum) percentage of women in a job from set

⁵For the sake of brevity, we concentrate on g_{PCA} ; the remaining gender directions (g_{diff_i}) yield similar results.

\mathcal{S} . Furthermore, let us call m the angular coefficient computed by (linearly) regressing $s_g(w_j)$ onto x_j over set \mathcal{S} . We compute the full-scale influence of semantic gender on direction \mathbf{g} (based on set \mathcal{S}) as

$$\Delta_s = |m(\max_{\mathcal{S}}(x) - \min_{\mathcal{S}}(x))|. \quad (8)$$

Visually, this corresponds to the vertical component of the blue regression line in Figure 1, clipped between $\min_{\mathcal{S}}(x)$ and $\max_{\mathcal{S}}(x)$.

Finally, we compute the relative strength of semantic and grammatical associations in the proposed gender direction as the ratio

$$k = \frac{\Delta_g}{\Delta_s}. \quad (9)$$

The first three rows of Table 4 report Δ_g , Δ_s and k for `wiki-cc` (first column) on the job dataset described in Section 4. The second column concentrates on a set of word embeddings derived from `wiki-cc` by removing information about grammatical gender from every word, via Equation 6.⁶ We label this new set of word embeddings `wiki-cc⊥`. In going from `wiki-cc` to `wiki-cc⊥`, Δ_g is reduced by over 40% while Δ_s decreases by less than 10%. This indicates that the orthogonal projection procedure reduces the influence of grammatical gender while retaining semantic information which is present in the original version of the WEs, hence the value of k decreases.

The final three rows of Table 4 report summary statistics for stereotypical associations described in Sections 5.1-5.3. Interestingly, the significance of each association is larger for `wiki-cc⊥` than for `wiki-cc`. In particular, the effect size for the Science-Arts WEAT becomes positive, in accordance with the stereotype. We interpret these results as evidence for the hypothesis that grammatical gender confounds and outweighs stereotypical associations in Italian WEs, in line with prior work on gendered languages (McCurdy and Serbetci, 2020).

6 Discussion

We successfully replicated prior analyses about gender-stereotypical associations in English WEs, finding them to be consistently stronger when computed on WEs trained on a weakly curated

⁶In this experiment, the grammatical gender subspace U is spanned by the first PC.

	wiki-cc	wiki-cc [⊥]
Occupations (\mathbf{g}_{PCA})		
Δ_g	0.41	0.22
Δ_s	0.23	0.20
k	1.79	1.09
$r(p)$	-0.63 (1.3×10^{-4})***	-0.68 (2.2×10^{-5})***
Science & Arts (all)		
$d(p)$	-0.17 (6.3×10^{-1})	0.73 (9.7×10^{-2})*
Career & Family (all)		
$d(p)$	0.84 (5.3×10^{-2})*	1.21 (6.1×10^{-3})***

Table 4: Importance of semantic and grammatical gender before (`wiki-cc`) and after (`wiki-cc⊥`) projecting WEs onto the orthogonal complement of the grammatical gender subspace (Equation 6). Where applicable, statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

corpus. To the best of our knowledge, this is a novel result.

For Italian WEs, the picture is more nuanced and tied to grammatical gender. WEs for occupations, which are ungendered or admit a dual form, are robustly associated with gender along a stereotypical direction. Compared against the other stereotypes analysed in this work, this is the strongest association, confirming results from prior work on English WEs (Fabris et al., 2020). In the Science-Arts WEAT, science-related words are all feminine nouns, departing from the expected stereotypical association. Semantic associations with gender are outweighed by grammatical gender in this WEAT, in accordance with prior work on gendered languages (McCurdy and Serbetci, 2020). Our analysis in Section 5.4 demonstrates the importance of grammatical gender in Italian. On the other hand, the Career-Family WEAT features a more balanced distribution of grammatical gender, resulting in a differential association which is in line with gender stereotypes, especially for `wiki`, less so for `wiki-cc`.

In Italian WEs, we find that `wiki` embeddings contain stronger stereotypical associations than `wiki-cc` embeddings for the Career-Family WEAT. This disconfirms our hypothesis that WEs trained on a less curated corpus (`wiki-cc`) would encode stereotypes more strongly. Finally, we find no consistent property connected to specific gender directions $\mathbf{g}_{\text{diff}_i}$. Across different corpora and stereotypes, the aggregated analyses (la-

belled “all” and \mathbf{g}_{PCA}) provide a reasonable summary of the stereotypical associations encoded in the single gender directions \mathbf{g}_{diff_i} .

7 Conclusion

Overall, we have analyzed gender bias in Italian WEs, adapting existing techniques and gathering data where required. We looked for stereotypical associations with gender-imbalanced professions, Career and Family, Science and Arts, finding significant associations in 2 out of 3. As expected from prior work (Gonen et al., 2019; McCurdy and Serbetci, 2020), grammatical gender is a strong confounder in these analyses.

We draw the following preliminary conclusions: (i) Italian WEs seem to have less potential than their English counterparts to systematically reinforce the tested gender stereotypes, mostly due to grammatical gender. However, (ii) the influence of grammatical gender on WEs may cause different harms. As an example, in the context of job search, masculine is likely to be the default choice for queries of recruiters (*male as norm* - e.g. “psicologo” [psychologist]). Those queries would likely be closer to male candidates’ CVs than equivalent female ones, in some embedded text representations, potentially putting women at a systematic disadvantage. Both points above require further analysis of text retrieval/classification systems based on Italian WEs. Finally, (iii) isolating stereotypical concepts and gendered associations in Italian WEs along a single direction is challenging. The tested WEs show little promise as a reliable measurement tool for gender-stereotypical associations, unless combined with approaches to mitigate the influence of grammatical gender.

References

- Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.
- Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Census Bureau. 2019. Current population survey. Accessed = 2020-02-12.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnamurthy Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19*.
- Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69(1):275–298.
- Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July. Association for Computational Linguistics.
- Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- Hila Gonen, Yova Kementchedjiev, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

- Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM.
- Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.
- Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.
- Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When are search completion suggestions problematic? In *Computer Supported Collaborative Work and Social Computing (CSCW)*. ACM, August. Pre-print, paper accepted to CSCW’20.
- Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.
- Rocco Tripodi and Stefano Li Pira. 2017. Analysis of italian word embeddings. In *CLiC-it*.

Automatic Induction of FrameNet Lexical Units in Italian

Silvia Brambilla[‡], Danilo Croce[†], Fabio Tamburini[‡], Roberto Basili[†]

[†]University of Rome Tor Vergata, [‡]University of Bologna

{silvia.brambilla2, fabio.tamburini}@unibo.it

{croce, basili}@info.uniroma2.it

Abstract

In this paper we investigate the applicability of automatic methods for frame induction to improve the coverage of IFrameNet, a novel lexical resource based on Frame Semantics in Italian. The experimental evaluations show that the adopted methods based on neural word embeddings pave the way for the assisted development of a large scale lexical resource for our language.

1 Introduction

When dealing with large-scale lexical resources, such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), VerbNet (Schuler, 2005) or VerbAtlas (Di Fabio et al., 2019), the semi-automatic association between predicates and lexical items (also known as Lexical Units or LUs) is crucial to improve the coverage of a resource while limiting the costs of its manual annotation. Several approaches to this semi-supervised task exist, as discussed in QasemiZadeh et al. (2019). In particular, Pennacchiotti et al. (2008) exploited distributional models of lexical meaning (Sahlgren, 2006; Croce and Previtali, 2010) to induce new LUs consistently with the Frame Semantics theory (Baker et al., 1998), representing words meaning and semantic frames through geometrical word spaces. As a result, this approach allows to induce new LUs when applied to the English version of FrameNet. However, this is a quite consolidated resource with many existing LUs connected to each semantic predicate, i.e., each frame. The applicability of this method in scenarios where only one or two LUs are available for each frame is still an open issue. At the same

time, since the work of Pennacchiotti et al. (2008), the application of neural approaches to the acquisition of word embeddings (Mikolov et al., 2013; Baroni et al., 2014; Ling et al., 2015) significantly improved in terms both of representation capability and scalability of geometrical models of lexical semantics.

In this paper we thus investigate the applicability of the method proposed in Pennacchiotti et al. (2008) to boost the coverage of a novel and still limited lexical resource based on Frame Semantics in Italian. This resource has been developed within the IFrameNet (IFN) project (Basili et al., 2017), which aims at creating a large coverage FrameNet-like resource for Italian and to come up with a complete dictionary in which every lexical entry¹ is linked to all the frames it can evoke (i.e., the frames for which it is a LU). At this moment, while the resource counts more than 7,700 lexical items associated to more than 1,048 frames, each lexical item is connected, on average, to only 1.3 frames, and it is problematic if considering the high polysemy of Italian words (Casadei, 2014).

The experimental evaluation shows that neural word embeddings enable the effective application of the distributional approach from Pennacchiotti et al. (2008) to improve the coverage of IFN. Moreover, the adopted distributional framework allowed to develop a graphical semantic browser to support annotators while assigning new LUs to frames. This study paves the way to the semi-automatic development of IFN and investigates about the applicability of neural word embeddings to the incremental semi-automatic LU induction process.

2 Related Work

In the development of FrameNet and FrameNet-like resources for new languages, one important

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Where with the term *lexical entry* we denote a lemma, with its Part of Speech tag, that activates at least one LU.

task is the creation of a large-scale dictionary, in order to guarantee an effective application in semantic analyses or NLP tasks. In fact, the limited coverage of FrameNet has been addressed as one of the main reason of failures (Pennacchiotti et al., 2008; Pavlick et al., 2015). For these reasons and given the high costs of manual annotation, both in terms of time and resources (i.e., human annotators), the automatic (or semi-automatic) expansion of the dictionary for FrameNet and FrameNet-like resources has received attention during the years. Several methods to support the population of frames in FrameNet (Baker et al., 2007; Pavlick et al., 2015; Ustalov et al., 2018; QasemiZadeh et al., 2019; Anwar et al., 2019; Arefyev et al., 2019; Yong and Torrent, 2020), and FrameNet-like resources (Johansson and Nugues, 2007; Tonelli et al., 2009; Tonelli, 2010; Johansson, 2014; Hayoun and Elhadad, 2016) with new Lexical Units have been widely investigated. Some of the methodologies proposed in order to automatically expand FrameNet have exploited the alignment between WordNet and FrameNet data (Johansson and Nugues, 2007; Pennacchiotti et al., 2008; Ferrández et al., 2010). Another strategy is the one adopted by Pavlick et al. (2015) where the scholars enlarge FrameNet coverage using automatic paraphrase. The majority of the works dealing with automatic frame induction, however, exploits distributional methods, for example the work on which this research relies the most, i.e., the work of Pennacchiotti et al. (2008) or some of the most recent works such as the ones of Ustalov et al. (2018), Arefyev et al. (2019) and Yong and Torrent (2020). Ustalov et al. (2018), for example, model the frame induction problem as a tri-clustering problem and use dependency triples automatically extracted from a Web-scale corpus. Arefyev et al. (2019) propose to combine dense representations from hidden layers of a masked language model with sparse representations based on substitutes for the target word in the context for the creation of vector representations.

3 IFrameNet status

The IFrameNet project (Basili et al., 2017), relied, as a starting point, on the achievements of previous researches on the development of Italian resources annotated according to Frame Semantics (Tonelli and Pianta, 2009; DeCao et al., 2010), i.e., a set of automatically induced LUs that were covering

554 frames of the 1,224 frames in FrameNet.

Since the beginning, our main objective has been to improve the coverage of the resource in terms of annotated frames, increasing the number of the LUs and the number of annotated sentences representing each predicate. Starting from the results achieved in 2017, we enlarged the dictionary and provided an initial set of LUs for those frames without any annotation. We also revised the whole dictionary and expunged the LUs whose lemma had low frequency² in CORIS (Corpus di Italiano Scritto) (Rossini Favretti et al., 2002). Since CORIS is a large-scale and general-purpose Italian corpus (without biases to any domain), we speculate that not represented LUs can hardly characterize a frame in Italian. Moreover, we worked on the frame annotation of sample sentences taken from the CORIS corpus. We relied on CORIS because it is domain independent and suitable to represent the generic notion of frames. Currently, the resource contains:

- **7,776 lexical entries** of which: 1,130 adjectives, 4,309 nouns and 2,337 verbs;
- **10,379 LUs** (nouns, verbs and adjectives) validated in terms of pairs of lexical entries and evoked frame(s);
- **1,048 frames with at least one LU** among which **743 frames are represented with at least one sentence**. Among the 176 frames that still do not have any LU in their dictionary, 134 are marked as Non-Lexical in FrameNet, 12 do not have any LU in FrameNet, but are not explicitly marked as Non-Lexical, 18 are not represented in FrameNet by any noun, verb or adjective and finally, for just 8 frames, it was difficult to find LUs in Italian (e.g. IMPROVISED_EXPLOSIVE_DEVICE or SHORT_SELLING);
- **5,208 sentences** annotated and validated with at least one LU;
- an average of **9.9 LUs assigned to each frame**;
- an average of **1.3 frames associated to each LU**. Among the existing LUs, 5,960 are assigned to only one frame. Given that Italian language is highly polysemous, it is probable that many LUs evoke more than one frame. This work aims at reducing this limitation.

²Less than 20 occurrences in the corpus.

4 Automatic Frame Induction

For the Frame Induction we rely on distributional methods as in Pennacchiotti et al. (2008), described hereafter.

Distributional representation. As a first step, we obtain a distributional representation of the CORIS corpus and represent in the wordspace each LU as a vector \vec{l} . We investigated three slightly different approaches for the acquisition of the wordspaces: the Continuous Bag-of-Words model (CBOW), the Skip-gram model (Mikolov et al., 2013) and the Structured Skip-gram (sskip-gram) model (Ling et al., 2015). The sskip-gram is a modification of the skip-gram model, sensitive to the positioning of the words and, thus, more suitable for capturing syntactic properties of the words (Ling et al., 2015). Our hypothesis is that this last model would be more suitable for capturing LUs frame properties since syntax is, in general, in agreement with semantic arguments (i.e., Frame Elements, FEs) and their order.

“Framehood” representation. As a second step, we exploit the obtained embeddings to represent the meaning of frames. We assume that a frame f can be described by the set of its LUs $l \in F$ and that LUs vectors \vec{l} can be thus used to acquire a distributional representation for each frame. In a nutshell, for each frame we: (i) select all the LUs of its dictionary, (ii) apply to LUs vectors \vec{l} a clustering algorithm. A frame will be then represented as a set of clusters: given that each frame can have various nuances and that it can be representative of non overlapping senses, sparse in the semantic space, we represent it through its “clusters of senses”. This captures, in the semantic space, the possible “framehood” distributions, as dense regions of LUs. In this work, we applied standard K-means (Hartigan and Wong, 1979), so that each frame is represented as a set of k clusters. For each frame k is empirically set to the square root of the number of LUs l in that frame: $k = \sqrt{|l|}$, where $|l|$ denotes the count of l per frame. In this way, each f will have k clusters depending on the number of its LUs and the centroid of each cluster will represent the prototype for a subset of the senses of a frame.

New LU induction. Once obtained the distributional representations for frames and LUs, the third step involves the automatic induction of frames given a candidate lexical item. For each

POS	1	2	5
<i>a</i>	295	207	65
<i>n</i>	631	463	250
<i>v</i>	675	514	245
<i>a-n-v</i>	1,041	916	511

Table 1: Number of frames considered according to different filtering policies. In column the threshold applied to the number of required LUs.

candidate predicate word, we computed the distance between its vector and the sets of clusters representing the frames. The “nearest” clusters will be the ones containing a set of LUs more closely related to the input lexical item, so that the corresponding frames will be suggested as its evoking frames.

5 Experimental Evaluation

In order to assess the quality of the proposed method, we evaluate its capability in rediscovering the frames manually associated to a lexical item. We apply a leave-one-out schema: for each candidate lexical item, we eliminate it from the dictionary and query the model to “suggest” up to 10 frames. In practice, we rebuild the clusters and then compute the distance between the lexical item’s vector and the set of clusters representing all frames. Then, we compare the suggested frames with the frames that were originally linked to the LU. As in Pennacchiotti et al. (2008), we compute Accuracy as the fraction of LUs that are correctly re-assigned to the original frame. Accuracy is computed at different levels b : a LU is correctly assigned if one of its gold standard frames appears among the best- b frames ranked by the model. In fact, as LUs can have more than one correct frame, we deem as “correct” an assignment for which at least one of the correct frames is among the best- b .

The model is evaluated by sampling the test bed according two dimensions, as reported in Table 1. First, we considered the Part-of-Speech (POS) of the LUs (i.e., rows in Table 1). In fact, lexical items having different POS are generally projected in different sub-spaces within word spaces. We thus evaluate the model considering separately LUs and frames containing adjectives (*a*), nouns (*n*) or verbs (*v*). For the sake of completeness, we also evaluated the model without any selection by POS (row *a-n-v*). When a frame does not contain any LU represented in the wordspace with a required POS, it is discarded during the evaluation: as an example, the actual dictionary contains 631

POS	th	$b-1$	$b-2$	$b-3$	$b-4$	$b-5$	$b-6$	$b-7$	$b-8$	$b-9$	$b-10$
<i>a</i>	1	32%	41%	47%	50%	52%	53%	55%	56%	57%	58%
	2	41%	54%	62%	65%	68%	70%	72%	74%	75%	76%
	5	60%	75%	83%	87%	89%	90%	91%	93%	94%	95%
<i>n</i>	1	42%	53%	59%	62%	65%	66%	68%	69%	70%	71%
	2	48%	61%	67%	71%	73%	75%	76%	78%	79%	80%
	5	59%	72%	79%	82%	85%	87%	88%	89%	90%	91%
<i>v</i>	1	25%	35%	41%	44%	47%	49%	50%	52%	53%	54%
	2	32%	43%	50%	53%	57%	59%	61%	62%	64%	65%
	5	42%	55%	63%	69%	72%	74%	76%	78%	79%	81%
<i>a-n-v</i>	1	36%	47%	53%	56%	59%	61%	62%	64%	65%	66%
	2	41%	52%	58%	63%	65%	67%	68%	70%	71%	72%
	5	49%	61%	68%	73%	75%	78%	79%	81%	82%	83%

Table 2: Accuracy on LU induction according to the best- b ranking and split according to number and POS tag of LUs.

frames containing at least one noun.

Then, we filtered frames by applying a threshold to the number of LUs a frame should be connected to, in order to be considered (columns in Table 1), as it follows: first, we considered all frames containing at least one LU whose lemma occurred at least 20 times in CORIS, without applying any other restriction (column 1); then we filtered frames with at least 2 valid LUs³ (column 2); finally we filtered frames with at least 5 valid LUs (column 5). Both filter policies can be combined and the stricter these policies are, the lower the number of frames considered in the evaluation. As a consequence, the Accuracy baseline of a model which randomly assigns LUs to frames depends on the number of selected frames: when no filter is applied (row $a-n-v$ and column 1) a random assignment would achieve $0.09\% = \frac{1}{1,041}$ of Accuracy, or $0.4\% = \frac{1}{250}$ when only frames containing at least 5 nouns are selected.

Table 2 reports the experimental results of a model derived using a `sskip-gram` model (Ling et al., 2015)⁴. If we consider the performance over only nouns (n) we see that, when a reasonable threshold is set (row $th = 2$), in 48% of cases in first position we find one of the original frames evoked by the noun under analysis (column $b-1$). If we consider the first two frames proposed by the system ($b-2$) the Accuracy rises up to 61% and it keeps increasing as we consider more frames. It is impressive if considering that the corresponding random baseline is $0.2\% = \frac{1}{463}$ and $0.4\% = \frac{2}{463}$. If we jointly consider nouns, verbs and adjectives

($a-n-v$) the performance is slightly lower: for example, with the same threshold $th = 2$ and considering only two suggested frames ($b-2$) the Accuracy is 61%. It means that, on average, the model capability of assigning LUs (ignoring their POS) to frames is slightly lower. This is confirmed by the general drop obtained when only verbs or adjectives are considered: for verbs, considering only the best suggestion ($b-1$) we measured 25%, if we don’t apply any threshold, to 32%, if we consider $th = 2$, to 42% if we consider $th = 5$. This is mainly due to higher polysemy characterizing verbs and adjectives with respect to nouns (Casadei, 2014). Anyway, this result is straightforward if considering that for verbs, the baseline in the setting $th = 2$ and $b = 1$ corresponds to $0.2\% = \frac{1}{514}$.

Discussion. It is worth noting that our dictionary is largely incomplete and thus some of those counted as “incorrect assignments” are instead frames that are evoked by the LU under analysis and that *should be added* to the dictionary. Moreover, we can see that many of the $b-10$ frames are often related at different degrees with the lexical entry under analysis and with the frames for which it is a LU.

For example, when considering the lexical entry “*impiccare.v*” (*hang.v*) the model does not retrieve among the $b-10$ suggestions the only “correct” frame, i.e., the frame EXECUTION. Anyway, the closest frame identified is the frame KILLING that not only is linked with EXECUTION with an Inheritance relation, but also appears to be evoked by “*impiccare.v*”. Again, the system is not able to re-assign the lexical entries “*innalzarsi.v*” (*raise.v* and *rise.v*), “*innocenza.n*” (*innocence.n*) and “*radiazione.n*” (*radiation.n* or *expulsion.n*). Anyway, in the $b-10$

³This threshold also overcomes the intrinsic limitation of the leave-one-out schema; when considering frames with only one LU, it becomes impossible to spot the original frame in the test data because it will not be represented by any LU.

⁴This method outperformed the CBOW and `skip-gram`, not reported here for lack of space.

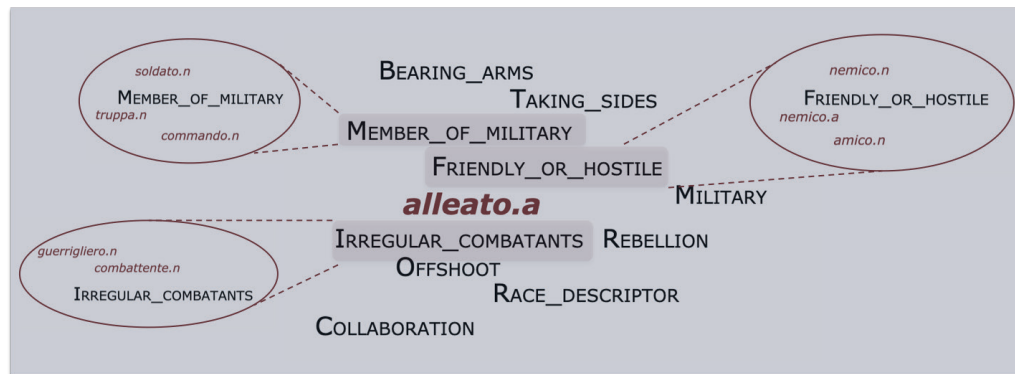


Figure 1: An example of the IFrameNet Navigator for the LU *alleato.a*

of “*innalzarsi.v*” appears in fourth position the frame `CHANGE_POSITION_ON_A_SCALE` that can be evoked by “*innalzarsi.v*” in sentences such as “*La marea si innalzava*” (*The tide was rising*) and in the *b* – 10 of “*innocenza.n*” appears, in first position, the frame `CANDIDNESS` that is evoked by this LU in sentences such as “*Lei rispose con innocenza*” (*She answered genuinely*). The term “*radiazione.n*” is present in the dictionary only with the meaning *expulsion.n* and it is linked only to `EXCLUDE_MEMBER`. Nevertheless, the system proposes the frame `NUCLEAR_PROCESS` in first position and retrieves one correct meaning of a LU like “*radiation.n*”. For “*alleato.a*” (*ally.n*, also shown in Figure 1) the system proposes a “correct” frame in ninth position. Anyway, we find in second position the frame `MEMBER_OF_MILITARY` that can be plausibly evoked. Moreover the LU “*agnello.n*” (*lamb.n*) evokes in the dictionary only the frame `FOOD`; anyway, as correctly suggested by the system, it is also LU of the frame `ANIMALS`. Moreover for “*agnello.n*” the system proposes also, in sixth position, `PEOPLE_BY_MORALITY` that recalls the idea of innocence and righteousness that represents (at least for the Italian language) a metaphorical extension of the meaning of “*lamb.n*”, strongly influenced by the religious image of the lamb.

In some other cases, the system suggests relations between frames. For example, if we consider the lexical entry “*identico.a*” (*identical.a* from `IDENTICALITY`) we see in the best-10 frames that the system proposes frames such as `SIMILARITY` (first position) or `DIVERSITY` (seventh position). If we look at the frame-to-frame relations in FrameNet, we see that `IDENTICALITY` and `SIMILARITY` or `IDENTICALITY` and `DIVERSITY` are not directly connected even if they appear, at a close analysis, strictly related.

6 IFrameNet Navigator

In order to make the model valuable for the annotators, we also developed a Graphical User Interface, called IFrameNet Navigator. It allows querying and navigating the geometrical representation of semantic phenomena as it displays, for each lexical entry in the dictionary, the best-10 frames. These can be also selected to browse the set of LUs assigned to the cluster underlying the frame, as shown in Figure 1. Finally, each LU can be selected to browse the list of corresponding annotated sentences.

The objectives of the Navigator are: (i) to support the analysis of the currently modeled lexical entries (and the corresponding LUs); (ii) to support the validation of the current sentence classification; (iii) the mining of the CORIS corpus for improving the semantic coverage of the resource for the Italian language; (iv) in perspective, to offer support towards crowd sourcing.

This tool will be publicly released to trigger collaborative validation and annotation as an extension of the IFrameNet and the CORIS resources.

7 Conclusions and Research Perspectives

In this work, we presented the actual state of the IFrameNet project, which aims at developing a large-scale lexical resource based on Frame Semantics in Italian. Moreover, we investigated the applicability of a method for the automatic Induction of FrameNet Lexical Units to improve the coverage of the actual resource, in terms of number of frames assigned to the almost 8,000 existing lexical entries.

With respect to previous work, i.e., Pennacchiotti et al. (2008) we empirically demonstrate the beneficial impact of neural word embeddings in the overall workflow in Italian. The robustness

of the adopted model is confirmed also when applied to a resource with a limited average number of frames associated to Lexical Units. The experimental evaluations in many cases showed the valuable support of the method in discovering new Lexical Units by suggesting novel evoked frames. Moreover, the error analysis suggested that most of the “discarded” frames still entertain various kinds of relationships with the “correct” ones as defined in FrameNet, such as Inheritance or Usage. In some cases, it also highlighted metaphorical meanings that the lexical entries could assume.

As a future work, we will certainly exploit the produced IFrameNet Navigator to extend the current LU Italian dictionary, support the annotation of novel sentences and introduce frame-to-frame relations in Italian. Another path that might worth investigating is the exploitation of dependency-based word embeddings for the distributional representation of LUs and frames. This may be beneficial since dependency-based contexts highlight more functional similarities (Levy and Goldberg, 2014). Finally, we plan to use the derived frame distributions to augment existing contextualized embeddings in support of Frame Induction (Sikos and Padó, 2019) or Semantic Role Labeling (Shi and Lin, 2019) tasks.

References

- Saba Anwar, Dmitry Ustalov, Nikolay Arefyev, Simone Paolo Ponzetto, Chris Biemann, and Alexander Panchenko. 2019. Hhmm at semeval-2019 task 2: unsupervised frame induction using contextualized word embeddings. *arXiv preprint arXiv:1905.01739*.
- Nikolay Arefyev, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko. 2019. Neural granny at semeval-2019 task 2: A combined approach for better modeling of semantic relationships in semantic frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 31–38.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proc. of COLING-ACL*, Montreal, Canada.
- Collin F Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.
- Roberto Basili, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. Developing a large scale framenet for italian: the iframenet experience. *CLiC-it 2017 11-12 December 2017, Rome*, page 59.
- Federica Casadei. 2014. La polisemia nel vocabolario di base dell’italiano. *Lingue e Linguaggi*, 12:35–52.
- Danilo Croce and Daniele Previtali. 2010. Manifold learning for the semi-supervised induction of FrameNet predicates: An empirical investigation. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 7–16, Uppsala, Sweden, July. Association for Computational Linguistics.
- Diego DeCao, Danilo Croce, and Roberto Basili. 2010. Extensive evaluation of a framenet-wordnet mapping resource. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. Verbatlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637.
- Oscar Ferrández, Michael Ellsworth, Rafael Munoz, and Collin F Baker. 2010. Aligning framenet and wordnet based on semantic neighborhoods. In *LREC*, volume 10, pages 310–314.
- J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108.
- Avi Hayoun and Michael Elhadad. 2016. The hebrew framenet project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4341–4347.
- Richard Johansson and Pierre Nugues. 2007. Using wordnet to extend framenet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, pages 27–30.
- Richard Johansson. 2014. Automatic expansion of the swedish framenet lexicon: Comparing and combining lexicon-based and corpus-based methods. *Constructions and Frames*, 6(1):92–113.

- Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.
- Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–413.
- Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of framenet lexical units. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 457–465.
- Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- Rema Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. 2002. Coris/codis: A corpus of written italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world*, pages 27–38.
- Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.
- Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.
- Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.
- Jennifer Sikos and Sebastian Padó. 2019. Frame identification as categorization: Exemplars vs prototypes in embeddingland. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 295–306, Gothenburg, Sweden, May. Association for Computational Linguistics.
- Sara Tonelli and Emanuele Pianta. 2009. Three issues in cross-language frame information transfer. In *Proceedings of the International Conference RANLP-2009*, pages 441–448, Borovets, Bulgaria, September. Association for Computational Linguistics.
- Sara Tonelli, Daniele Pighin, Claudio Giuliano, and Emanuele Pianta. 2009. Semi-automatic development of framenet for italian. In *Proceedings of the FrameNet Workshop and Masterclass, Milano, Italy*.
- Sara Tonelli. 2010. *Semi-automatic techniques for extending the FrameNet lexical database to new languages*. Ph.D. thesis, Università Ca’Foscari Venezia.
- Dmitry Ustalov, Alexander Panchenko, Andrei Kutuzov, Chris Biemann, and Simone Paolo Ponzetto. 2018. Unsupervised semantic frame induction using triclustering. *arXiv preprint arXiv:1805.04715*.
- Zheng Xin Yong and Tiago Timponi Torrent. 2020. Semi-supervised deep embedded clustering with anomaly detection for semantic frame induction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3509–3519.

Detecting Early Onset of Depression from Social Media Text using Learned Confidence Scores

Ana-Maria Bucur

University of Bucharest, Romania
ana-maria.bucur@drd.unibuc.ro

Liviu P. Dinu

University of Bucharest, Romania
ldinu@fmi.unibuc.ro

Abstract

English. Computational research on mental health disorders from written texts covers an interdisciplinary area between natural language processing and psychology. A crucial aspect of this problem is prevention and early diagnosis, as suicide resulted from depression being the second leading cause of death for young adults. In this work, we focus on methods for detecting the early onset of depression from social media texts, in particular from Reddit. To that end, we explore the eRisk 2018 dataset and achieve good results with regard to the state of the art by leveraging topic analysis and learned confidence scores to guide the decision process.¹

1 Introduction

Mental illnesses are a common problem of our modern world. More than one in ten people was living with mental health disorders in 2017 (Ritchie and Roser, 2018), with women being the most affected. These disorders affect people's way of thinking, mood, emotions, behaviour and their relationships with others. Most mental illnesses remain undiagnosed because of the social stigma around them.

Depression is one of the main causes of disability globally², it affects people of all ages. Prevention is used to reduce depression and to save the lives of people at risk of suicide, but prevention is only limited to raising awareness and programs to cultivate positive thinking in case of depression and monitoring people who attempted suicide or self-harm.

With the rise in social media use, more computational efforts are made to detect mental illnesses

such as depression (De Choudhury et al., 2013) and PTSD (Coppersmith et al., 2015), but also to detect misogyny (Anzovino et al., 2018), irony and sarcasm (Khokhlova et al., 2016) from users' texts.

People tend to talk more about their emotions and mental health problems online and to seek support. The sources of mental health cues used for detection are Twitter, Facebook, Reddit and forums (Calvo et al., 2017). Reddit³ is a social media site very similar to forums. It is organized in subreddits with specific topics, some dedicated to mental health problems. The use of throwaway accounts to maintain anonymity promotes disclosure, and users are more likely to share problems they have not discussed with anyone before. The use of these accounts makes it difficult for users to receive more social support because the majority of them are used only for one post (Calvo et al., 2017).

In this work, we choose to tackle the problem of detecting early onset of depression from users' posts on social media, specifically from Reddit. As such, we explore the eRisk 2018 dataset through topic analysis by means of Latent Semantic Indexing (Deerwester et al., 1990) and learned out-of-distribution confidence scores (DeVries and Taylor, 2018). Due to the nature of the dataset, we repurpose the learned confidence score to make a decision on whether to label the user as depressed or non-depressed or to wait for more data, as test chunks were progressively released every week.

2 Related Work

Recent studies for depression detection from text are reviewed by Guntuku et al. (Guntuku et al., 2017). People diagnosed with mental illnesses from the datasets are identified using screening surveys, self-reported posts about diagnosis from social media or by their membership in different forums related to mental health. The most used fea-

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.who.int>

³<https://www.reddit.com>

tures are topic modelling, n-grams, Linguistic Inquiry and Word Count (LIWC), emotion and meta-data. The most used methods are Support Vector Machines (SVM), Logistic Regression, Random Forests and Neural Networks.

Coppersmith et al. (2016) show the differences in emoticons use between suicidal users and controls, neurotypicals using emojis with a much higher probability than a user before an attempt. Prior to the suicide attempt, the users at risk tend to use a more self-focused language, same as the people diagnosed with depression. The authors highlight different changes in post emotions before and after the suicide attempt. Users are also more likely to talk about suicide after an attempt than before it.

Sekulić et al. (2018) indicate that users diagnosed with bipolar disorders use more first-person singular pronouns, same as depressed people. They also use more words associated with emotions; words associated with positive emotions as well as words associated with negative emotions explained by alternating episodes of mania and depression.

Nalabandian et al. (2019) show that depressed persons tend to use more negative words and a self-focused language when writing about their interactions with a close romantic partner than when writing about other people around them. This is because people experience different symptoms of mental illness based on their interactions with other people.

Loveys et al. (Loveys et al., 2018) show the differences in language use of users with depression from different cultures to avoid cultural biases. Even if depression affects people all over the world, the way they experience and express it is shaped by their cultural context. Users from some ethnic groups does not address mental health issues online as much as the others and this can make the depression task more difficult. After topic modeling, the authors show that the words from each topic vary for each ethnic group, people discussing different themes relevant to their culture.

For diagnosis before the onset of the mental health disorders, Eichstaedt et al. (2018) use users' posts from Facebook to predict a future depression diagnosis. De Choudhury et al. (2013) use a classifier to predict users' depression likelihood ahead of the onset of illness, with different measures used: language, linguistic style, emotion, ego-network, demographics and user engagement.

We chose to tackle the problem of detecting early onset of depression from users' Reddit posts. To that end, we focus our efforts into processing the eRisk 2018 dataset (Losada et al., 2018), given its success at the Workshop for Early Risk Detection on the Internet⁴ within The Conference and Labs of the Evaluation Forum (CLEF) and its fruitful submissions from participants.

The teams from this workshop had different detection systems, based on bag of words ensembles (Trotzek et al., 2018), machine learning models with hand-crafted features (Trotzek et al., 2018; Ramiandrisoa et al., 2018; Cacheda et al., 2018; Ramírez-Cifuentes and Freire, 2018) or with different text embeddings (Trotzek et al., 2018; Ramiandrisoa et al., 2018; Ragheb et al., 2018), on sentence-level analysis to detect self references and extract different features (Ortega-Mendoza et al., 2018), on Latent Dirichlet Allocation (LDA) topic modelling (Maupomé and Meurs, 2018), models combining Term Frequency — Inverse Document Frequency with Convolutional Neural Networks (Wang et al., 2018) or other machine learning models. Most systems took the decision after the last chunk, only a few were able to emit a decision in the first chunks.

Several works addressing depression (Schwartz et al., 2014; Resnik et al., 2015) and PTSD (Coppersmith et al., 2015; Preoțiuc-Pietro et al., 2015) use a topic modelling approach showing that topics encountered texts have important discriminative power to make the distinction between persons suffering from mental illnesses and healthy controls.

3 Dataset

Early Risk Detection on the Internet (eRisk) workshops organized by CLEF explore the technologies that can be used for people's health and safety and the issues related to building tests collections (Losada et al., 2018). eRisk 2018 has two tasks, for early detection of depression and anorexia. We choose to focus on the task of detecting early onset of depression of social media users.

This task consists of sequentially processing chunks of Reddit posts from depressed users and controls. Submissions from each user are encoded in an xml file, one subject xml per chunk of data. Each xml contains the id of the subject and his posts and comments. Each submission has the posting time and the actual text. If a submission does

⁴<https://early.irlab.org/>

not have a title, it is considered a comment. The goal is to detect depression as early as possible and the dataset has to be processed in chronological order. The test collection of posts from depressed and non-depressed users is split into 10 chunks. As training data, the teams had access to data from eRisk 2017, both train and test. The test chunks were released one every week. Every week the teams had to decide whether to label the user as depressed or non-depressed or to wait for the test data of the following week.

The dataset contains 125 depressed users and 752 non-depressed users as training data and 79 depressed users and 741 non-depressed users as test data. The dataset has more posts and comments from people without depression than from users diagnosed with depression. From a total of 531,349 submissions, only 49,557 submissions are from users diagnosed with depression. The average time from the first to the last submission is between 2 and 3 years, so the posts were collected over a long period of time (Losada et al., 2018).

4 Method

Our methodology for early diagnosis of depression follows a classical Natural Language Processing pipeline. To clean the users' texts, we transform them into lowercase, we remove the punctuation and stopwords, the numbers and URLs are replaced with specific tokens and we perform stemming with Porter Stemmer (Porter, 1980). To reduce the dimension of the dictionary, we use collocations (Bouma, 2009) to extract meaningful bigrams and trigrams.

The number of posts and comments from non-depressed users is much higher than those from depressed users. To balance the two classes, we downsample the majority class to a ratio of 2:1.

We train our Latent Semantic Indexing model with 128 topics on every users' post. We use this model to extract topic modelling embeddings from users' texts and use them as input to our fully connected neural network architecture. The neural network has three hidden layers of 512, 256 and 256 neurons respectively, Leaky ReLU activation and we use Dropout for regularization. We use a random sample of 20% of the training data provided by the organisers of the competition for validation.

The network has two outputs, one for classifying if the user is depressed or not and one for confidence estimation. The motivation for using this

architecture is to learn the confidence (DeVries and Taylor, 2018) of our predictions and use it to make a decision on whether to label a user or wait for the next chunk of data. The learned confidence, besides its use case in out-of-distribution detection, can be used as a measure for how much the model trusts its classification output to be correct. As such, we consider the classification output only if the confidence exceeds a certain threshold. As indicated by DeVries et al. (2018), the network loss is computed by interpolating the predicted probabilities p with the target y , using the computed confidence score c , as follows:

$$p'_i = c \cdot p_i + (1 - c)y_i \quad (1)$$

The final loss is then given by:

$$\mathcal{L} = - \sum_{i=1}^M \log(p'_i)y_i - \lambda \log(c) \quad (2)$$

Where, in our case, $M = 2$, is the number of classes. The loss includes an additional term that forces the predicted confidence to be as high as possible. We performed an ablation study on the validation data on the confidence penalty λ .

A recent study by Hein et al. (2019) shows that neural networks with ReLU activation functions tend to be overconfident on incorrectly classified samples, thus we can not rely only on the output probabilities, and the predicted confidence offers a more reliable measure of uncertainty of the classification.

As the number of submissions seen by the model increases, we want to make a decision as early as possible and thus we use a decaying function that decreases progressively the fixed threshold for confidence. The decision function is defined as follows:

$$D_w(x) = \begin{cases} \text{decide for } x & \text{if } c > T * e^{-sw^2} \\ \text{wait for data} & \text{otherwise} \end{cases} \quad (3)$$

Where x is the embedding for the current user's posts, w is the week number (i.e. the current chunk), s is a scaling factor and T is the initial threshold. We choose $T = 85\%$ and progressively scale it down to 40%. The scaling factor is computed such that, at the final chunk, the threshold is less than the smallest confidence encountered on the training data.

At the test phase, the proposed model does not make an independent decision for each chunk of data in the test set. In the first chunk of data, if the model is not confident enough to make a final decision regarding the depressed or non-depressed status of a user, then, starting with the second chunk of data, we concatenate the current chunk with the previously available chunks for the current user. This way, the LSI model has more data for making better informed predictions.

5 Results

Our results on eRisk 2018 dataset are presented in Table 1. Even if F_1 is a standard evaluation measure used for imbalanced classification, it does not include the time component of the early detection task, thus Losada and Crestani (2016) propose an evaluation metric better suited for this task, the Early Risk Detection Error (ERDE).

ERDE is defined as:

$$ERDE_o(d, k) = \begin{cases} c_{fp} & \text{if } d = FP \\ c_{fn} & \text{if } d = FN \\ lc_o(k) \cdot c_{tp} & \text{if } d = TP \\ 0 & \text{if } d = TN \end{cases} \quad (4)$$

The use of false positive (FP), false negative (FN), true positive (TP) and true negative (TN) for prediction d is to avoid the classifiers that always predict the label of the majority class. $lc_o(k) \in [0, 1]$ encodes a cost for the delay in detecting TP. For the eRisk datasets, where the number of negative labels is greater than positive labels, the value of c_{fn} is 1 and c_{fp} is 0.1296, set according to the proportion of depressed users in eRisk 2017 dataset (Losada et al., 2018). c_{tp} is set to c_{fn} because the late detection of people at risk of depression can have serious consequences, a late detection is considered as equivalent to not detecting the depressed user at all. The late detection of TN cases does not affect the effectiveness of the system.

The goal of the system is to detect as early as possible people at risk of depression. For the detection of non-depressed users, the time of the detection is not relevant. The latency cost function, which grows with k (the number of submissions seen by the algorithm), is defined as:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \quad (5)$$

o represents the number of posts after which the cost grows more quickly.

Method		ERDE ₅	ERDE ₅₀	F ₁	Prec	Rec
Baseline LSI		9.98%	8.29%	0.25	0.22	0.29
LSI _c $\lambda = 0.01$		14.19%	11.41%	0.25	0.15	0.87
LSI _c $\lambda = 0.1$		11.12%	9.09%	0.28	0.20	0.48
LSI_c $\lambda = 0.2$		10.24%	7.74%	0.30	0.25	0.38
LSI _c $\lambda = 0.4$		11.15%	8.53%	0.25	0.17	0.47
LSI _c $\lambda = 0.6$		12.67%	10.17%	0.25	0.15	0.71
LSI _c $\lambda = 0.8$		10.53%	8.08%	0.30	0.21	0.56
Funez al.(2018)	et	8.78%	7.39%	0.38	0.48	0.32
Trotzek al.(2018)	et	9.50%	6.44%	0.64	0.64	0.65

Table 1: Classification results on the detection of early onset of depression task from eRisk 2018 dataset.

The detection task is difficult, as seen in the low values of F_1 and Precision. However, the task is to predict *early* onset of depression, and for that, the ERDE metrics are more appropriate, as they are a measure of prediction delay. ERDE₅ metric is very sensitive to delays, after the first 5 submissions from the user the penalties grow quickly. In contrast to ERDE₅, for ERDE₅₀ the penalties grow only after the first 50 submissions from the user. The difference between ERDE₅ and ERDE₅₀ is very important in practice because of the consequences of late detection of depression signs. As the task suggests, the detection should be made as early as possible.

To measure the impact of our learned out-of-distribution confidence from the neural network, we also trained a plain ReLU network with cross-entropy loss. For this model, we employed a hard threshold on the output probabilities for whether to wait for more data or classify the sample. As shown by Hein et al. (2019), ReLU networks can be overly confident on misclassified examples. This is shown in Table 1: the model has a low ERDE₅ score as the output probabilities mostly have extreme values, which means that for most users the model makes a decision from the first chunk of data.

We trained our model with different λ values in order to see the impact of the confidence component on the results. Larger values for λ make the model overly confident, as expected from Equation 2, the best performing model being the one with $\lambda = 0.2$. Smaller values of λ generate a wider confidence distribution on the training examples, facilitating the decision process, as extreme values either make the model overly-confident on every example, or not confident at all. This is consistent

with findings by DeVries et al. (2018).

In Table 1 we also present the best two submission from the eRisk 2018 Workshop, the one from Funez et al. (2018), having the best results for the ERDE₅ metric, and the one from Trotzek et al. (2018) having the top ERDE₅₀ score.

We can assume from these results that topics encountered in user writings have important discriminatory power. Depressed users mostly write about different subjects than non-depressed subjects, consistent with results from the work of Resnik et al. (2015). The writings from users diagnosed with depression are more focused on their feelings and their life events. Topics related to those themes contain words such as *someone kill, bad though, never able to get, forever alone, life save, stay sober, i am sad, still can't, improve life. new hope, oneself, tell anything, happy sad, hope one day*. Texts from non-depressed users are found in topics related to their hobbies containing specific words: *black mirror, first season, movie adaptation, hologram, nine inch nails, jimi hendrix, artist name, vlog, game, fallout, terra mistica, way to make money, paid time, really proud, amazon wishlist, food industry, white bread*.

6 Conclusion

In this paper, we use the eRisk 2018 dataset on Early Detection of Signs of Depression for depression classification from Reddit posts. Our method uses Latent Semantic Indexing for topic modelling and to generate the embeddings used as input for our neural network, but focuses on using a learned out-of-distribution confidence score alongside the classification output to decide whether to label the user or wait for more data. Besides its initial use case in out-of-distribution detection, we repurposed the confidence score as a measure for how much the model trusts its classification output to be correct. We showed that, in general, there is a significant difference in writing topics depending on the users' mental health, to the extent that it contains enough information for use in classification.

Acknowledgements

We would like to thank our reviewers for their useful comments and suggestions that helped us improve this paper and also to the organizers of the eRisk Workshop for their efforts in encouraging the research on mental illnesses detection from social media.

Liviu P. Dinu was supported by a grant of the Romanian Ministry of Education and Research, CCCDI—UEFISCDI, project number 411PED/2020, code PN-III-P2-2.1-PED-2019-2271, within PNCDI III.

References

- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.
- Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.
- Fidel Cacheda, Diego Fernández Iglesias, Francisco Javier Nóvoa, and Victor Carneiro. 2018. Analysis and experiments on early detection of depression. *CLEF (Working Notes)*, 2125.
- Rafael A Calvo, David N Milne, M Sazzad Husain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.
- Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.
- Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.
- Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoȃuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

- Dario G Funez, Maria José Garciarena Ucelay, Maria Paula Villegas, Sergio Burdisso, Leticia C Cagnina, Manuel Montes-y Gómez, and Marcelo Errecalde. 2018. Unsl’s participation at erisk 2018 lab. In *CLEF (Working Notes)*.
- Sharath Chandra Guntuku, David Yaden, Margaret Kern, Lyle Ungar, and Johannes Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 12.
- Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50.
- Maria Khokhlova, Viviana Patti, and Paolo Rosso. 2016. Distinguishing between irony and sarcasm in social media texts: Linguistic observations. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–6. IEEE.
- David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.
- David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th International Conference of the CLEF Association, CLEF*.
- Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87.
- Diego Maupomé and Marie-Jean Meurs. 2018. Using topic extraction on social media content for the early detection of depression. *CLEF (Working Notes)*, 2125.
- Taleen Nalabandian and Molly Ireland. 2019. Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or friends. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 62–73.
- Rosa María Ortega-Mendoza, Adrián Pastor López-Monroy, Anilu Franco-Arcega, and Manuel Montes-y Gómez. 2018. Peimex at erisk2018: Emphasizing personal information for depression and anorexia detection. In *CLEF (Working Notes)*.
- Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Daniel Preotiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30.
- Waleed Ragheb, Bilel Moulahi, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2018. Temporal mood variation: at the clef erisk-2018 tasks for early risk detection on the internet. In *Proceedings of the 9th International Conference of the CLEF Association*.
- Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara, and Véronique Moriceau. 2018. Irit at e-risk 2018. In *Proceedings of the 9th International Conference of the CLEF Association*.
- Diana Ramírez-Cifuentes and Ana Freire. 2018. Upf’s participation at the clef erisk 2018: Early risk prediction on the internet. In *Cappellato L, Ferro N, Nie JY, Soulier L, editors. Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum; 2018 Sep 10-14; Avignon, France.[Avignon]: CEUR Workshop Proceedings; 2018. p. 1-12. CEUR Workshop Proceedings*.
- Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.
- Hannah Ritchie and Max Roser. 2018. Mental health.
- H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.
- Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. *arXiv preprint arXiv:1811.04655*.
- Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF (Working Notes)*.
- Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*.

Building a Treebank in Universal Dependencies for Italian Sign Language

Gaia Caligiore¹, Cristina Bosco², Alessandro Mazzei²

¹Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Università degli Studi di Torino

²Dipartimento di Informatica, Università degli Studi di Torino

gaia.caligiore@edu.unito.it, {alessandro.mazzei|cristina.bosco}@unito.it

Abstract

The Italian Sign Language (LIS) is the natural language used by the Italian Deaf community. This paper discusses the application of the Universal Dependencies (UD) format to the syntactic annotation of a LIS corpus. This investigation aims in particular at contributing to sign language research by addressing the challenges that the visual-manual modality of LIS creates generally in linguistic annotation and specifically in segmentation and syntactic analysis. We addressed two case studies from the storytelling domain first segmented on the ELAN platform, and second syntactically annotated using CoNLL-U format.

1 Introduction and research goals

The Italian Sign Language (LIS) is the natural language used by the Italian Deaf community. Signed languages have been extensively studied in the last years (Brentari, 2010). From a theoretical point of view, Signed languages are of interest to the linguistic domain since they are *multi-channels* natural languages, where the coexistence of different articulators (hands, face, lips, posture, feet, etc.) is the test-bed for the formalization of new linguistic theories or objects (e.g. (Huenerfauth, 2006)). From a practical point of view, there is a real necessity to design and realise automatic translators for Deaf communities (Bragg et al., 2019). We want to investigate LIS with the same means used for Vocal Languages (VL) and verify if, in doing so, LIS can be properly represented. In this context, language-specific strategies and resources should be developed. The reference framework in

this work is the *Universal Dependencies* formalism¹ (UD), a *de facto* standard for syntactic annotation.

The main goals of this research are three. The first goal is theoretical. We want to investigate the expressiveness of UD tags and its relations with LIS. Signed languages have peculiar forms of lexicons and syntax and we want to experimentally verify the expressive power of the UD formalism in representing this richness. The second goal is theoretical as well. We want to determine the extent of the similarities and differences in syntactic constructions between Italian (as reported in the Italian-UD (Simi et al., 2014)), Swedish Sign Language (SSL, as reported in the SSL-UD treebank (Mesch and Schonstrom, 2018)) and LIS. At the present moment, the SSL is the only sign language that has been annotated on UD. The SSL treebank is comprised of 203 sentences taken from the Swedish Sign Language Corpus (SSLC) (Mesch and Schönström, 2018; Mesch and Wallin, 2015). Being the only reference for the construction of a treebank for a sign language, the SSL treebank was a fundamental resource for the choice of the direction to follow in the annotation of LIS, particularly with regard to the first step of the process, i.e. the segmentation on ELAN (see section 2.1). The third goal is more practical. We want to create a UD compliant resource for the syntactic annotation of LIS: the first LIS-UD treebank. To our knowledge, apart from the domain-specific bilingual corpus developed in the projects ATLAS and LIS4ALL on automatic translation (Mazzei et al., 2013; Geraci et al., 2014; Mazzei, 2015), this is the first attempt to use dependency relations for representing LIS syntax.

For building the corpus, we selected two case studies in the storytelling domain: all of the sentences of two LIS videos, namely the fairy tale *Cappuccetto Rosso* (Little Red Riding Hood) and

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://universaldependencies.org/>

the story *I tre fratelli* (The three brothers, written by the Italian writer Grazia Deledda), were collected in the novel treebank that, at the moment, is comprised of 257 dependency trees. While in the *Cappuccetto Rosso* story the signer signs the story without an Italian reference text, in the *I tre fratelli* story the signer is translating from a well defined written Italian text. By using the full original version of these stories, we had to face the challenge represented by unrestricted real data. For instance, very long and complex sentences were annotated and translated into LIS.

Considering the intrinsic complexity of the task and the novelty of the project, in the preliminary release of data described in this paper we only addressed some of the features of LIS. For instance, the location in space of a sign is only annotated in the portion of analysis carried out on ELAN but was not transferred in the CoNLL-U files. Furthermore, non-manual elements – which are one of the fundamental means used by LIS signers to convey meaning (Volterra, 2004) – are not included in any annotation layer developed in this project. This is the result of a lack of a more appropriate annotation strategy that is specific to sign languages within the UD framework, which is originally developed for the analysis of VLS and, by default, does not include the possibility to annotate the features of a language that go beyond the alphabetical construction of a word (or gloss, in this case).

The paper is organized as follows. In the Section 2, the data collection and the morphological and syntactic annotation processes will be described. In Section 3, language-specific morpho-syntactic phenomena are discussed mainly focusing on pointing signs as Highly Iconic Structures (henceforth HIS). The strategies used to annotate signs and their dependency relations are justified. Section 4 concludes the paper by providing some issue on the future development of the project.

2 Data Annotation

In this Section, we describe the main steps for the realization of the annotation of the UD-LIS, the pre-processing, which consists in the analysis with ELAN, and the application of the tags and relations of the UD format for the generation of the CoNLL-U format of each LIS sentence.

2.1 Analysis on ELAN

Following the same annotation procedure applied for the SSL treebank (Mesch and Schonstrom, 2018), the ELAN platform was used for the identification, segmentation and definition of each sign of our corpus. ELAN (EUDICO Linguistic Annotator)² is a computer software initially released in 2000 by the Max Planck Institute for Psycholinguistics in the Netherlands. ELAN is used to annotate audiovisual files manually and semi-automatically and allows annotators to tag video material frame by frame with information arranged on multiple lines that can be defined by the program itself or personalized by the annotator (Brentari, 2010). It is also a useful tool in multimodality research since it allows the user to manually create multimodal annotations, useful for the analysis of sign languages (Wittenburg et al., 2006). Mesch and Wallin – creators of the SSL treebank – state that “ELAN allows researchers to provide time-aligned annotations of a video file on parallel tiers, making it useful for representing individual articulators on separate tiers as they are used simultaneously to produce a single sign.”. For these reasons, the software is considered to be the most used for sign language annotation (Branchini et al., 2013). As a result, four tiers were developed to describe the main qualities of a sign in this context: *segno*, *luogo*, *UD POS* *Tag*, *traduzione*.

Defining a strategy to gloss the signs included in the *Segno* (sign) tier was a challenging task since the final aim was that of providing unambiguous and easily retrievable glosses. The rule that is generally followed when writing down a sign gloss is to write the translation of the sign as it would be normally written in the vocal language, but in capital letters. This writing strategy might cause ambiguity since different variations of a sign can be translated with the same gloss. At this stage of the annotation process, glossing strategies adopted for LIS and for the SSL treebank were compared, in order to minimise ambiguity and create a system that refers to both languages. The developers of the SSL treebank started from the SSL corpus that, in turn, referred to the SSL Dictionary (Ostling et al., 2017). For this reason, it was decided that the most appropriate source for sign gloss annotation would be a well-established dic-

²<https://archive.mpi.nl/tla/elan/download>

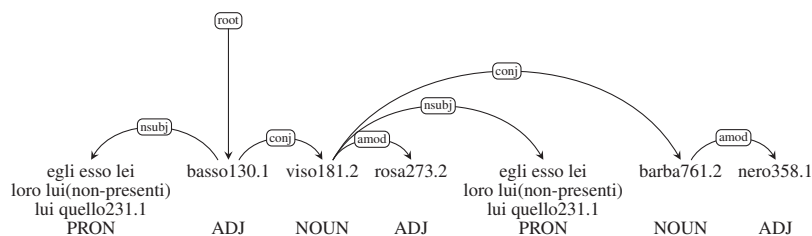


Figure 1: Lui è basso ed ha il viso rosa e la barba nera.

Piccolo e roseo, con una gran barba nera incolta.

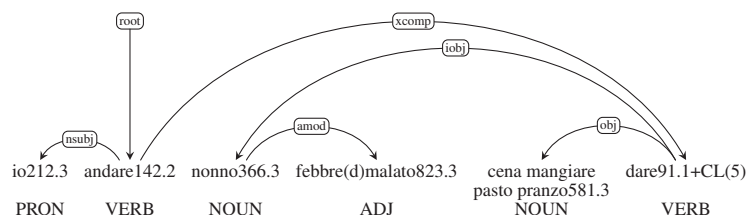


Figure 2: Io vado a dare da mangiare alla nonna malata.

tionary: the *Dizionario bilingue elementare della Lingua dei Segni Italiana LIS* (Radutzky, 1992) in its digital version, which is considered as the most easily retrievable and unambiguous collection of glosses within the context of LIS, including more than 2500 terms signed by native signers. In this dictionary, a gloss – made up of a translation of the sign into Italian and a sequence of numbers – is associated to each sign. For instance, the LIS sign for RED is glossed as “*rosso.202.1*”; for signs not found in the dictionary, glosses are taken from other resources, such as SpreadTheSign³. In this case, a translation in caps lock of the sign into Italian is associated to the code “-STS”, as in the gloss “*NEMICO-STS*” (ENEMY). Lastly, if a sign is not found in any of the mentioned resources, it is autonomously developed with SignWriting (Renzo et al., 2010) and glossed with the code “SW-”, as was done to gloss the sign TO PUT ON a sleeping cap: “*SW-indossare-cuffia*”.

Luogo (location) defines where the sign is articulated in the signing space and is based on the 16 sign locations identified by Radutzky (Radutzky, 1992) which are *parte superiore del capo* (upper part of the head), *faccia* (face), *occhi* (eyes), *naso* (nose), *orecchie* (ears), *guancia* (cheeks), *bocca* (lips), *mento* (chin), *spalla* (shoulder), *petto* (chest), *gomito* (elbow), *polso* (wrist), *mano non dominante* (non-dominant hand), *tronco inferiore* (waist). The 16th and most used location of a sign is the Neutral Signing Space that is described by

the acronym SN (*Spazio Neutro*).

UD POS Tag includes the UD Part of Speech Tag⁴ associated to the sign. A peculiarity of this tier is that the POS Tag might not correspond to the form of the sign gloss. For the annotation of this tier, priority is given to the function of the sign within the sentence, rather than its gloss. In fact, in some cases the gloss associated to a sign will not correspond to the role of the sign in that specific syntactic context. For instance, the gloss of the sign *POESIA* (poesia-STS) suggests that the sign is a NOUN. Yet, in the *I tre fratelli* video, *POESIA* plays the role of an adjective. In the sign sequence *POESIA TEMPO*, the previously mentioned sign has the meaning of “*tempo poetico*” (poetic time). In this case, the UD POS Tag for the sign *POESIA* will be ADJ and not NOUN. This is because priority was given to the function of the sign within the sentence, rather than its gloss.

Traduzione (translation) provides a translation of the LIS sentence into spoken Italian. If the signs are a direct translation from Italian, as in the *I tre fratelli* video, the information included in this tier will be a word-for-word transcription of the spoken or written text. If the signer is not translating, as in the *Cappuccetto Rosso* video, the information included in the tier will be a translation into Italian that imitates the structure of the sentence in LIS as closely as possible. This tier was included to facilitate the understanding of a signed sentence given that the sequence of sign glosses will look

³<https://www.spreadthesign.com/it.it/search/>

⁴<https://universaldependencies.org/u/pos/>

fragmented. By providing a linear translation into spoken Italian of each sentence, a non-signer will be able to have a general understanding of the sentence. Furthermore, as mentioned in the previous section, POS Tags might not correspond to sign glosses. Therefore, by providing this translation, any discrepancies between sign glosses and POS Tags will be justified.

2.2 Annotation in UD :

The information encoded in ELAN was transferred in CONLL-U files and split into its ten columns, except for the LEMMA, DEPS and MISC columns, where no information is included.

2.2.1 Morphological Annotation

Sign glosses and UPOS Tags were included respectively in the FORM and UPOS columns, and coarse-grained tags taken from the Tanl POS Tagset⁵ were included in the XPOSTAG column. Language-specific annotation strategies can be found in the FEATS column where specific symbols and labels taken from different sources were used to provide more information on the peculiarities of a sign or of its production. Based on SSLC tags (Mesch and Wallin, 2015), sign types were marked with *@b* for finger-spelled signs and *@g* for gesture-like signs. Reduplicated signs were signalled with the feature tag *@RDP=true*, adapted from (Mesch and Schonstrom, 2018). Role shift is marked with *RS=* followed by the symbols *<* and *>* and the codes 3a, 3b, 1 or 2 which are used to identify the positioning of the signer in the orthogonal signing space (Pfau et al., 1987).

2.2.2 Syntactic Annotation

The annotation of the HEAD and DEPREL columns of the CoNLL-U format is at the very core of the research. As for the definition of a root node, the UD standard indicates that, in Italian, the root is usually a verbal predicate or a noun. If the verbal predicate is not present due to ellipsis, the root is moved to the leftmost dependent of the verbal predicate⁶.

The strategy for finding the syntactic structure of each sentence consists in looking at Italian and SSL treebanks, choosing the most appropriate annotation with respect to the specific phe-

nomenon to be annotated or the context where it occurs. For instance, for the annotation of particular signs we were inspired by the solution adopted in SSLC (see 2.2.1). Nevertheless, if no annotation is deemed to be adequate, a novel independent solution that seems most fitting is applied. If a sentence presented a unique combination of signs and no corresponding or similar dependency relations were found in these other treebanks, the construction of the dependency tree was based on an independent solution that is compliant with the general criteria adopted for LIS in the novel treebank.

In the following Section, a small selection of phenomena encountered in the treebank are described.

3 Annotation of Language-specific Morpho-syntactic Phenomena

In this section we discuss two LIS phenomena, i.e. pointing signs and repetitions, which we addressed in the development of the novel resource.

3.1 Pointing signs

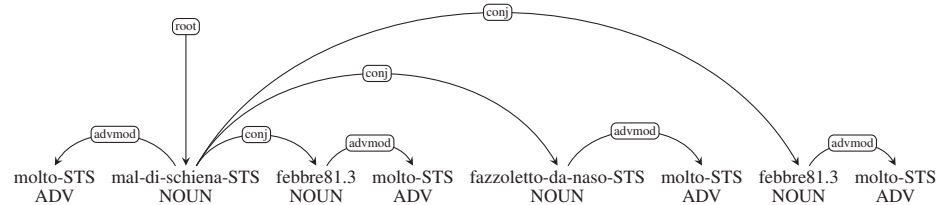
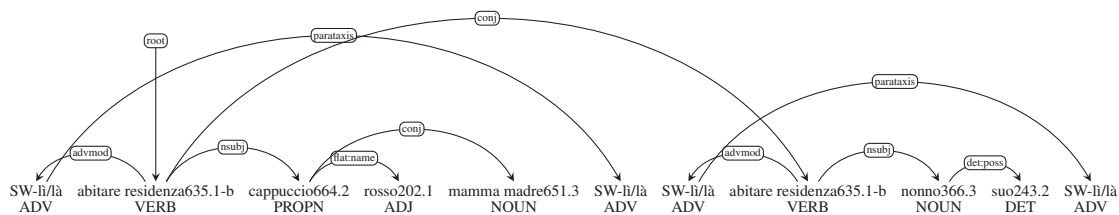
In LIS, pointing signs are HIS realized using an extended index finger and carrying out several functions. For example, a pointing sign can function as a pronoun, as a determiner or as an adverb (Cormier et al., 2013) depending on the context in which it is performed. They are in all cases deictics found in PE-clauses⁷, used to establish a location in space of a certain referent and create agreement in space. In the annotation of dependency relations, the second repetition of a deictic sign with an anaphoric function was either marked as dependent on the first one, or – as can be seen in figure 1 – as dependent on a noun (*viso181.2*) that, in turn, is attached to the root.

Additionally, pointing signs can also be used by the signer to refer to himself or herself during role shift, that is, while impersonating a character, as in Figure 2. A sub-type of pronominal or determiner pointing signs are demonstrative pointing signs (Cormier et al., 2013, p.232), which could be compared to Italian demonstrative pronouns. When a pointing sign was used as an adverb of

⁵http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

⁶<https://universaldependencies.org/it/dep/root.html>

⁷PE-clauses are labelled as such by Branchini and Donati (Branchini and Donati, 2009). These clauses can be compared to relative clauses in spoken Italian and include a PE-marker that is a sign “[...] realized manually with the index finger stretched out and shaken downwards [...]” and is “[...] coreferential with an NP within the clause, and this coreferentiality can be realized through agreement in space”.



References

- Carl Borstell. 2011. Revisiting Reduplication. Toward a description of reduplication in predicative signs in Swedish Sign Language. Master's thesis, Stockholm University, Faculty of Humanities, Department of Linguistics.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2019, Pittsburgh, PA, USA, October 28-30, 2019*, pages 16–31.
- Chiara Branchini and Caterina Donati. 2009. Relatively different: Italian sign language relative clauses in a typological perspective. In Anikó Lipták, editor, *Correlatives Cross-Linguistically*, pages 157–191. John Benjamins Publishing Company, Amsterdam.
- Chiara Branchini, Carlo Cecchetto, and Isabella Chiari. 2013. La lingua dei segni italiana. In Gabriele Iannaccaro, editor, *La linguistica italiana all'alba del terzo millennio (1997-2010)*, pages 369–404. Bulzoni.
- D. Brentari. 2010. *Sign Languages*. Cambridge Language Surveys. Cambridge University Press.
- Kearsey Cormier, Adam Schembri, and Bencie Woll. 2013. Pronouns and pointing in sign languages. *Lingua*, Volume 137:230–247.
- Carlo Geraci, Alessandro Mazzei, and Marco Angster. 2014. Some issues on Italian to LIS automatic translation. The case of train announcements. In *Proc. of CLiC-it 2014, first Italian conference on computational linguistics*, December.
- Matt Huenerfauth. 2006. Representing coordination and non-coordination in American Sign Language animations. *Behav. Inf. Technol.*, 25(4):285–295.
- Alessandro Mazzei, Leonardo Lesmo, Cristina Battaglino, Mara Vendrame, and Monica Bucciarrelli. 2013. Deep Natural Language Processing for Italian Sign Language Translation. In *AI*IA 2013: Advances in Artificial Intelligence - XIIIth International Conference of the Italian Association for Artificial Intelligence, Turin, Italy, December 4-6, 2013. Proceedings*, pages 193–204.
- Alessandro Mazzei. 2015. Translating Italian to LIS in the rail stations. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 76–80, Brighton, UK, September. Association for Computational Linguistics.
- Johanna Mesch and Krister Schonstrom. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 121–126. European Language Resources Association.
- Johanna Mesch and Krister Schönström. 2018. From Design and Collection to Annotation of a Learner Corpus of Sign Language. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*. European Language Resources Association, may.
- Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics (IJCL)*, Volume 20:102–120.
- Robert Ostling, Carl Borstell, Moa Gardenfors, and Mats Wiren. 2017. Universal Dependencies for Swedish Sign Language. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 303–308, Gothenburg. Linköping University Electronic Press.
- Roland Pfau, Marcus Steinbach, and Annika Herrmann (Eds.). 1987. *A matter of complexity. Subordination in sign languages*. Walter de Gruyter Inc. and Ishara Press, Boston/Berlin and Preston, UK.
- Elena Radutzky. 1992. *Dizionario bilingue elementare della lingua italiana dei segni: Oltre 2.500 significati*. Kappa, Roma.
- Alessio Di Renzo, Luca Lamano, Tommaso Lucoli, Barbara Pennacchi, Gabriele Gianfreda, Giulia Pettita, Claudia Savina Bianchini, Paolo Rossini, and Elena Antinoro Pizzuto. 2010. *Scrivere la LIS con il Sign Writing. Manuale Introduttivo*. Consiglio Nazionale delle Ricerche, Roma.
- Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of LREC 2014*, page 83–90.
- Virginia Volterra, editor. 2004. *La lingua dei segni italiana*. Il Mulino.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genova, Italy, May. European Language Resources Association (ELRA).

Analysis of lexical semantic changes in corpora with the Diachronic Engine

Pierluigi Cassotti and Pierpaolo Basile and Marco de Gemmis and Giovanni Semeraro

Department of Computer Science

University of Bari Aldo Moro

Bari, Italy

{firstname.surname}@uniba.it

Abstract

English. With the growing availability of digitized diachronic corpora, the need for tools capable of taking into account the diachronic component of corpora becomes ever more pressing. Recent works on diachronic embeddings show that computational approaches to the diachronic analysis of language seem to be promising, but they are not user friendly for people without a technical background. This paper presents the *Diachronic Engine*, a system for the diachronic analysis of corpora lexical features. *Diachronic Engine* computes word frequency, concordances and collocations taking into account the temporal dimension. It is also able to compute temporal word embeddings and time-series that can be exploited for lexical semantic change detection.

1 Motivation and Background

Synchronic corpora are widely used in linguistics for deriving a set of abstract rules that govern a particular language under analysis by using statistical approaches. The same methodology can be adopted for analyzing the evolution of word meanings over time in the case of diachronic corpora. However, this process can be very time-consuming. Usually, linguists rely on software tools that can easily explore and clean the corpus, while highlighting the more relevant linguistic features. Sketch Engine¹ (Kilgarriff et al., 2004; Kilgarriff et al., 2014) is the leading tool in the corpus analysis field. Beyond several interesting features, Sketch Engine includes *trends* (Kilgarriff et al., 2015), which allow for diachronic

analysis based on the frequency distribution of words. Trends rely on merely frequency features, ignoring word usage information. Moreover, the Sketch Engine interface does not provide temporal information about concordances and collocations. NoSketchEngine² is an open-source version of SketchEngine. It requires technical expertise for the setup and, contrarily to SketchEngine, it does not support word sketches, terminology, thesaurus, n-grams, trends and corpus building. An interesting system is DiaCollo³ (Jurish and der Wissenschaften, 2015), a software tool for the discovery, comparison, and interactive visualization of target word combinations. Combinations can be requested for a particular time period, or for a direct comparison between different time periods. However, DiaCollo is focused exclusively on the extraction and visualization of collocations from diachronic corpora.

In recent works about computational diachronic linguistics, techniques based on word embeddings produce promising results. In Semeval Task 1 (Schlechtweg et al., 2020), for instance, type embeddings rich high performances on both subtasks. However, these techniques are not included in any aforementioned linguistic tool. In order to bridge this gap, we try to build a tool that includes approaches for the analysis of diachronic embeddings. The result of our work is Diachronic Engine (DE), an engine for the management of diachronic corpora that provides tools for change detection of lexical semantics from a frequentist perspective. DE includes tools for extracting diachronic collocations, concordances in different time periods as well as for computing semantic change time-series by exploiting both word frequencies and word embeddings similarity over time.

The rest of the paper is organized as follows:

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.sketchengine.eu/>

²<https://nlp.fi.muni.cz/trac/noske>

³<https://www.clarin.eu/showcase/diacollo>

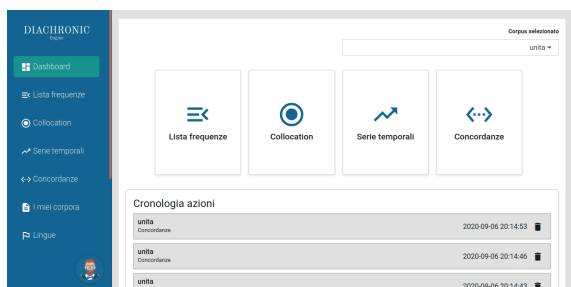


Figure 1: Diachronic Engine web interface.

Section 2 describes the technical details of DE, while Section 3 shows some use cases of our engine that encompass that address time-series. We also present the results of a preliminary evaluation about the system’s usability in Section 4. Conclusions and future work close the paper.

2 Diachronic Engine

Diachronic Engine (DE) is a web application for lexical semantic change analysis in diachronic corpora. The DE pipeline needs diachronic corpora to compute statistics about the corpus. A diachronic corpus must include a temporal feature (e.g., year or timestamp of the publication date); DE exploits that feature to sort the documents.

We adopt the vertical format to represent word information, as specified for the IMS Corpus Workbench (CWB). In a vertical corpus, each word is in a new line. In each line, fields, called p-attributes, are separated by tabs. In DE the default p-attributes are word, lemma, PoS tag and syntactic dependency. Non-recursive XML tags (s-attributes) on a separate line can be used for representing sentences, paragraphs and documents.

Corpora can be served in vertical format⁴ or in plain-text mode; in the latter case, the plain-text is transformed in vertical format using the Spacy UDPipe⁵ (Straka, 2018) tool, which splits plain-text into sentences and then predicts the PoS-tag, the lemma and the syntactic dependency for each token. UDPipe is a dependency parser that provides models for several languages. Models are built by using the Universal Dependencies⁶ datasets as training data. Input files’ names must contain the temporal tag of the period to which they refer. DE automatically detects temporal pat-

⁴https://www.sketchengine.eu/my_keywords/vertical/

⁵<https://pypi.org/project/spacy-udpipe/>

⁶<http://universaldependencies.org>

terns in the name of the files. In particular, the last sequence of numbers in the file name is used to sort the documents.

Corpora are stored and managed by the CWB, a tool for the manipulation of large, linguistically annotated corpora. In particular, DE relies on the Corpus Query Processor (CQP) (Christ et al., 1999), a specialized search engine for linguistic research.

For building temporal word embeddings, DE exploits Temporal Random Indexing (TRI) (Basile et al., 2014; Basile et al., 2016) that computes a word vector for each time period by summing shared random vectors over all the periods. TRI is able to produce aligned word embeddings in a single step and it is based on Random Indexing (Sahlgren, 2005), where a word vector (word embedding) $sv_j^{T_k}$ for the word w_j at time T_k is the sum of random vectors r_i assigned to the co-occurring words taking into account only documents $d_l \in T_k$. Co-occurring words are defined as the set of m words that precede and follow the word w_j . Random vectors are vectors initialized randomly and shared across all time slices so that word spaces are comparable.

Future versions will include other approaches, such as Procustes (Hamilton et al., 2016), Dynamic Word Embeddings (Yao et al., 2018), Dynamic Bernoulli Embeddings (Rudolph and Blei, 2018) and Temporal Referencing (Dubossarsky et al., 2019).

The DE architecture is based on the client-server paradigm. The back-end of DE has been developed with Flask, a web framework written in Python. Concordances are retrieved by CQP, that indexes the corpus as soon as it is uploaded to the server, while collocations and frequencies are computed in Python. The back-end provides a set of services by a REST API where the input/output is based on JSON messages.

The back-end consists of three macro components: User Handler, Corpus Handler and Diachronic Operations. The User Handler manages registered users information such as username and passwords. Admitted operations on users are creation, read, update and delete. The Corpus Handler Component manages corpora information such as name, language, the list of fields in the vertical files, corpus visibility. Moreover, it deals with corpora types: each corpus has a label indicating if it is synchronic or diachronic. For di-

achronic corpora also the temporal range is stored. Operations admitted on corpora are: creation, update, delete, search and read. The Diachronic Operations component shows frequency lists, collocations of words, time-series, change-points and concordances. This component relies on CWB that indexes vertical files.

The Diachronic Operations component architecture is sketched in Figure 2.

The front-end of DE has been developed with JHipster⁷, using Spring⁸ for server-side applications and Angular for client-side applications. The front-end communicates with the back-end by the means of the REST API.

The front-end design is inspired by the Google’s Material Design and the Sketch Engine interface. The user interface provides multilingual support in Italian and English, but we plan to extend it to other languages.

This architecture allows the independence between the back-end and the front-end, in this way is possible to develop a different front-end or connect the front-end to a different implementation of the back-end. The only constraint is the REST API interface.

A screenshot of the DE homepage is provided in Figure 1. The homepage provides an easy access to all corpora owned by the logged user with links to available tools. The front-end provides also tools for creating and managing users and corpora. In particular, it is possible to define different grant permissions for each corpus.

The tool is distributed as open-source software under the GNU v3 license⁹.

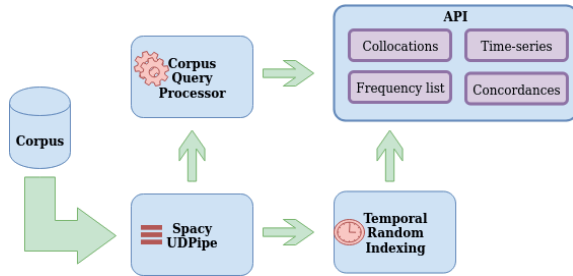


Figure 2: Diachronic Engine corpora manager.

⁷<https://www.jhipster.tech/>

⁸<https://spring.io/>

⁹<https://github.com/swapUniba/Diachronic-Engine>

2.1 DE tools

DE provides a set of tools for managing and querying diachronic corpora. The core of the back-end is based on the IMS Open Corpus Workbench (CWB)¹⁰, which allows querying the indexed corpora by using the powerful CQP. Other tools have been integrated to facilitate the analysis of a diachronic corpus:

Word frequency Many works show a correlation between lexical semantic change and frequency differences between time periods. Google Ngram Viewer (Michel et al., 2011) uses n-grams frequencies over time to show the change in the semantics of n-grams. SketchEngine exposes the Trends tool, which uses a linear regression of frequencies to predict words that appear to be changed. In DE, queries can be filtered by part-of-speech, as well as by time periods. We use normalized frequencies, that can be filtered by time period.

Collocations Collocations have shown to be an effective tool in diachronic analysis (Basile et al., 2019). A collocation is a sequence of words that occurs more often than would be expected. In order to compute the collocation strength we use the logDice (Rychlý, 2008):

$$\log \frac{2f_{xy}}{f_x + f_y}$$




logDice takes into account the frequency of the word f_x , of the collocate f_y and the frequency of the whole collocation f_{xy} . Collocation results can be grouped by the PoS tag.

Concordances Concordances offer a way to find “the evidence” directly in the text by exploiting the context. The Concordances tool lists instances of a word with its immediate left and right context and the period the collocation belongs to. An example of concordances from “L’Unità” (Basile et al., 2020), is shown in Figure 3.

Time-series A time-series $\Gamma(w)$ of a word w is an ordered sequence of cosine similarities between the word vector at time k (v_w^k) and the previous one at time $k - 1$ (v_w^{k-1}):

$$\Gamma(w)_k = \frac{v_w^k \cdot v_w^{k-1}}{|v_w^k| |v_w^{k-1}|}$$

¹⁰<http://cwb.sourceforge.net/>

#	Source	Date	Left context	KWIC	Right context	Copy
1	unita	1948-01-01	Forze Aeree Israelite L aereo	pilotato	da ufficiali ebrei era diretto	
2	unita	1951-01-01	casa , su tm aereo	pilotato	da lo stesso comandante e	
3	unita	1951-01-01	I apparecchio , che era	pilotato	da il tenente Augusto Sb*rtoli	




#	Source	Date	Left context	KWIC	Right context	Copy
581	unita	2008-01-01	. Il presidente che ha	pilotato	gli Usaversodue conflitti da gli	
582	unita	2009-01-01	aveva parlato di « incidente	pilotato	e programmato » , a	
583	unita	2009-01-01	di Milano , che avrebbe	pilotato	un' asta giudiziaria per assegnare	

Figure 3: DE shows the KWIC (Keyword in the context) “pilotato”, shifted from meaning *driven* to meaning *manipulated*.

Diachronic Engine relies on word vectors computed by Temporal Random Indexing, but it is possible to integrate other approaches. In order to detect change points, we use the Mean Shift algorithm (Taylor, 2000). According to this model, we define a mean shift of a general time series Γ pivoted at time period j as:

$$K(\Gamma) = \frac{1}{l-j} \sum_{k=j+1}^l \Gamma_k - \frac{1}{j} \sum_{k=1}^j \Gamma_k \quad (1)$$

In order to understand if a mean shift is statistically significant at time j , a bootstrapping (Efron and Tibshirani, 1994) approach under the null hypothesis that there is no change in the mean is adopted. In particular, statistical significance is computed by first constructing B bootstrap samples by permuting $\Gamma(t_i)$. Second, for each bootstrap sample P , $K(P)$ is calculated to provide its corresponding bootstrap statistic and statistical significance (p-value) of observing the mean shift at time j compared to the null distribution. Finally, we estimate the change point by considering the time point j with the minimum p-value score. The output of this process is a ranking of words that potentially have changed meaning. Time-series is able to compare multiple

words at the same time and allows to filter words by time period.

3 Use cases

In this section, we describe two use cases concerning both historical and computational linguistics. DE is an extension of existing tools for synchronic corpora. It shares many of the use cases already available on those tools, such as applications in lexicography, terminology and linguistics.

Time series

Search of terrorismo from 1960-01-01 to 1984-01-01

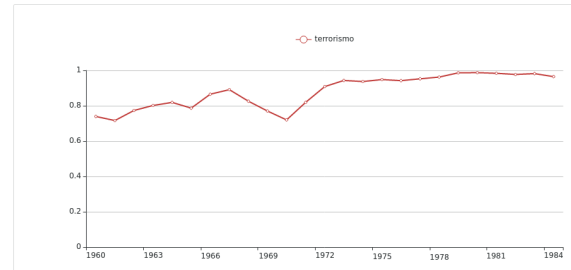


Figure 4: DE shows time-series of the word “terrorismo”.

3.1 Event detection through time-series

Lexical semantic changes can reveal aspects of real-world events, such as global armed conflicts (Kutuzov et al., 2017). DE provides several tools

to help events detection through time-series:

- the comparison of two time-series for highlighting potential correlations between lexical-semantic changes
- the plot of the time-series of cosine similarity between two word vectors over time, showing how the relatedness between two words changes over time
- the detected change points can bring out hidden information

In Figure 4, the time-series of “terrorismo” (*terrorism*) is shown. The time-series appears to be influenced by real-world events happening in Italy. In particular, we can observe a decrease in similarity starting in 1968 and culminating in 1970 during a crucial moment in Italy: “Anni di piombo” (*Years of Lead*), years marked by terrorism and violent clashes carried out by political activists.

3.2 Annotation of semantic shifts

The manual annotation of lexical-semantic shifts can be very expensive. Although robust frameworks (Schlechtweg et al., 2018) for the annotations already exist and are successfully used in evaluation tasks (Schlechtweg et al., 2020), no tools for facilitating the annotation are available yet.

DE can provide useful tools for the annotation of semantic shifts:

1. Frequencies over time can be preliminary exploited to filter words that have good coverage in the years under analysis;
2. Change points in time-series offer an overall and intuitive idea of the potential semantic shifts;
3. Diachronic concordances and collocations can support the identification of the type of change (Blank, 2012), such as when a word gains or loses a meaning.

4 Evaluation

We place a particular focus on the usability of our tool by giving a satisfactory experience. To understand the strength and weakness of the user interface, we conduct a preliminary usability test, according to the eGLU protocol (Simone et al., 2015). We use 21 participants. As a first step

of the evaluation, we want to test the system’s usability by measuring the task success rate: the ratio of users able to accomplish a set of predefined tasks. We ask participants to perform four tasks and we compute the average task success over all the 21 participants. During the evaluation, all participants complete their tasks without difficulties except for the showing frequency list task, where they had some problems with the corpus selection. We have already fixed this issue: the user is warned to choose a corpus from those available if no corpus is selected.

Results of the evaluation are reported in Table 1.

Task	Avg. task success
User registration	1
Login and show user information	1
Add a corpus	1
Show frequency list	.8095
Overall	.9523

Table 1: Results of the usability evaluation.

Moreover, we designed and dispensed a questionnaire for measuring user satisfaction. The questionnaire is composed of ten questions about the usability and the design of DE with a Likert scale of five values. The questionnaire results return an average score of 84.05/100. The system appear likeable to use.

5 Conclusions

In this paper, we present the Diachronic Engine, a tool for the analysis of lexical semantic change. DE integrates and extends current tools for corpus analysis enabling the study of corpus diachronic features. DE includes tools not included in other systems, such as time-series and change points detection based on diachronic word embeddings.

As future work, we plan to provide pre-loaded corpora such as Google Ngram, Diacoris (Onelli et al., 2006) and the integration of other approaches for computing diachronic word embeddings. Moreover, we plan to add a tool for the annotation of lexical-semantic shifts inspired by DUREL (Schlechtweg et al., 2018).

Acknowledgments

The authors would like to thank Dr. Ferrante and Dr. Lopatriello for supporting the preliminary development of the Diachronic Engine (Ferrante, 2019; Lopatriello, 2020). This research has been partially funded by ADISU Puglia under the post-graduate programme “Emotional city: a location-aware sentiment analysis platform for mining citizen opinions and monitoring the perception of quality of life”.

References

- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *First Italian Conference on Computational Linguistics CLiC-it (CLiC-it 2014)*. CEUR.org.
- Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the Italian language exploiting google ngram. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, page 56. CEUR.org.
- Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2019. Kronos-it: a dataset for the Italian semantic change detection task. In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR.org.
- Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. A diachronic Italian corpus based on “L’Unità”. In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020)*. CEUR.org.
- Andreas Blank. 2012. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, volume 285. Walter de Gruyter.
- Oliver Christ, Bruno M Schulze, Anja Hofmann, and Esther Koenig. 1999. The ims corpus workbench: Corpus query processor (cqp): User’s manual. *University of Stuttgart*, 8.
- Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470. Association for Computational Linguistics (ACL).
- Bradley Efron and RJ Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.
- Francesco Ferrante. 2019. Diachronic-engine: Un tool per la gestione dei corpora diacronici. B.Sc. degree Thesis in Metodi per il Ritrovamento dell’Informazione.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1489–1501.
- Bryan Jurish and Berlin-Brandenburgische Akademie der Wissenschaften. 2015. Diacollo: On the trail of diachronic collocations. In *Proceedings of the CLARIN Annual Conference*, pages 28–31.
- Adam Kilgariff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105:116.
- Adam Kilgariff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.
- Adam Kilgariff, Ondřej Herman, Jan Bušta, Vojtěch Kovář, et al. 2015. Diacran: a framework for diachronic analysis.
- Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36.
- Gabriele Lopatriello. 2020. Diachronic engine: A tool for the management of diachronic corpora. Master Thesis in Intelligent Information Access and Natural Language Processing.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.
- Corinna Onelli, Domenico Proietti, Corrado Seidenari, and Fabio Tamburini. 2006. The diacoris project: a diachronic corpus of written Italian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1212–1215.
- Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.
- Pavel Rychlý. 2008. A lexicographer-friendly association score. *RASLAN 2008 Recent Advances in Slavonic Natural Language Processing*, page 6.
- Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering*.
- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of

- lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.
- Borsci Simone, Boscarol Maurizio, Cornero Alessandra, et al. 2015. Il Protocollo eGLU 2.1. Il Protocollo eGLU-M. Come realizzare test di usabilità semplificati per i siti web ei servizi online delle PA. Glossario dell’usabilità.
- Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Wayne A Taylor. 2000. Change-point analysis: a powerful new tool for detecting changes.
- Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM 2018)*, pages 673–681.

Hate Speech Detection with Machine-Translated Data: The Role of Annotation Scheme, Class Imbalance and Undersampling

Camilla Casula
Fondazione Bruno Kessler
Trento, Italy
ccasula@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

While using machine-translated data for supervised training can alleviate data sparseness problems when dealing with less-resourced languages, it is important that the source data are not only correctly translated, but also follow the same annotation scheme and possibly class balance as the smaller dataset in the target language. We therefore present an evaluation of hate speech detection in Italian using machine-translated data from English and comparing three settings, in order to understand the impact of training size, class distribution and annotation scheme.¹

1 Introduction

The task of detecting hate speech on social media has been attracting increasing attention due to the negative effects this phenomenon can have on online communities and society as a whole. The development of systems which can effectively detect hate speech has therefore become increasingly important for academics and tech companies alike.

One of the difficulties of producing accurate hate speech detection systems is the need for large, high-quality datasets, the creation of which is time and resource-consuming. English can count on the highest number of hate speech detection datasets, as well as the ones with the largest sizes, with up to 150k posts for a single dataset (Gomez et al., 2020). Other languages such as Italian, on the other hand, can count on fewer datasets which tend to be smaller (Vidgen and Derczynski, 2020). Given that machine learning methods are typically used for this task, the use of small datasets can lead to overfitting problems due to the lack of linguistic variation (Vidgen and Derczynski, 2020).

One possible solution to alleviate data sparseness is the use of machine translated data from English to less resourced languages for training classifiers, exploiting the large amount of data available for English. This has already been used in the context of hate speech detection (Sohn and Lee, 2019; Casula et al., 2020) but results have not been consistent across languages.

An additional issue is the fact that there is no shared fixed definition within the NLP community of what type of language constitutes hate speech. Indeed, there are typically large differences among hate speech and abusive language datasets in terms of annotation frameworks and their applications in practice (Caselli et al., 2020). In addition to this, there can be large variations between datasets in terms of size and class balance. Possible issues affecting the behaviour of classifiers trained on machine-translated data, such as different class distribution in source and target language, or different annotation scheme, have not been analysed.

In order to fill this gap, we explore the impact of these differences between datasets when performing hate speech detection in Italian using machine-translated data from English. Our goal is to address the three following questions:

- What performance can we expect by using only machine translated data, given that translation quality for social media language may be problematic?
- Is it better to use a larger translated set for training, even by merging slightly different classes, or a smaller, more precise one?
- What is the impact of class imbalance, and to what extent can undersampling be effective?

The above questions are addressed by comparing three experimental settings that are described in Section 4 and evaluated in Section 5.

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Related Work

In recent years, the number of research works focused on the detection of hate speech on social media has remarkably increased, mostly due to the growing awareness regarding the societal impact these platforms can have.

Computational methods for detecting the presence of hate speech on the web have become necessary due to the extremely large amounts of user-generated content being posted each day. These methods typically rely on supervised learning, in the form of both traditional machine learning (e.g. support vector classifiers) and deep learning approaches (Schmidt and Wiegand, 2017). Given the increased attention towards this topic, more and more shared tasks regarding hate speech and abusive language detection have emerged, such as the HaSpeeDe task at Evalita 2018 (Bosco et al., 2018), OffensEval (Zampieri et al., 2019) and HatEval (Basile et al., 2019) at SemEval 2019, and the multilingual OffensEval at SemEval 2020 (Zampieri et al., 2020).

Systems based on Transformers architectures such as BERT (Devlin et al., 2019) have proven effective for hate speech detection and classification in both English (Zampieri et al., 2019) and Italian (Polignano et al., 2019a). These systems are generally pre-trained on large unlabeled corpora through two self-supervised tasks (next sentence prediction and masked language modeling) to create language models which can then be fine-tuned to a variety of downstream tasks using labeled data.

AIBERTO (Polignano et al., 2019b) is a BERT-based system which was pre-trained on Italian Twitter data, and it currently defines the state of the art for hate speech detection in Italian (Polignano et al., 2019a).

Recently, more attention has been directed towards the quality of hate and abuse detection systems. Vidgen et al. (2019) investigate the flaws presented by most abusive language detection datasets in circulation: they can contain systematic biases towards certain types and targets of abuse, they are subject to degradation over time, they typically present very low inter-annotator agreement, and they can vary greatly with respect to quality, size, and class balance. Vidgen and Derczynski (2020) further analyse the role of datasets in the detection of abuse, addressing issues such as the use of different task descriptions and annotation

schemes across corpora, as well as similar annotation schemes being applied in different ways.

3 Data

Since tweets containing hate speech or abusive language constitute a very small subset (between 0.1% and 3% depending on the label used) of all tweets being posted (Founta et al., 2018), random samples are generally not used for annotation, because the final datasets would contain an extremely low number of positive class examples, which would make classification difficult. The typical solution to this is to preselect posts that are likely to contain hateful language by searching for specific hate-related keywords. While this method is effective for gathering more instances of hate speech, it can make datasets biased, which is a main issue in hate speech datasets (Wiegand et al., 2019).

The dataset we chose for training our system is described in Founta et al. (2018). This dataset was not created starting from a set of predefined offensive terms or hashtags in order to reduce bias, which was an important factor in our choice. The method used by Founta et al. (2018) to increase the percentage of hateful/abusive tweets is boosted random sampling, in which a portion of the dataset is “boosted” with tweets that are more likely to belong in the minority classes. The boosted set of tweets is created using text analysis and machine learning (Founta et al., 2018).

The dataset was annotated through crowdsourcing using the labels *hateful*, *abusive*, *spam*, and *normal*. The definition of *hate speech* given by Founta et al. (2018) to the annotators, based on existing literature on the topic, is:

Hate Speech: Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.

The *abusive* label, on the other hand, is the result of three separate labels (*abusive*, *offensive*, and *aggressive*) being combined. In preliminary annotation rounds, Founta et al. (2018) found that these three labels were significantly correlated, so they grouped them together. The definition of *abusive language* given to the annotators is:

Abusive Language: Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion.

While the Founta et al. (2018) dataset was originally comprised of 80k tweets, Twitter datasets can often be subject to degradation due to tweets being removed over time and not accessible anymore through tweet IDs (Vidgen et al., 2019). After retrieving all available tweets and after removing tweets annotated as spam, the total number of tweets we use for training is 12,379, of which 727 are annotated as hateful and 1,792 as abusive. Before translating the data into Italian, we preprocess it using the Ekphrasis tool² to tokenise the text and normalise user mentions, URLs (replaced by `<user>` and `<url>` respectively), as well as numbers, which are substituted with a `number` tag. We then use the Google Translate API to translate the data into Italian, in order to use it as training data for our classifier.

For testing, we use the test portion of the Twitter dataset used in the Hate Speech Detection (HaSpeeDe) task at Evalita 2018 (Bosco et al., 2018), consisting of 1,000 Italian tweets manually annotated for hate speech against immigrants. This dataset is a simplified version of the dataset described in (Sanguinetti et al., 2018), in which more fine-grained labels are used.

4 Experimental Setup

We experiment with the fine-tuning of AIBERTO (Polignano et al., 2019b), a BERT-based language model pre-trained on Italian Twitter data, using data that was automatically translated from English. This model has achieved state-of-the-art results when fine-tuned on the training data from the HaSpeeDe task at Evalita 2018 (Polignano et al., 2019a).

Our goal is that of exploring the impact of different annotation schemes and class balance when using machine-translated data for hate speech detection. Indeed, merging fine-grained classes into coarser ones has been a common and accepted practice when creating larger training sets from a smaller one (e.g. Founta et al. (2019)). This step has been performed also to compare classification in different languages (Corazza et al., 2020).

²<https://github.com/cbaziotis/ekphrasis>

In order to investigate this, we compare three different experimental settings. In the first one, we fine-tune AIBERTO on the translated tweets in Founta et al. (2018) after merging the *hateful* and *abusive* classes together, mapping them to a single hateful class as required by the binary classification task at Evalita 2018. In a second setting, AIBERTO is fine-tuned on the *hateful* class alone, discarding all tweets annotated as *abusive* in Founta et al. (2018). We hypothesize this setting may perform better when tested on the HaSpeeDe data, given the higher similarity in annotation framework.

Simply removing tweets annotated as abusive, however, can throw off the balance between classes. More specifically, when training the system on both abusive and hateful tweets the hateful+abusive class constitutes about 20% of our data, while when we only use tweets annotated as hateful this percentage drops to 7%, potentially affecting classification results. In particular, the data we use for testing has a different class balance, with 30% of tweets marked as hateful. In order to assess the impact of class imbalance on our results, we further evaluate each setting using undersampling (Kubat, 2000; Sun et al., 2009), a technique typically used for imbalanced classification, in which we reduce the number of tweets belonging to the majority class, so that the overall percentage of tweets containing hate increases.

Given that undersampling our data reduces the total size of tweets available for training, the resulting datasets for each annotation scheme considerably differ in size. We therefore consider a third setting, in which we use further random undersampling (Kubat, 2000; Sun et al., 2009) to match the larger dataset (hateful+abusive) with the smaller one (hateful only), so that the two annotations can be effectively compared in a setting with equal class balance and sample size.

In summary, the three data settings we train our system on are:

1. Hateful and abusive tweets, using undersampling to progressively lower class imbalance;
2. Hateful only tweets, again using undersampling to progressively lower class imbalance;
3. Hateful and abusive tweets, both using undersampling to progressively lower class imbalance as in the previous settings, and using

further random undersampling to match the low sample sizes of setting 2.

Our AIBERTo fine-tuning architecture consists of a pooling layer for extracting the AIBERTo hidden representation for each sequence, followed by a dropout layer (dropout rate 0.2), two dense layers of size 768 and 128 and, finally, a softmax layer. We use L2 regularization ($\lambda=0.01$), Adam optimizer ($2e-5$ learning rate), and categorical cross-entropy loss. We train the system for 5 epochs with batch size 32.

5 Results and Discussion

We measure the classification results using both macro-F1 score and minority class F1 score. We repeat each run five times in order to compensate for random initialization, and we report the average scores of these runs.

5.1 Setting 1: Hateful + Abusive Tweets

The classification results obtained when fine-tuning AIBERTo on both abusive and hateful tweets combined can be observed in Table 1.

Setting 1: Hateful + abusive			
% hate	Size (tweets)	Macro-F1	Hate class F1
20%	12,379	0.40	0
30%	8,397	0.64	0.52
40%	6,298	0.63	0.57

Table 1: Scores obtained when fine-tuning AIBERTo on both hateful and abusive tweets.

The class balance of the dataset prior to undersampling is 20% hateful + abusive tweets and 80% non-hateful, which amounts to 12,379 tweets total. With this class balance, the system performs the worst, classifying every tweet as belonging to the majority non-hateful class. On the other hand, with a higher percentage of minority class instances, the classification results improve, in spite of the considerably smaller amount of training data available. These results suggest that consistency in class balance can play a bigger role than training data size in classification results in this context.

5.2 Setting 2: Hateful Only Tweets

The performance of the system when fine-tuned on tweets labeled as hateful only is reported in Table 2. As previously mentioned, only 7% of tweets in the dataset we use are labeled as hateful. The

classes are therefore extremely imbalanced before undersampling. Predictably, with the classes being this imbalanced, the system identifies all test instances as belonging to the majority class. This again happens with the minority class comprising 20% of the training data.

Setting 2: Hateful only			
% hate	Size (tweets)	Macro-F1	Hate class F1
7%	10,587	0.40	0
20%	3,635	0.40	0
30%	2,423	0.65	0.54
40%	1,818	0.52	0.56

Table 2: Scores obtained when fine-tuning AIBERTo on tweets labeled as hateful only.

Similarly to Setting 1, the best classification performance in this case is achieved with 30% of minority class tweets. Interestingly, the best performance is comparable to the one obtained in Setting 1, even though in this case the number of training samples available is much lower, suggesting that more task-specific training instances can impact performance. We can note a difference with the minority class at 40% of total data, in which the performance drops in terms of macro-F1 score, likely due to the very small number of samples available for training and the consequent lack of linguistic variation. The hate class F1 score, however, remains stable.

State-of-the-art results obtained by fine-tuning AIBERTo on the same Evalita dataset as reported in Polignano et al. (2019a) reach 0.80 macro-F1 and 0.73 F1 on the hate class, which we can consider an upper-bound for our task, obtained in a fully-supervised monolingual setting. On the other hand, the most frequent label baseline is 0.40 macro-F1, which is clearly outperformed using only machine-translated data.

5.3 Setting 3: Hateful + Abusive Tweets (Random Undersampling)

Since there are large differences in size between the hateful+abusive annotation and the hateful-only annotation, we randomly undersample the hateful+abusive training data so that it matches the size of the hateful-only training data, in order to allow us to effectively compare the impact of each annotation framework on our results. The classification performance is reported in Table 3.

If we compare the results of Setting 3 with those of Setting 2, it is clear that using more task-

Setting 3: Hateful + abusive (random undersampling)

% hate	Size (tweets)	Macro-F1	Hate class F1
30%	2,423	0.58	0.38
40%	1,818	0.59	0.51

Table 3: Scores obtained when fine-tuning ALBERTo on tweets labeled as hateful and abusive, after random undersampling.

specific data, in this case hateful-only tweets, can lead to a larger improvement in performance when the amount of training data is the same. This suggests that consistency in annotation between training and test data can have a positive impact on classification, although it is not fundamental to help classification of hate speech detection with machine translated data. In fact, other aspects such as class balance can also play an important role.

5.4 Qualitative Analysis

Another aspect affecting classification, which we have not considered so far, is the quality of machine translation, a particularly challenging task on social media data (Michel and Neubig, 2018). In order to assess the impact of translation quality on our results, two annotators with linguistic background manually analysed 500 samples from the training data, consisting of 300 tweets annotated as normal, 100 as hateful, and 100 as abusive. Each annotator checked manually 250 random tweets from this sample. Translation quality was evaluated using the semantic adequacy annotation scheme proposed in Dorr et al. (2011, p. 807). Annotations are judged on a scale between -3 and 3, with scores below 0 for inadequate translations and above 0 for adequate ones. The averaged annotations for each class are reported in Table 4.

	Normal	Hateful	Abusive	Overall
Average	0.438	0.527	-0.043	0.368

Table 4: Average translation quality scores.

Overall, translations tend towards adequacy, but the average scores are below 1 for all classes. Interestingly, tweets annotated as abusive show poorer translation quality than other classes. This could help explain the small differences in classification performance between our experiments.

A major role is played in this context by profanities, which are often used to offend a target but can also appear in non derogatory messages exchanged among members of the same community

(Pamungkas et al., 2020). In the case of abusive tweets, we observe that the offenses are less direct and therefore slurs tend to be translated poorly. See for example the following sentence, which is labeled as abusive in the Founta et al. (2018) dataset:

- (1) use that ugly ass design [...]
 utilizzare quel disegno asino brutto [...]
use that design donkey ugly [...]

Here, “ass” is translated with “asino” (“donkey”), effectively removing the profanity in the translated tweet and changing completely the meaning of the message.

On the other hand, when profanities are used in a more direct way, or when they are expressed through unambiguous words such as “idiot” and “stupid”, they tend to be translated correctly, contributing to a correct classification. Example 2 shows a hateful tweet which was translated almost correctly, retaining its offensiveness in the target language.

- (2) what happens when you put idiots in charge
 cosa succede quando si mette idioti in carica

6 Conclusions

In this paper we analysed the impact of machine-translated data on Italian hate speech detection in a zero-shot setting. Our experiments show that when using machine-translated data for training it is possible to learn a classification model that clearly outperforms the most-frequent baseline, even if translation quality is affected by the jargon used in social media data. We found that using more task-specific data can have a positive impact on classification performance even with lower sample sizes compared to larger, less targeted datasets.

Consistency in class distribution of training and test data can have a bigger impact than the size of the training set, or the annotation scheme. Indeed, using only the original training set translated into Italian, without undersampling, classification performance would be poor.

In the future, we plan to extend this kind of evaluation to new language pairs and new datasets, to check whether the findings obtained on the English – Italian pair are confirmed also with other languages.

References

- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9, Turin, Italy. CEUR.
- Tommaso Caselli, Valerio Basile, Jelena Mitrovic, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6193–6202. European Language Resources Association.
- Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Fbk-dh at semeval-2020 task 12: Using multi-channel bert for multilingual offensive language detection. In *Proceedings of Offenseval*.
- Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Techn.*, 20(2):10:1–10:22.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Bonnie J Dorr, Joseph Olive, John McCary, and Caitlin Christianson. 2011. *Machine Translation Evaluation and Optimization*, pages 745 – 843. Springer New York.
- Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media*.
- Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science, WebSci ’19*, page 105–114, New York, NY, USA. Association for Computing Machinery.
- Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, 03.
- M. Kubat. 2000. Addressing the curse of imbalanced training sets: One-sided selection. *Fourteenth International Conference on Machine Learning*, 06.
- Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6237–6246. European Language Resources Association.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019a. Hate speech detection through alberto italian language understanding model. In *NL4AI@ AI* IA*.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019b. AIBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International*

- Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.
- Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.
- Y. Sun, A. Wong, and M. Kamel. 2009. Classification of imbalanced data: a review. *Int. J. Pattern Recognit. Artif. Intell.*, 23:687–719.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, abs/2004.01670.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffenseEval 2020). In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

UDante: First Steps Towards the Universal Dependencies Treebank of Dante’s Latin Works

Flavio M. Cecchini, Rachele Sprugnoli, Giovanni Moretti, Marco Passarotti

CIRCSE Research Centre, Università Cattolica del Sacro Cuore

Largo Agostino Gemelli 1, 20123 Milano

{flavio.cecchini, rachele.sprugnoli,
giovanni.moretti, marco.passarotti}@unicatt.it

Abstract

English. This paper¹ presents the early stages of the development of a new treebank containing all of Dante Alighieri’s Latin works. In particular, it describes the conversion of the original TEI-XML files to CoNLL-U, the creation of a gold standard, the process of training four annotators and the evaluation of the syntactic annotation in terms of inter-annotator agreement and LA, UAS and LAS. The aim is to release a new resource, in view of the celebrations for the 700th anniversary of Dante’s death, which can support the development of the *Vocabolario Dantesco*.

1 Introduction

The research field of treebanking (i.e. the building of corpora enhanced with syntactic metadata) has evolved substantially since the time when the first large-scale syntactically annotated corpus, the Penn Treebank for English, was published between the late Eighties and the early Nineties (Taylor et al., 2003). Across the last two decades, the range of languages for which a treebank is available has increased considerably. The grammar framework behind the most widespread annotation style currently used in treebanking has also changed: treebanks annotated according to various styles of dependency grammars have been increasingly outnumbering those based on constituency (or phrase-structure) grammars, as demonstrated by the current status of the *Universal Dependencies* initiative (UD) containing more than 160 treebanks and 90 languages which follow the same, dependency-based, annotation style (Nivre et al., 2016).

The set of textual genres covered by currently available treebanks is quite diverse. While the first corpora were built mostly collecting texts from news, the last decade has seen a substantial growth of treebanks of different genres, including literary texts, mostly written in ancient or historical languages.² The first available treebanks for ancient languages were those for Ancient Greek and/or Latin, namely the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2019) and the Ancient Greek and Latin Dependency Treebank from the Perseus digital library (Bamman and Crane, 2011). With regard to Latin, the available treebanks in UD cover just a minimal subset of the Latin texts that have survived the centuries and which show a wide diversity, mostly due to Latin’s *lingua franca* role played all over Europe up until the 1800s (Leonhardt, 2009). So far, the treebanks for Latin include only portions of the Classical and Late Latin canon of texts (Perseus and PROIEL (Eckhoff et al., 2018)), a set of Early Medieval charters from Tuscia (Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011; Cecchini et al., 2020)) and a selection of Late Medieval philosophical-theological texts by Thomas Aquinas (IT-TB), for a total of more than 800 000 nodes.

Among the many Latin texts that still lack syntactic annotation are those by Dante Alighieri (1265-1321). Given the importance of Dante in the history of Italian literature (and beyond) and in the light of the celebrations for the upcoming 700th anniversary of his death, we have started a project (called *UDante*) aimed at performing a UD-compliant syntactic annotation of all his Latin texts. The syntactic annotation of Dante’s *opera omnia* in Latin fits into the larger project of the *Vocabolario Dantesco*, which aspires to provide a detailed description of the entire (both Vulgar and

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²Examples among the UD treebanks are the Kyoto treebank of Classical Chinese (Yasuoka, 2019) and the *Scriptorium* treebank of Coptic (Zeldes and Abrams, 2018).

Latin) lexicon of Dante Alighieri.³ Indeed, during the composition of entries for the vocabulary, lexicographers will benefit from having the possibility to run syntactic queries on Dante’s works.

The choice of using the UD formalism in the *UDante* project is motivated by a number of benefits implied by the inclusion of a new set of annotated texts into such a large collection of treebanks sharing the same annotation style, among others the use of the several tools developed by the UD community with the goal of querying, editing, visualizing and (automatically) processing the (meta)data of the treebanks.⁴ Particularly, a remarkable added value is the possibility to run common queries on the almost 100 different languages provided with at least one treebank in the current version of UD (v2.6, released on May 15th, 2020). Furthermore, adopting a well known and widely used data format (CoNLL-U) and part-of-speech tagset (UPOS) fosters the dissemination and use of a treebank of Dante’s Latin works in the community of computational linguistics, laying the foundation for a closer collaboration with that of Italian philology and, more generally, with scholars in the Humanities, leading to a mutual benefit.

This paper presents the process behind the development of the manually annotated UD treebank containing the full collection of Latin works of Dante Alighieri. More specifically, we describe the conversion of the original TEI-XML files into the CoNLL-U format, we give details on the creation of a gold standard⁵ and we report on the training of four annotators with no previous knowledge of the UD formalism, providing an evaluation of their annotation work.

2 Treebank Development

The texts of the Latin works by Dante Alighieri (*De monarchia*, *De vulgari eloquentia*, *Eclogues*, *Epistulae* and *Quaestio de aqua et terra*) are made available by the *DanteSearch* corpus (Tavoni, 2012).⁶ All texts come already manually lemmatised and morphologically tagged by a team of young scholars at the University of Pisa, and are encoded in TEI-XML.⁷ The original files are

converted into the CoNLL-U format⁸ and then revised and syntactically annotated using ConlluEditor (Heinecke, 2019).

2.1 From TEI-XML to CoNLL-U

We implement an own developed script to automatically convert the TEI-XML files of the *DanteSearch* corpus into the CoNLL-U format. First of all, the script analyses the XML tag structure to identify the internal organisation of the text (i.e. the division of the work in books, chapters etc.): this information is stored in the `MISC` field so as to facilitate the recoverability of the original structure of the text starting from the CoNLL-U file. Then, sentences are split and the tag `<LM>`, which for each token contains morphological information, is parsed in order to extract lemma, part of speech and morphological traits, and to convert the codes used in *DanteSearch* into UPOS tags⁹ (originally inspired by (Petrov et al., 2012)) and UD features respectively. An example is:

```
<LM lemma="resono"
catg="valcis3">resonaret</LM>
```

In particular, the part-of-speech tag and the morphological traits are derived from the values of the `catg` attribute, while those fields of the CoNLL-U format dedicated to syntactic information are filled with underscores (`_`) and left for manual annotation. The conversion of `catg` attribute is challenging, because the string-type values of its slots do not follow a fixed-position strategy, thus the string ends up having a variable length, and the same morphological trait can occupy different positions according to the given part of speech. In general, UD requires a more fine-grained annotation of morphological traits compared to the one originally provided by the *DanteSearch* corpus. For example, the value `valcis3` of *resonaret* (active imperfect subjunctive third-person singular of *resono* ‘to resound’) is converted into the UD formalism as follows:

```
v → VERB
a → Voice=Act
```

tagger for XIIIth century Italian by means of TreeTagger and the Stanford POS tagger (Basile and Sangati, 2016).

⁸<https://universaldependencies.org/format.html>

⁹<https://universaldependencies.org/u/pos/index.html>

³<http://www.vocabolario.dantesco.it>

⁴See the website of UD for the list of tools: <https://universaldependencies.org>.

⁵<https://github.com/CIRCSE/UDante>.

⁶<https://dantesearch.dantenetwork.it>

⁷The *DanteSearch* corpus also contains the Vulgar works by Dante Alighieri that were used to develop a part-of-speech

`1` → VerbClass=LatA¹⁰
`ci` → Aspect=Imp|Mood=Sub|Tense=Past
 |VerbForm=Fin
`s` → Number=Sing
`3` → Person=3

Ad hoc rules are added to cover specific cases. For example, in *DanteSearch* the lemma *prius* ‘before’ is marked only with the grammatical category `r`: a rule converts `r` into the UPOS tag `ADV` and adds the morphological feature `Degree=Cmp` (comparative degree).

Annotators, in addition to annotating syntax from scratch, have to check the correctness of the automatic conversion and to manually modify or add items not covered by it. For example, annotators have to: (i) modify the grammatical category of population names (such as *Veronenses* ‘inhabitants of Verona’), which are marked as proper names in *DanteSearch*, contrary to UD recommendations, for which they should be considered as adjectives;¹¹ (ii) check the ambiguous case of some pronouns in the neutral gender which in *DanteSearch* have mistakenly been marked as nominative instead of accusative (e. g. *quod* ‘that’), or viceversa; (iii) disambiguate the `PronType` feature in the case it has more than one value: this happens because the types of pronouns in *DanteSearch* cannot always be matched to only one `PronType` value (e. g. *quis* ‘who/any’, interrogative or indefinite).

2.2 Gold Standard Creation

An important part of the *UDante* project consists in training a group of annotators on the formalism of UD with the goal of providing them with adequate competences to pursue the complete syntactic annotation of Dante’s works. To this aim, for each of the two parts of our training (Section 2.3) a small number of sentences is singled out from all across Dante’s Latin texts to be used as a common (first part) or individual (second part) benchmark for the assessment of the annotators’ progress.

The first part of the training makes use of 33 sentences out of the total 1 662 (corresponding to

950 tokens out of 55 666). These sentences are not chosen to be consecutive, nor do they follow a particular order, but they are allocated into three different groups of increasing complexity, corresponding to the three distinct phases of this part of the training. The distribution is of 15 sentences in the first, introductory group, 5 in the second, intermediate group and 10 in the third, more challenging group. The first two groups are rather homogeneous and mostly draw from the *De vulgari eloquentia*, while in the third one each work is represented by 2 sentences, and the *Eclogues* are featured only here.

The differences in complexity can be understood in terms of number of nodes, depth, and breadth of the resulting syntactic trees. While a sentence of the first group has a median number of 11 nodes, a median depth of 4 layers and most nodes (not counting the root) tend to be at depth ca. 3, for the second group the same figures are respectively 42, 7 and ca. 4; for the third group they are 46.5, 7.5 and ca. 4.5. The difference is especially marked between the first group and the other two. Besides such quantitative factors, other more qualitative ones, like difficult syntactic structures, contribute to the overall complexity.

As for the second part of our training, which consists of only one phase, textual cohesion substitutes increasing complexity as the main selection criterion: as such, each annotator is assigned, from the work they will respectively take care of, the first 10 sentences which have not been previously annotated. The complexity of the single sentences is thus more variable in this phase, but still well represents the whole corpus. In particular, we use sentences 1-4 and 7-12 from book I of the *De monarchia*, sentences 4-5 and 7-14 from book I of the *De vulgari eloquentia*, sentences 1-10 from the first of Dante’s *Eclogues*, and 1-10 from Epistle VII of the *Epistulae*. The *Quaestio de aqua et terra*, of uncertain authorship, is not assigned to any annotator at the moment.

All the selected sentences are priorly syntactically annotated by hand by a UD expert applying language-specific features and subrelations developed for Latin, while lemmas, parts of speech and morphological traits are corrected or enhanced where needed with respect to the CoNLL-U conversion (see Section 2.1). This way, on the one hand a tripartite, scaled gold standard is created for common evaluation, while on the other hand

¹⁰We add VerbClass as a language-specific feature to encode traditional verb conjugations. The value LatA indicates the first conjugation, which has thematic vowel ‘a’.

¹¹From <https://universaldependencies.org/u/pos/ADJ.html>: “ADJ is also used for ‘proper adjectives’ such as European (‘proper’ as in proper nouns, i. e., words that are derived from names but are adjectives rather than nouns).”

each annotator will be tested on an individual gold standard in the last phase (Section 2.4).

2.3 Tripartite Training Process and Control

The training of the annotators (all with no background in treebanking, but provided with a solid knowledge of Dante’s works and academic background in Latin and Italian philology) is split into two main parts: three “training proper” phases (phase 1 to 3), and one further “control” phase (phase 4).

The first part is meant to lay out a common training ground where the annotators can learn the specifics of the UD annotational scheme, and their progress is overseen and periodically reviewed to prompt improvements. In the first phase, the basics of the UD formalism are presented, and the annotators are required to manually annotate a first group of sentences as a way to evaluate their understanding of the UD principles.¹² In the second and third phases, various aspects of the performed annotation get to be discussed and more complex syntactic structures are introduced, each time assigning new, more challenging sentences for an overall evaluation of the annotator’s performance and their inter-annotator agreement (see Section 2.2). At every step, the focus is primarily on the syntactic level, since most aspects regarding lemmatisation, parts of speech and morphology are already mostly dealt with during the conversion phase (Section 2.1).

In contrast to the first three phases, the last, control phase is carried out individually for each annotator on separate sets of sentences (see Section 2.2), as a prelude to their actual annotation work.

	Phase 1	Phase 2	Phase 3
EDGES	80%	83%	79%
DEPRELs	84%	92%	91%

Table 1: Inter-annotator agreement.

2.4 Evaluation and Analysis

Table 1 reports the overall inter-annotator agreement (IAA) for each of the first three phases in terms of Fleiss’ kappa,¹³ with regard to the struc-

¹²<https://universaldependencies.org/guidelines.html>

¹³Our script for calculating IAA on CoNLL-U files is available at https://github.com/johnnymoretti/CoNLL-U_Fleiss_Kappa.

ture of syntactic trees (EDGES) and the choice of dependency relations (DEPRELs), whereas in Table 2 the correctness of the annotator’s analyses are compared for all phases to the gold standard according to label accuracy (LA), unlabelled attachment score (UAS) and labelled attachment score (LAS) (Buchholz and Marsi, 2006). Table 3 presents the macro-average F-measure on the assignment of dependency relations,¹⁴ again for all four phases. Both these scores and the IAA are computed over basic relations only, i.e. disregarding any subrelation (e.g., the dependency relations *obl*, *obl:agent* and *obl:arg* all count as *obl*) so we can focus on the syntactic soundness of the annotations, since more specific subrelations are often related to secondary language-specific, lexical and semantic factors.

For what concerns the IAA, the scores are rather good (always >75%) and, together with the equally positive scores in Table 2, show that the basic principles of UD have been uniformly acquired by all the annotators during the first part of the training, especially the UD scheme of dependency relations. In general, going from phase 1 to phase 3, we notice that all scores are quite stable, and we only observe a slight decrease of the EDGES score in phase 3 which mirrors the noteworthy complexity of the corresponding test sentences (see Section 2.2); the same general decrease shown in Table 2. However, this is more than compensated by generally markedly improved scores for all annotators in phase 4: taking into account the greater variability in sentence complexity, these data show that all annotators have reached a good degree of confidence both with UD’s syntactic formalism and with the specific annotation guidelines developed for Latin, which have been constantly updated during this project.

In particular, if we consider only the labelling of single nodes (LA), we register a decided mean improvement in the last phase (89% vs. 79.75%), showing that the annotators have factually improved their assessment of syntactic dependency relations. A similar trend for IAA in the first three phases and quite close scores in Table 3 point to the fact that those cases where annotators disagree are also those for which they have greater uncertainties at the syntactic level; this leads us to con-

¹⁴Metrics calculated with MaltEval (Nilsson and Nivre, 2008).

	Phase 1			Phase 2			Phase 3			Phase 4		
	LA	UAS	LAS	LA	UAS	LAS	LA	UAS	LAS	LA	UAS	LAS
Ann1	91%	96%	87%	89%	80%	75%	82%	70%	63%	87%	78%	73%
Ann2	72%	65%	54%	87%	77%	70%	83%	75%	69%	91%	91%	85%
Ann3	78%	83%	72%	86%	84%	78%	83%	76%	70%	92%	87%	82%
Ann4	78%	83%	72%	86%	84%	78%	79%	76%	68%	86%	85%	77%

Table 2: Annotators’ performances versus gold standard.

	Phase 1	Phase 2	Phase 3	Phase 4
	F	F	F	F
Ann1	86%	80%	71%	70%
Ann2	54%	69%	73%	86%
Ann3	69%	70%	72%	72%
Ann4	69%	70%	60%	67%

Table 3: Macro-average F-measure on dependency relations with respect to the gold standard.

clude that most errors might stem from the same sources. In particular, while basic core relations, especially for nominals (*nsubj*, *obj*), and the choice of the *root* all score well, we observe most discrepancies, persisting through all phases, with regard to the labelling of clausal dependents, such as *advcl* and notably clausal complements (*ccomp*, *xcomp*). This pairs with minor confusions regarding the labelling and the attachment of connective elements, i.e. both co-ordinating and subordinating conjunctions. These persistent difficulties are reflected by Table 3, which, as a macro-average that does not take into account the actual frequencies of single dependency relations, has lower scores than LA in Table 2. Considering that the array of syntactic relations in the later phases is much more varied than in the first one, we still observe a quite stable, if not slightly improving, trend.

The decrease of UAS and LAS in the third phase, when compared to the good results of the second phase, has to be expected, as the sentences of phase 3 are chosen to be particularly challenging and in some cases present open problems of syntactic annotation.¹⁵ Despite this, the differences between phase 1 and phase 3 still show a rather stable quality of the annotation from this angle. Then again, the last control phase registers much

improved performances also for UAS and LAS, displaying the good level of assurance reached by the annotators at all levels of annotation.

3 Conclusion and Future Work

In this paper, we describe the preliminary steps towards the creation of a UD-compliant treebank of the Latin works by Dante Alighieri. To this end, we create a gold standard and we train and evaluate the work of a team of four annotators by means of a tripartite common set of sentences of increasing complexity annotated by a UD expert, complemented by specific gold standards for each annotator in a final control phase before the actual annotation work takes place.

Besides supporting the objectives of the *Vocabolario Dantesco* project, the development of a treebank based on Dante’s Latin works also serves a wider scope, i.e. the inclusion of these latters into the *LiLa Knowledge Base*, which makes distributed linguistic resources for Latin interoperable through the Linked Data paradigm (Passarotti et al., 2020).¹⁶ At the same time, the efforts put into this project will hopefully bring forth some much-needed recommended guidelines for the UD-style annotation of Latin.

The complete annotation of Dante’s Latin works will provide the community with a new, manually annotated dataset of higher quality than any automatic system. Table 4 reports LAS scores computed on the sentences of our gold standard and processed with UDPipe using the UD v2.5 models for Latin (Straka and Straková, 2017). The scores clearly show that current models are not good enough to parse the Latin of Dante.

	IT-TB	Perseus	PROIEL
LAS	40.83%	24.93%	29.98%

Table 4: UDPipe scores (based on UD v2.5) for gold standard sentences (all four phases).

¹⁵See for example the open issue on how to deal with singular subjects and plural copula at <https://github.com/UniversalDependencies/docs/issues/714>.

¹⁶<https://lila-erc.eu>

The addition of Dante’s Latin works into the thriving and expanding UD project and the newly acquired possibility to interact with a large number of other Latin texts of different genres and time periods makes us hope for a breakthrough of the world of treebanking into the wider community of the Humanities, which today can benefit from accessing a huge set of connected textual (meta)data like never before.

Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme via the “LiLa: Linking Latin” project - Grant Agreement No. 769994. The authors also want to thank Prof. Mirko Tavoni (University of Pisa) and the annotators Martina De Laurentis, Federica Favero, Giulia Pedonese and Elena Vagnoni.

References

- David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer.
- Angelo Basile and Federico Sangati. 2016. D (h) ante: A New Set of Tools for XIII Century Italian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2825–2828.
- Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.
- Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020. A new latin treebank for universal dependencies: Charters between ancient latin and romance languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 933–942.
- Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.
- Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93.
- Timo Korkiakangas and Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.
- Jürgen Leonhardt. 2009. *Latein. Geschichte einer Weltsprache*. Beck.
- Jens Nilsson and Joakim Nivre. 2008. MaltEval: an Evaluation and Visualization Tool for Dependency Parsing. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666.
- Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212.
- Marco Passarotti. 2019. *The Project of the Index Thomisticus Treebank*, pages 299 – 320. De Gruyter Saur, Berlin, Boston.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- Mirko Tavoni. 2012. *DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica*. Università degli Studi di Napoli” L’Orientale”, Il Torcoliere-Officine.
- Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: an Overview. In *Treebanks*, pages 5–22. Springer.
- Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. In

DADH2019: 10th International Conference of Digital Archives and Digital Humanities, pages 20–28. Digital Archives and Digital Humanities.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proceedings of the Universal Dependencies Workshop 2018*, pages 192–201, Brussels.

“Spotto la quarantena”: per una analisi dell’italiano scritto degli studenti universitari via social network in tempo di COVID-19

Francesca Chiusaroli¹, Johanna Monti², Maria Laura Pierucci¹, Gennaro Nolano²

¹Università di Macerata

²UniOr NLP Research Group, Università degli Studi di Napoli L’Orientale

f.chiusaroli@unimc.it

marialaurapierucci@gmail.com

{jmonti, gnolano}@unior.it

Abstract

Italiano Per un’analisi dell’italiano scritto degli studenti universitari, un contesto privilegiato si trova in quei canali social di condivisione tra pari noti con la titolatura di “Spotted”. Attivi particolarmente su Instagram, si tratta di canali di interazione pubblica informale che forniscono elementi significativi all’indagine sulle scritture spontanee dei giovani in rete. Questa ricerca presenta un corpus di oltre trentamila testi da Spotted universitari italiani. Concentrandoci sulle fasi COVID (e sul periodo immediatamente precedente), intendiamo illustrare come il corpus *Spotted-Ita* appaia funzionale a un’indagine riferita alla comunicazione linguistica nell’occasione di un’esperienza unica e straordinaria vissuta dagli studenti universitari.

English “Spotted” posts represent one of the most popular forms of Computer-mediated Communication (CMC) among university students in Italy, and as such, they represent a privileged context to analyze the Italian language used by students on the Web. This kind of informal communication channels is active especially on Instagram, and provides relevant insights on the spontaneous writing of young people. Based on *Spotted-ITA Corpus*, a corpus of over 30,000 posts retrieved from Italian Spotted accounts on Instagram, this paper presents a focus on the period of lockdown in Italy due to COVID pandemics and its immediately preceding pe-

riod, showing how this corpus can be useful in analyzing the way Italian university students communicated during this unique and extraordinary experience.

1 Introduzione

Tra le manifestazioni della lingua giovanile, un ruolo peculiare è assolto dalle scritture degli studenti universitari, i quali, pur nelle differenze tra i percorsi formativi (convenzionalmente distinti in umanistici e scientifici), attestano competenze di scrittura riguardanti lo strato più colto delle giovani generazioni, il più edotto sugli usi standard della lingua, e parimenti esposto alle condizioni comunicative informali del web. Su queste premesse, numerose sono le analisi e le ricerche accademiche rivolte a definire gli equilibri, le differenze, le interferenze, che si innescano tra le varie espressioni della lingua scritta all’interno del diasistema delle scritture, dai luoghi della più rigida formalità (come le tesi di laurea) a quelli della massima spontaneità (comunicazioni attraverso social network e piattaforme di messaggistica istantanea). La naturale permeabilità e l’osmotica relazione tra i livelli formale e informale, o pubblico e privato, fanno rilevare, nella scrittura dello studente, la tipica apertura a forme neologiche e non standard (per grafia, grammatica e lessico), indotte dalla cosiddetta lingua di Internet (“scritture brevi” digitali), ed evidenziano l’annessa connotazione emozionale che di tali adozioni costituisce la diretta implicazione espressiva².

²Oltre all’ampia bibliografia sul tema della lingua e dei linguaggi giovanili, di cui solo pochi titoli saranno citati nella bibliografia del presente contributo, interessa qui citare, come punto di riferimento per il tema generale, il Progetto di Ricerca Nazionale 2017, “UniverS-Ita. L’italiano scritto degli studenti universitari: quadro sociolinguistico, tendenze tipologiche, implicazioni didattiche”, attualmente in corso: (<https://site.unibo.it/univers-ita/it>). Per la nozione di “scritture brevi” digitali, cfr. Scritture Brevi, progetto di ricerca e blog, su www.scritturebrevi.it.

Per una analisi della scrittura degli studenti universitari di oggi, un contesto privilegiato si trova in quei canali social di condivisione noti con la titolarità di “Spotted”. Tali luoghi virtuali costituiscono ambiente di interazione elettivo degli studenti, in particolare per ciò che riguarda la piattaforma *Instagram*, che - rispetto ad esempio al caso di *Facebook* - garantisce all’analisi la qualità di bacheca attualmente più popolare e diffusamente utilizzata tra i giovani, idonea allo scopo di riversare i propri pensieri e anche sottoporre dubbi e quesiti indirizzandoli al gruppo di pari³.

Di generale diffusione su scala nazionale, *Spotted* costituisce un format grafico-testuale che si caratterizza innanzitutto per l’anonimizzazione dei mittenti assicurata dagli amministratori che fungono da editor al momento della pubblicazione. In questo modo, dunque, l’intervento dell’autore di ogni post appare svincolato da eventuali condizionamenti esercitati da un presidio esterno – o del mondo “adulto” (in particolare, nel caso specifico, educatori e docenti) – così che, in *Spotted*, si offre all’analisi una scrittura informale e allo stesso tempo pubblica, gergale ma intesa ad ampia diffusione, interessante per la forma oltre che per i contenuti condivisi.

Il presente contributo è così organizzato: la sezione 2 riguarda il corpus *Spotted-Ita*, sulla cui composizione si forniscono informazioni dettagliate insieme ad una descrizione delle modalità di raccolta dei post; la sezione 3 invece riguarda l’analisi linguistica condotta su due subcorpora relativi ai post pubblicati dagli studenti nei due mesi prima e nei due mesi durante il lockdown; seguono infine le conclusioni.

2 Il corpus *Spotted-Ita*

Il corpus denominato *Spotted-Ita* è costituito da 33.865 testi (“post”) pubblicati su Instagram nel

Il contributo di F. Chiusaroli e M.L. Pierucci in questo paper consiste nell’impianto teorico e di analisi e si inserisce nell’ambito del PRIN “UniverS-Ita”. J. Monti e G. Nolano hanno curato la raccolta del corpus, le collocazioni, la realizzazione dei grafici e tabelle e il più generale framework computazionale. I risultati del contributo presentato sono condivisi da tutti gli autori.

³Si vedano i report Blogmeter e Nielsen relativi all’anno in corso, disponibili online rispettivamente agli indirizzi <https://www.blogmeter.it/it/reports/italiani/-e-social-media-2020> e <https://www.nielsen.com/us/en/insights/report/2020/the-nielsen-total-audience-report-august-2020/>.

periodo che va dal 10/01/2019 al 04/05/2020⁴, per un totale di 1.063.905 token e 56.973 type.

La selezione degli account da cui raccogliere i post è stata in gran parte determinata dall’elenco degli atenei che costituiscono i soggetti della ricerca PRIN UniverS-Ita. Nello specifico, gli account *Spotted* considerati ai fini della creazione del corpus sono riportati nella Tabella 1 assieme al numero di post da essi pubblicati, mentre la Figura 1 mostra la distribuzione geografica dei post presenti nel corpus.

Account	Num.post
spotted.unibo	7.791
spotted.unicz	7.259
spotted.unimc	5.946
spotted.unina	3.024
spotted.unitorvergata	2.903
spotted.unipa	2.389
spotted.uniud	1.637
spottedbicocca	902
spotted.unipas	670
spotted.unito	418
spotted.uniss	303
spotted.lasapienza	240
spotted.unisa	212
unimi.spotted	58
spotted.univpm	50
spottedunicampania	32
spottedpoliba	31
Totale	33.865

Tabella 1: Numero post per account *Spotted*.

La raccolta di dati linguistici da *Instagram* è stata particolarmente elaborata date le difficoltà intrinseche dovute alla natura dei post, che sono pubblicati sotto forma di immagini e che, per poter essere analizzati, devono essere preliminarmente convertiti in formato testuale.

Il processo di raccolta dei post è stato quindi il seguente: per scaricare le immagini e i metadati dei post interessati è stata utilizzata la libreria *Instaloader*⁵; successivamente i post sotto forma di immagini sono stati convertiti in testo attraverso il tool OCR *pytesseract*⁶. Per quanto i testi estratti con questo strumento presentino un certo margine

⁴Un ulteriore gruppo di circa 600 post, tutti pubblicati in periodo di lockdown in Italia e non inclusi nell’analisi di quest’articolo, costituisce oggetto di nuove analisi in corso sul tema. Si tratta di testi estratti da una serie di account universitari ancora di *Instagram* denominati *Poivorre* (una sorta di estensione degli *Spotted*), caratterizzati per il richiamo alla dimensione nostalgica dell’esperienza universitaria (@poivorreitorvergata usare la lavagna delle aule per studiare).

⁵<https://pypi.org/project/instaloader/>.

⁶<https://pypi.org/project/pytesseract/>.

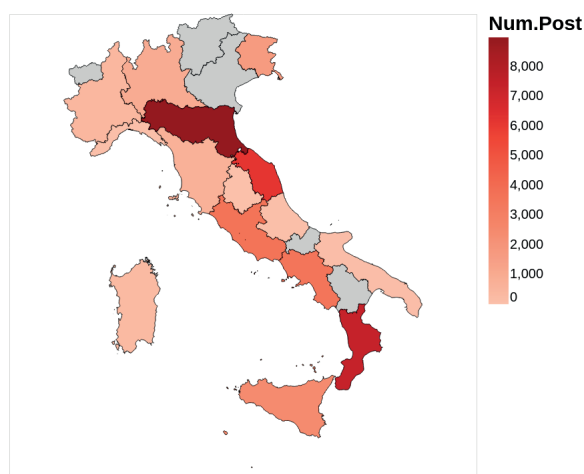


Figura 1: Mappa di calore sulla provenienza dei post *Spotted*.

di errore, sia riguardo elementi puramente testuali come punteggiatura, spazi e lettere accentate che riguardo elementi non puramente testuali (come possono essere gli emoji), tuttavia hanno mostrato un grado di leggibilità accettabile e ancorché grezzi, risultano idonei all'analisi.

È utile premettere che ciascun account *Spotted* istituisce e rappresenta comunità di scriventi idealmente delimitate dall'appartenenza ai relativi atenei del territorio nazionale.

Nonostante il carattere anonimo dei testi raccolti, la loro assimilabilità ai fini della presente analisi si istituisce dunque a partire dal criterio dell'appartenenza logicamente ipotizzabile per gli autori, ovvero l'età anagrafica, il ruolo di studenti universitari, l'appartenenza all'ateneo di riferimento dell'account, e le attività per le quali si pubblica richiesta o denuncia.

Infine un'ultima precisazione riguarda la natura dei contenuti dei post considerati per il corpus, che non sono stati filtrati per argomento. Il corpus *Spotted-Ita*, costituito in base a tali premesse, si presta ad essere interpretato a partire da molteplici prospettive di osservazione e di analisi, che molto possono dire sulla qualità dell'italiano degli studenti di livello universitario in una scrittura eterodiretta e pubblica, ma insieme libera e spontanea.

3 La lingua della pandemia: metodi per l'analisi del corpus

Una prima prospettiva di analisi del corpus è rappresentata dallo studio di come gli studenti univer-

sitari italiani abbiano vissuto e narrato il lockdown in Italia dovuto alla pandemia da COVID-19 e, di conseguenza, come questa esperienza senza precedenti si rifletta nella lingua usata nei loro post sui canali presi in esame.

A tal fine si è deciso di estrapolare due subcorpora da *Spotted-Ita*: il primo, indicato come 'pre-lockdown', copre il periodo che va dal 09/01/2020 al 07/03/2020; mentre il secondo, indicato come 'durante il lockdown', include il periodo che va dall'08/03/2020 al 04/05/2020. In tal modo, entrambi i subcorpora coprono un periodo di 58 giorni, e sono composti, rispettivamente, da 6.536 e 3.992 testi. Informazioni dettagliate sui due subcorpora (numero di token, numero di type e type/token ratio (TTR)) sono riportate nella Tabella 2.

Sezione	Num.Token	Num.Type	TTR
Pre-Lockdown	168.600	34.357	20,38%
D.Lockdown	113.185	25.449	22,48%

Tabella 2: Descrizione dei subcorpora Pre- e Durante-Lockdown.

I due subcorpora, così definiti, costituiscono l'oggetto del presente contributo, che intende soprattutto illustrare la densità del corpus, così come le potenzialità del materiale raccolto, la sua originalità sul piano dei contenuti, consentendo allo stesso tempo di verificare le consuetudini scritte favorite dal mezzo.

I due subcorpora sono stati sottoposti ad una operazione di pre-processing dei dati che ha riguardato la eliminazione di punteggiatura e stopword⁷ tramite la libreria di Python NLTK⁸, insieme a una lemmatizzazione e un tagging delle Part of Speech (*PoS*) con lo strumento *treetaggerwrapper*⁹.

Utilizzando una lista di parole chiave considerate esemplificative per il periodo preso in esame, si è proceduto all'analisi delle collocazioni¹⁰ così da analizzarne il cambiamento nei due periodi in esame. Le parole chiave possono essere suddivise nelle seguenti categorie:

⁷La lista utilizzata è quella proposta da NLTK, cui è stato aggiunto il termine 'spotto' vista l'altissima frequenza nel corpus.

⁸<https://www.nltk.org>.

⁹<https://treetaggerwrapper.readthedocs.io/en/latest/>.

¹⁰L'estrazione delle collocazioni è stata effettuata attraverso la libreria per Python NLTK, con frequenza minima = 1, e il tool online SketchEngine.

- parole della pandemia: *virus*, *Corona virus* e *corona virus* (anche unverbati), *Covid*, *lockdown*, *quarantena*;
- parole della didattica a distanza (DAD): *videolezione/i*, *lezione/i online*, *segreteria*, *Teams*, *Zoom*;
- verbi del sentimento nostalgico: *manca*, *mancano*, *mancate*, *mancare*, *tornare*.

Nella Tabella 3, vengono riportati alcuni esempi di come cambiano le collocazioni nei due periodi di riferimento presi in considerazione insieme ai relativi valori di Pointwise Mutual Information (PMI).

Termine	Pre-lockdown	D. il lockdown
covid	c. furia (11,42)	c. commovente (10,71)
	c. provvedimento (11,42)	c. god (10,71)
	epidemiologico c. (11,42)	defunto c. (10,71)
	mentana c. (11,42)	epidemiologico c. (10,71)
	c. insegnamento (10,20)	morgan c. (10,71)
quarantena	inverno q. (11,89)	angosciare q. (7,78)
	q. forzato (11,31)	combattere q. (7,78)
	q. ahahahah (11,89)	estenuante q. (7,78)
	q. volontario (10,08)	infrangere q. (7,78)
	beccare q. (9,72)	passatempo q. (7,78)
tornare	bestiame t. (10,94)	riuscire t. (9,33)
	chilometro t. (10,95)	t. delfino (9,33)
	t. canosa (10,95)	t. incazzarsi (9,33)
	t. patria (10,94)	t. patria (9,33)
	rancore t. (9,94)	t. rattusare (9,33)
mancare	m. cfu (11,79)	amiketti m. (8,23)
	finalizzare m. (10,79)	cocainomane m. (8,92)
	giustamente m. (10,79)	lisbona m. (8,92)
	m. spottedunina (10,21)	m. fiato (8,92)
	umilta m. (9,79)	m. persino (8,92)
videolezione	v. analisi (11,57)	smadonna v. (12,52)
	arrivare v. (10,43)	v. fissato (12,52)
		v. iniziare (10,93)
		v. accorgere (10,20)
		spottedlasapienza v. (9,93)

Tabella 3: Esempi di collocazioni nel corpus con indicazione del valore di PMI: la tabella è stata creata con i subcorpora normalizzati diacronicamente e lemmatizzati.

Si vedano, ne citiamo due, i seguenti post: “*Spotto il prof V. che in diretta nella pausa della lezione online ammette di aver sviluppato una dipendenza da whisky in questa quarantena*” e “*Spotto il Coronavirus e la quarantena che mi hanno fatto comprendere quanto abbia fatto bene Caino ad ammazzare quello stronzo di suo fratello!*”. È qui ben esemplificata la formula “Spotto+qualcuno/qualcosa+che” a introdurre elementi che contraddistinguono le conversazioni digitali sulla vita universitaria a distanza (*il/la prof.*, *lezione online*, *quarantena*, *Coronavirus*): si tratta di flussi grazie ai quali è possibile indagare la dimensione del *sentiment* tipicamente emergente dalla scrittura dei giovani via social, in una fase

in cui la componente emotiva risulta ancor più accentuata a causa del peculiare evento pandemico. L’incrocio dei dati fra la prospettiva temporale e quella geografica consente di osservare come, al diffondersi della pandemia e delle conseguenti restrizioni sul territorio nazionale, sia seguita una progressiva convergenza verso il tema del lockdown nell’uso della lingua in seno alle comunità afferenti alle varie sedi universitarie.

A tale riguardo, si vedano i dati riportati nelle Figure 2, 3 e 4 che mostrano le frequenze sul tempo delle parole chiave *DAD*, *nostalgia* e *pandemia* nelle tre università con frequenza maggiore di pubblicazione (rispettivamente UniBo, UniMc e UniNa), e quelli riportati nella Tabella 4, che fanno riferimento invece, alla percentuale di post *Spotted* nei quali compaiono le parole chiave *coronavirus*, *covid*, *lockdown* e *quarantena* all’interno dei due subcorpora considerati complessivamente.

Parallelamente a questo dato, che emerge nel confronto sulle tecniche pratiche (richiesta di informazioni e modalità per seguire le lezioni in tele-didattica, su tutte), si evidenzia in modo altrettanto netto l’emergenza di un sentimento “nostalgico” rilevabile dalla co-occorrenza quantitativamente marcata di lemmi riconducibili alla situazione sanitaria, come *quarantena*, con lemmi della sfera semantica delle emozioni come, ad esempio, *tornare*, *manca*, *mancano*, *mancate*, *mancare* a testimoniare il senso di solitudine e perdita (di tempo, di relazioni, di produttività negli studi).

3.1 Spotto la quarantena

Il corpus, annotato con le parti del discorso¹¹, ha inoltre permesso di effettuare ulteriori analisi relativamente all’uso ed ai contesti in cui occorre la parola *quarantena*.

Valutando il lemma nei testi estratti (cfr. Tabella 5) si nota come le co-occorrenze dalla frequenza più alta siano rappresentate dai sintagmi PRE+quarantena e PRO+quarantena, come in “in quarantena, in questa quarantena”, “durante la/questa quarantena”.

Con frequenza più bassa compaiono invece i sintagmi come “dopo questa quarantena” e “dopo la quarantena” (entrambi 5 occorrenze), e la forma semicolta “il post quarantena” (4 occorrenze), con valore temporale. Alcuni esempi dai testi:

¹¹Il tagset utilizzato è disponibile all’indirizzo <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt>.

	gennaio	febbraio	marzo	aprile	maggio
parole della pandemia	0.22%	3.69%	10.33%	9.18%	7.44%
parole della DAD	1.95%	2.99%	18.98%	9.44%	8.68%
verbi della nostalgia	0.8%	0.74%	3.34%	5.61%	6.2%

Tabella 4: Percentuale dei post nel corpus nei quali sono presenti le parole chiave.

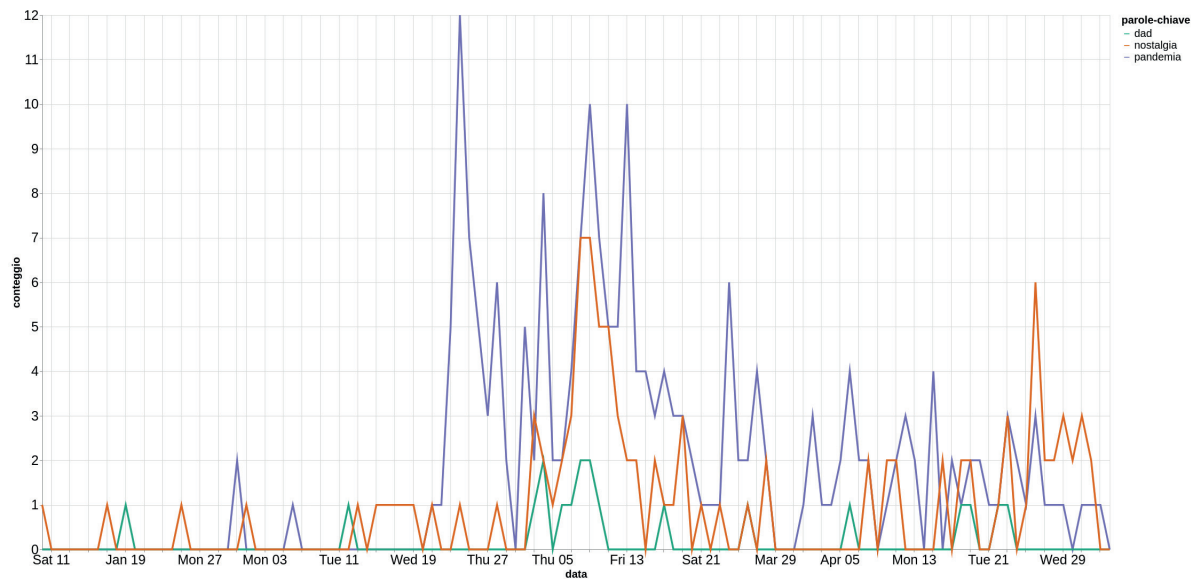


Figura 2: Frequenza delle parole chiave *DAD*, *nostalgia* e *pandemia* per i post di spotted_unibo.

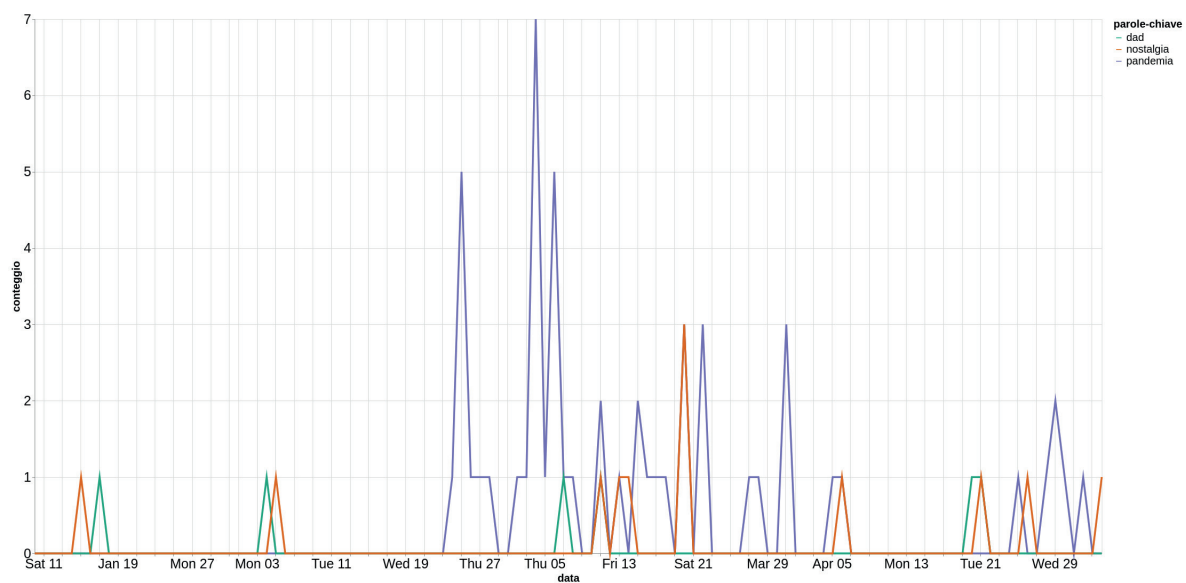


Figura 3: Frequenza delle parole chiave *DAD*, *nostalgia* e *pandemia* per i post di spotted.unimc.

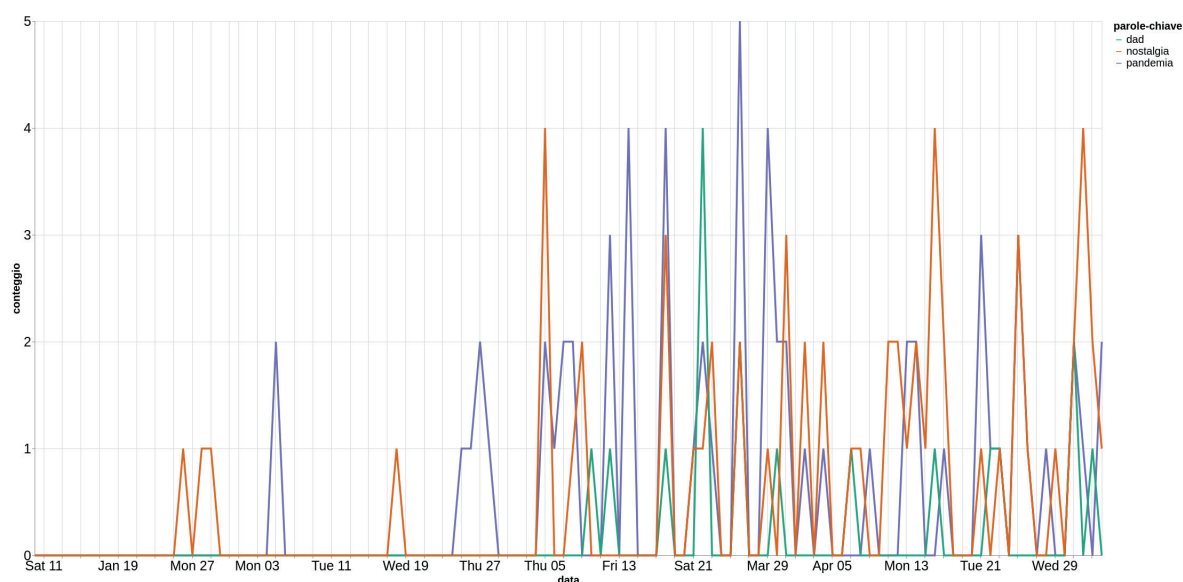


Figura 4: Frequenza delle parole chiave *DAD*, *nostalgia* e *pandemia* per i post di spottedunina.

Sintagma	Freq.
in_PRE questo_PRO:demo quarantena_NOM	34
durante_PRE il_DET:def quarantena_NOM	13
periodo_NOM di_PRE quarantena_NOM	13
giorno_NOM di_PRE quarantena_NOM	12
durante_PRE questo_PRO:demo quarantena_NOM	11

Tabella 5: PoS Tag delle co-occorrenze del termine *quarantena*.

- “*spo, data la situazione di disagio complessivo che spero si capisca, cerco tipa per il post **quarantena**; garantisco ottima compagnia fornitura illimitata di risorse 420, umorismo, prestazioni sessuali e cattive intenzioni, senza contare la mia enorme capacita di guardare al futuro, dato che la tipa la cerco gia da adesso... anonimo pis @ scrivi un messaggio... i) e)*”
- “*qualcuno disposto a due chiacchiere **in questa quarantena** forzata per tutti noi che siamo rientrati? (lo son rientrata il 24 febbraio, non faccio parte della mandria di gente tornata in queste ultime ore... incompetent!!!) comunque vi avverto che non sono una top model! dato che molti si fermano all’aspetto fisico anche per due chiacchiere*”
- “*organizziamo un concerto di gigione per il primo giorno **dopo la quarantena**? spotted unisa*”.

I sintagmi ADJ+NOM veicolano più di altri la

connotazione di *quarantena* in termini “emotivi”: “opprimente e sterile q.”, “noioso e insopportabile periodo di q.”, “noiosa q.”.

Altri esempi:

- “*spotto ragazza carina e simpatica che abbia voglia di chiacchierare per magari fare nascere qualcosa durante questa **opprimente e sterile quarantena** jç (sono un ragazzo a detta di molti simpatico e non di brutto aspetto haha) contatto i like —*”
- “*spotto una ragazza single con cui uscire una volta terminato questo **noioso e insopportabile periodo di quarantena!!!** contatto i likes*”
- “*spotto ragazze che mi tengano compagnia virtualmente in questa **noiosa quarantena**”*
- “*spotto qualcuno con cul fare amicizia in questa **noiosa quarantena**”.*

Ancora, il termine *quarantena* funziona da catalizzatore delle conversazioni degli studenti introducendo nella scrittura stati d’animo come il senso di solitudine, la nostalgia data dalla lontananza dalle sedi universitarie e dai compagni di corso e di sede, e il desiderio del ritorno alla vita universitaria in sede. Alcuni esempi:

- “*Spotto chi come me sta passando la **quarantena lontana** dalla persona che ama e che sta male per il fatto che probabilmente dopo il*

4 maggio non potremo ancora spostarci tra regioni. **Non siete soli**"

- "Sotto il casino che facevamo per strada, sia a Bologna che da altre parti (NICOLOOOO), e spero che possiamo tornare a farlo ancora: Non so perche me he sono uscito. con ste cose mielose ma non mi'sto sopportando piu in questa cazzo di quarantena e **voglio tornare** al piu presto alla normalita con voi ragazzi".

Il lemma *quarantena* (328 occorrenze), termine selezionato da Treccani fra le "parole del Coronavirus"¹² nell'ambito della rubrica online "Le parole valgono", appare la verbale espressione della centralità dell'argomento COVID nel corpus di *Spotted*.

Come si vede in Figura 5, "quarantena" è entrato più tardi nell'uso rispetto allo stesso termine "coronavirus" che mostra un primo picco già alla fine di febbraio 2020. In compenso, "quarantena" risulta essere anche il termine più stabilizzato e longevo tanto da apparire con regolare frequenza fino a maggio. "Coronavirus", invece, scompare quasi del tutto dall'uso dopo i primi giorni di marzo. Interessante anche vedere come il termine più specialistico *COVID* veda un uso più raro sin dal suo primo utilizzo a febbraio, pur restando presente nella comunicazione degli *Spotted* fino alla fine del periodo preso in esame.

Nell'ottica di una valutazione della scrittura giovanile, alla quale sovente si tende a imputare la responsabilità del declino della lingua rispetto agli influssi linguistici esogeni, ci sembra qui infine interessante rilevare la pressoché nulla frequenza d'uso del termine *lockdown*, che viene impiegato in soli 2 post fra quelli raccolti nel corpus *Spotted-ita*, laddove esso risulta invece ampiamente e stabilmente adottato nello stesso periodo in ambito giornalistico e della comunicazione istituzionale.

Spostandoci sulle parole chiave della DAD, è interessante notare come il sintagma 'lezione online' veda un incremento molto basso nell'uso, come si vede dalle tabelle 6 e 7, che mostrano le frequenze dei trigrammi in cui compare il termine 'lezione'.

¹²https://www.treccani.it/magazine/parolevalgono/Le_parole_del_Coronavirus/index.html.

Sintagma	Freq.
durante_PRE il_DET:def lezione_NOM	81.0
il_DET:def lezione_NOM online_ADJ	65.0
il_DET:def lezione_NOM di_PRE	53.0
del_PRE:det lezione_NOM online_ADJ	27.0
lezione_NOM online_ADJ di_PRE	21.0
al_PRE:det lezione_NOM online_ADJ	17.0
del_PRE:det lezione_NOM di_PRE	16.0
lezione_NOM di_PRE diritto_NOM	12.0
lezione_NOM di_PRE storia_NOM	12.0
lezione_NOM online_ADJ e_CON	12.0

Tabella 6: PoS Tag delle co-occorrenze del termine "lezione" nel periodo pre-lockdown.

Sintagma	Freq.
durante_PRE il_DET:def lezione_NOM	67.0
il_DET:def lezione_NOM online_ADJ	47.0
il_DET:def lezione_NOM di_PRE	46.0
al_PRE:det lezione_NOM online_ADJ	12.0
del_PRE:det lezione_NOM online_ADJ	12.0
lezione_NOM online_ADJ di_PRE	12.0
seguire_VER:infi il_DET:def lezione_NOM	11.0
del_PRE:det lezione_NOM di_PRE	9.0
lezione_NOM online_ADJ sotto_VER:pres	9.0
sotto_VER:pres il_DET:def lezione_NOM	9.0

Tabella 7: PoS Tag delle co-occorrenze del termine "lezione" nel periodo d. il lockdown.

Nonostante si abbassino le frequenze delle forme *il_DET:def lezione_NOM online_ADJ*, *al_PRE:det lezione_NOM online_ADJ* e *lezione_NOM online_ADJ di_PRE*, compaiono trigrammi nuovi come *lezione_NOM online_ADJ sotto_VER:pres* e *del_PRE:det lezione_NOM online_ADJ*.

Alcuni esempi dai testi pre-lockdown:

- *Sotto lezione di fisica (prof M.) Ingegneria civile. Durante la lezione online e suonato l'allarme antincendio e la prof e scappata*
- *Sotto tutti quelli che si connettono 30 Min prima dell'inizio della lezione online. Avete paura di non trovare posto?.*

E alcuni esempi dai testi durante il lockdown:

- *Sotto la lezione online di economia politica da sotto le coperte con pigiamae caffè sul comodino alle 12*
- *Sotto l'eroe che durante la lezione online di diritto internazionale, accorgendosi di aver lasciato un libro nella sede di Ravenna (nella quale non puo fare ritorno) ha esclamato VACCA BOIA.*

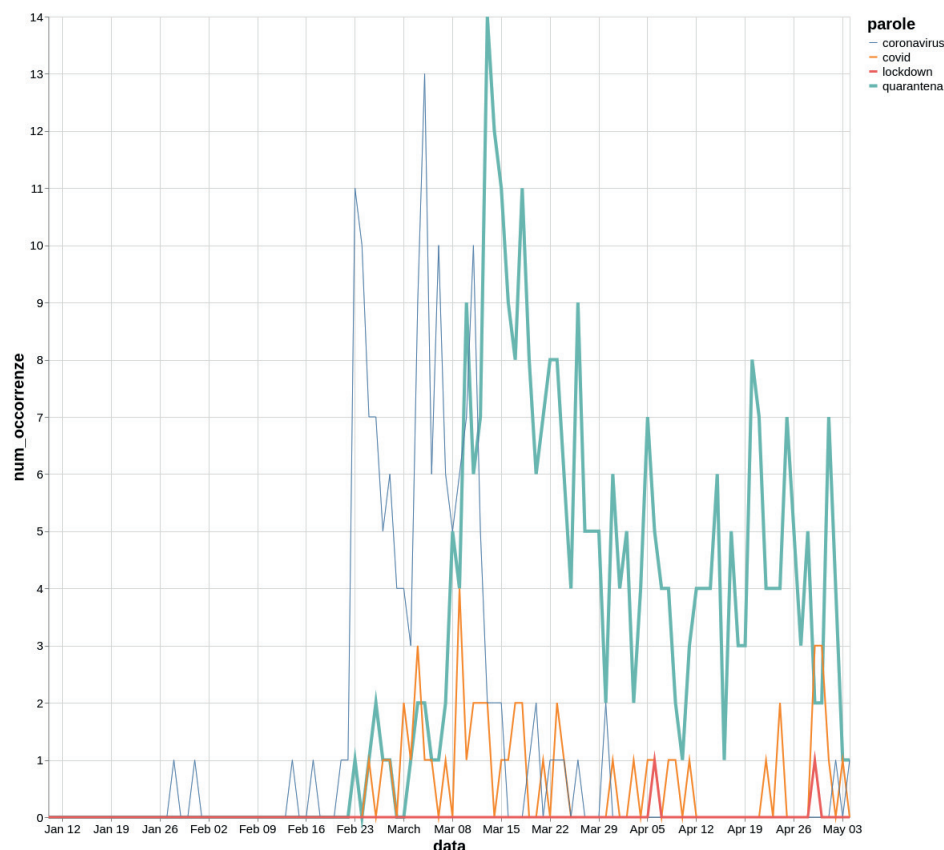


Figura 5: Conteggio delle keyword sul tempo

Infine, nell'analisi delle parole chiave della nostalgia si è notata la presenza di un fenomeno che riguarda il verbo 'mancare', che passa da 88 usi pre-lockdown a 364 durante il lockdown. Nel periodo pre-lockdown, a mancare possono essere i CFU o un/a ragazzo/a, come negli esempi seguenti:

- *Ciao spo, devo laurearmi e mi mancano Qc-fu, qualcuno mi pud dire se c'è un esame della triennale di qualsiasi facoltà molto facile? Grazie mille JI, © Scrivi un messaggio... eA*
- *Spotto Lumpa, non sei solo la migliore amica, sei la mia vita... Non posso pensare che stiamo buttando tutto via perche non abbiamo il coraggio di parlarne...mi manchi.*

Mentre, come ci si può aspettare, durante il lockdown a mancare è invece la quotidianità temporaneamente perduta:

- *Spotto la mia tremenda nostalgia di Bologna, della normalità, degli abbracci senza paura, dei baci al mio lui, della gente per la città. **Mi manca** tutto questo terribilmente, ma andrà tutto bene &*

- *Spotto chiara e melissa. Prima eravamo tutti I giorni insieme. Oggi la distanza Mantova Roma si sente. **MI mancate** tope.*
- *Spotto i miei fratelli dell'ed. 20, dal primo all'ultimo **mi mancate**. Ad alcuni quel posto mette ansia, per me era una seconda casa e spero di tornarci presto. Un bacio a tutti da 3p.*

4 Conclusioni

Analizzare l'italiano scritto degli studenti universitari attraverso i testi postati sulla piattaforma social Instagram, in particolare gli *Spotted*, fornisce un focus di ricerca nodale nel panorama degli studi accademici e scientifici sulla natura della lingua dei giovani.

In tale contesto, nel registrare l'andamento delle conversazioni nel pre- durante- e post quarantena, il corpus *Spotted-ita* fornisce un punto di vista privilegiato per l'analisi dell'esperienza dello studente nelle inedite ed esclusive condizioni conosciute nella fase sociale più significativa dell'emergenza sanitaria COVID-19.

Coadiuvata dagli strumenti della linguistica computazionale, l'analisi qui presentata illumina alcuni aspetti degli usi linguistici degli studenti, in contesti informali, tipicamente non-standard, ma altresì, per il gruppo dei pari, contesti pubblici, consentendo di ridefinire le esigenze comunicative rispetto alle nuove condizioni delle attività quotidiane e di studio svolte a distanza. Ne emerge un osservatorio peculiare e unico, dunque, che esemplarmente rispecchia e riproduce, in termini di rappresentazione linguistica, il vissuto di un periodo storico, mai altrimenti sperimentato nella contemporaneità, di disaggregazione fisica della comunità universitaria.

References

- Chiusaroli F. 2017. "Scritture brevi" nel diasistema delle scritture digitali, in C. De Santis/N. Grandi (eds.) *CLUB Working Papers in Linguistics*, Università di Bologna: 5-18.
- Cortelazzo M. 2010. Linguaggio giovanile, in *Enciclopedia dell'italiano*, disponibile all'indirizzo [\(Enciclopedia dell'Italiano\)](http://www.treccani.it/enciclopedia/linguaggio\-giovanile_) Ultimo accesso 10/09/2020.
- D'Achille P. 2016. Architettura dell'italiano di oggi e linee di tendenza, in Lubello (ed.), *Manuale di linguistica italiana*, Berlin/Boston, De Gruyter: 165-189.
- De Mauro T. 2010. La cultura degli italiani, a cura di F.Erbani, Roma/Bari, Laterza.
- Fiorentino G. 2015. Aspetti problematici del discorso accademico: un'analisi dei riassunti delle tesi di laurea, *Cuadernos de Filología Italiana* 22: 263-84.
- Grandi N. (ed.) 2015. La grammatica e l'errore, Bologna, BUP.
- Lubello S. (ed.) 2020. Homo scribens 2.0. Scritture ibride della modernità, Firenze, Franco Cesati.
- Piemontese M.E./Sposetti P. 2014. La scrittura dalla scuola superiore all'università, Roma, Carocci.
- Palermo M. 2017 *Italiano Scritto 2.0. Testi e Ipertesti*, Roma, Carocci
- Pistolesi E. 2014. Scritture digitali, in G. Antonelli/M. Motolese/L. Tomasin (eds.), *Storia dell'italiano scritto, III: Italiano dell'uso*, Roma, Carocci: 349-75.

Italian Counter Narrative Generation to Fight Online Hate Speech

Yi-Ling Chung

University of Trento
Fondazione Bruno Kessler
ychung@fbk.eu

Serra Sinem Tekiroğlu

Fondazione Bruno Kessler
tekiroglu@fbk.eu

Marco Guerini

Fondazione Bruno Kessler
guerini@fbk.eu

Abstract

English. Counter Narratives are textual responses meant to withstand online hatred and prevent its spreading. The use of neural architectures for the generation of Counter Narratives (CNs) is beginning to be investigated by the NLP community. Still, the efforts were solely targeting English. In this paper, we try to fill the gap for Italian, studying how to implement CN generation approaches effectively. We experiment with an existing dataset of CNs and a novel language model, recently released for Italian, under several configurations, including zero and few shot learning. Results show that even for under-resourced languages, data augmentation strategies paired with large unsupervised LMs can held promising results.

Italiano. *Le Contro Narrative sono risposte testuali volte a contrastare l'odio online e a prevenirne la diffusione. La comunità di NLP ha iniziato a studiare l'uso di architetture neurali per la generazione di CN. Tuttavia, gli sforzi sono stati rivolti esclusivamente all'inglese. In questo lavoro, cerchiamo di colmare la lacuna per l'italiano, mostrando come implementare efficacemente approcci di generazione di CN. Sperimentiamo con un dataset esistente di CN e un modello del linguaggio per l'italiano recentemente rilasciato, in diverse configurazioni, tra cui zero e few shot learning. I risultati mostrano che anche per lingue con poche risorse, strategie di data augmentation abbinate a potenti modelli del linguaggio possono offrire risultati promettenti.*

1 Introduction

The rise of online Hate Speech (HS) brings along the need for combating strategies as it can trigger harmful psychological effects on the target groups and more crimes against them. While research studies have been widely focusing on hate speech detection methodologies for social media platforms (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), a recent line of research has taken the problem a step further by addressing the automatic generation of counter responses, aka counter narratives (Qian et al., 2019; Tekiroğlu et al., 2020), in order to assist non-governmental organizations in their real-world online hatred combating efforts. An example of HS along with a possible CN are shown below:

HS: *Gli arabi sono tutti terroristi e vogliono conquistarci con la violenza e le bombe. Bisogna rispondere con il napalm.* [Arabs are all terrorists and they want to conquer us with violence and bombs. We must respond with napalm.]

CN: *Essere di origine araba non significa essere terroristi, evitiamo generalizzazioni che portano solo ad altro odio.* [Being of Arab descent does not mean being a terrorist, let's avoid generalizations that only lead to more hatred.]

Despite the encouraging results of the counter narrative generation task, experiments have been limited to English due to the scarcity of hate speech / counter narrative data in other languages. In this paper, we investigate counter narrative generation for Italian as a case study where zero or only a small amount of task specific in-language data is available. We first explore the portability of generation across languages, considering that recent neural machine translation (NMT) systems have shown outstanding performances. We pro-

pose utilizing off-the-shelf NMT models to synthesize silver data from other languages, and fine-tuning GePpeTto (Mattei et al., 2020), a recently developed GPT-2 based language model for Italian, on the silver data. We then examine the effect of combining silver with gold data on CN generation by experimenting with various gold data sizes. Our findings show that a proper combination of silver and gold data while fine-tuning LMs can drastically reduce the need for expert-annotator effort on target languages.

2 Related Work

In this section we briefly recap relevant works for our counter narrative generation task, including the problem of online hatred recognition, effectiveness of approaches to hatred intervention, methodologies for generating counter-arguments, and text generation for low-resourced languages.

Hate problem. A wealth of work has investigated online hateful content, aiming at creating datasets for hate speech identification (Warner and Hirschberg, 2012; Burnap and Williams, 2015; Silva et al., 2016). For instance, there are datasets collected from Facebook (Kumar et al., 2018), forums (Silva et al., 2016; de Gibert et al., 2018), and Twitter (Silva et al., 2016; Waseem and Hovy, 2016). Hate speech detection tasks are available at IberEval (Fersini et al., 2018) for Spanish and EVALITA (Del Vigna12 et al., 2017; Fersini et al., 2018) for Italian.

Hate countering. Counter narratives can be used as an effective approach to moderate hateful content on social media platforms such as Twitter (Munger, 2017; Wright et al., 2017), Youtube (Ernst et al., 2017; Mathew et al., 2019) and Facebook (Schieb and Preuss, 2016). Previous studies on hate countering cover several aspects of CNs. For example: defining counter narratives (Benesch et al., 2016), studying their effectiveness (Schieb and Preuss, 2016; Silverman et al., 2016; Ernst et al., 2017; Munger, 2017; Wright et al., 2017), linguistically characterizing online counter narrative accounts (Mathew et al., 2018), creating real or simulated CN datasets (Mathew et al., 2019; Chung et al., 2019; Qian et al., 2019; Tekiroğlu et al., 2020), and neural approaches to CN generation (Qian et al., 2019; Tekiroğlu et al., 2020).

Counter-argument Generation. This task share the same abstract goal as CN generation -

i.e. to produce the opposite or alternate stance of a statement. Previous works adopted sequence-to-sequence architectures to generate arguments (Rakshit et al., 2019; Hua et al., 2019; Rach et al., 2018; Le et al., 2018) targeting specific domains in which massive discussion is available, such as politics (Hua et al., 2019; Hua and Wang, 2018; Le et al., 2018), and economy (Le et al., 2018; Wachsmuth et al., 2018).

NLG for under-resourced languages. In spite of several studies addressing NLG, only a few have investigated the generation for languages other than English. For instance, there is the porting of SimpleNLG API (Gatt and Reiter, 2009) to Dutch (de Jong and Theune, 2018) and Italian (Mazzei et al., 2016), or Bilingual generation via combining NMT and Generative Adversarial Networks (Rashid et al., 2019).

3 Italian Counter Narrative Generation

Our main goal is to determine a methodology for Italian counter narrative generation considering the scarcity of gold standard data for training. Accordingly, we hypothesize that the availability of a decent amount of silver data can provide a kick-start for the generative models. Therefore, we resort to data augmentation through translation with the help of the existing datasets of hate speech / counter narrative pairs in other languages. For translation setting, we use DeepL¹, an off-the-shelf and well-performing MT system, to translate data from other languages to Italian. The translated pairs are used for fine-tuning a large Italian pre-trained generative model, i.e. GePpeTto, along with the original Italian gold standard pairs.

4 Dataset

For our study, we use CONAN dataset (Chung et al., 2019), which is a niche-sourced hate-countering dataset that consists of HS/CN pairs focusing on Islamophobia. The dataset provides pairs in English, French, and Italian, collected with the help of operators from three European NGOs specialized in online hate countering. Each pair in CONAN can either be an original or one of the 2 paraphrases of an original pair. In the experiments, we used the following splits:

1. 2142 pairs (original IT pairs and 1 IT paraphrase pair) as a training set made of gold

¹<https://www.deepl.com/translator>

standard data.

2. 5996 pairs as a training set made of silver data obtained by automatically translating FR and EN pairs to IT.
3. 1071 pairs (the rest of the IT paraphrased pairs) are kept for testing purposes.

5 Models

In order to inspect how Italian CN generation can be accomplished under different resource conditions, we test the effect of using (i) silver data, (ii) gold standard data, and (iii) their combination. In particular we experiment with the following configurations on which `GePpeTto` is fine-tuned:

GP-trans. `GePpeTto` is fine-tuned on the silver data obtained by translating EN and FR pairs to IT using DeepL. This configuration represents the worst case scenario, where no HS/CN pair is available in the target language, and corresponds to a zero-shot learning setting.

Gp-ita. We fine-tune `GePpeTto` on all the original IT pairs in CONAN. This represents our practical best-case scenario, despite the fact that more pairs might provide better results.

GP-hybrid. We conjecture that introducing even a small amount of gold standard examples can help LMs adapt to the domain-specific idiosyncrasies. Moreover, we inspect how generation performance varies with the size of gold standard data provided. In this regard, we conduct a second phase of fine-tuning on top of the GP-trans model using 100, 300, 500, 800, and full IT pairs of CONAN. Therefore, we can represent various intermediate conditions of few-shot learning where few to several pairs for the target language are available. Thus, we assess how much the pre-training with the silver data helps to reduce the amount of gold standard data needed to reach a proper generation performance.

5.1 Training Details

For all the experiments, we have used `GePpeTto` as the pretrained Italian language model adopted from HuggingFace’s Transformers library² and fine-tuned our models on a single K80 GPU using a batch size of 2048 tokens. The training pairs are represented as `[HS_start_token] HS [CN_start_token]`

²<https://github.com/huggingface/transformers>

`CN [CN_end_token]`. The hyperparameter tuning details are provided in the following. At test time, we employed nucleus sampling with a p value of 0.9 for the generation of CNs. Conditioned on HSs, the generated sequence of text tagged with `[CN_start_token]` `CN [CN_end_token]` is selected as output.

Training Epochs We have empirically chosen 5 epochs for training for all the configurations, tuned from {2, 3 and 5} on test set. Preliminary experiments show that while lower number of epochs grant higher novelty in the output, they also came at the cost of lower BLEU scores. A further manual evaluation confirmed that the generation with 5 epoch provides more suitable responses.

Learning rate Once defining the epochs, we experimented with different learning rates of [1,2,5]e-5 and chose 5e-5 for the best performing setting - preliminary experiments show that while producing less novel and slightly more repeated text, the learning rate of 5e-5 consistently has better results in terms of BLEU and ROUGE scores.

Fine-tuning steps. In case where multiple datasets (silver and gold standard) were used, we followed a multi-step fine-tuning procedure by first using the silver and then the gold standard dataset. Gururangan et al. (2020) showed that task-adaptive pretraining using curated datasets from a dataset with similar distribution with the end task, provides significant improvements. Our fine-tuning schema follows this finding by first fine-tuning `GePpeTto` with the silver data as the task adaptive pretraining with an augmented dataset. Our preliminary experiments confirmed that adapting fine-tuned models towards the language characteristics of the target corpus is more effective than mixing silver and gold data together in a single fine-tuning procedure.

5.2 Evaluation

For our experiments we report word-overlap metrics BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to evaluate the CN generation on the gold standard test set. As for the generation quality, we compute Repetition Rate (Bertoldi et al., 2013) and Novelty (Wang and Wan, 2018) to assess how *Diverse* a response is with reference to the given HS and how *Novel* the generation is concerning the training data.

We also conduct a human evaluation to compare the generation quality of the configurations based on 3 criteria: **(i) Suitableness.** How suitable the given CN is as a response for the input HS. **(ii) Specificity.** How specific the given CN is as a response. This metric is used to discern suitable responses that are nonetheless very generic. **(iii) Grammaticality.** How grammatically correct the given CN is. All scores were in a scale from 1 to 5.

6 Results and Discussion

Model comparison. Results in Table 1 show that using the silver data (GP-trans) provides a viable step towards a proper model. When gold standard data is also available (GP-hybrid), we obtain better quantitative performance in terms of BLEU and ROUGE scores in comparison to the best case scenario (GP-ita). Furthermore, mixing the silver translation and the Italian gold standard data (GP-hybrid) yields better performances also in terms of output diversity (RR 11.7 vs 12.8). On the contrary, the most novel output is obtained by GP-trans, which can be expected since EN and FR pairs usually have slightly different focus on the topic of Islamophobia (topics and tropes can vary across nations and cultures). In Table 2 we provide few examples of generated CNs.

Learning Curve Discussion. As can be seen in Figure 1, even 100 Italian pairs are enough to dramatically improve the performances of GePpeTto on the task of CN generation over the baseline GP-trans. If we continue fine-tuning GP-trans with more and more Italian pairs, soon we are able to outperform also GP-ita. The number of examples required to obtain a new state of the art CN generation in Italian comes within 200 and 300, which reduces the required amount of gold standard data by around 80%. Therefore, it becomes clear that a good NMT model can be of fundamental help while porting the generation task to new languages, especially if few or no gold standard examples are available in the target language. Considering the fact that the counter narrative data collection is an expert-based task requiring costly human effort (Chung et al., 2019), decreasing the required amount of expert data can be of remarkable importance for low-resource languages.

Human Evaluation. As annotators, we employed 2 Italian native speakers that are expert

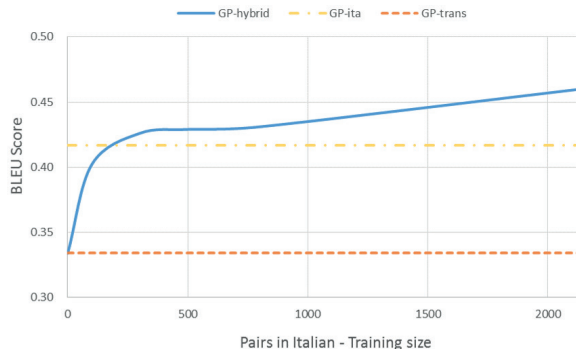


Figure 1: Learning Curve of GP-hybrid model while Italian pairs being added. GP-hybrid performance with no examples is shown as GP-trans.

in counter narrative production. The annotators were instructed in assessing CN *suitableness*, *specificity*, and *grammaticality* with respect to the paired hate speech. During training, we explained what a common and suitable counter narrative is, and then asked them to intuitively evaluate the generation without overthinking. We further presented 20 examples of HS/CN pairs to demonstrate the appropriate evaluation. In order to avoid comparison or primacy/recency effects, we have presented 20 random pairs from each condition to each annotator as a single randomized file and asked them to evaluate each counter narrative with respect to the 3 criteria. The results presented in Table 3 show that all models reach very high levels of grammaticality; most of the sentences were completely grammatical and few ungrammatical ones were due to dangling sentences. Moreover, using silver data alone can already provide a performance lower than but close to the GP-ita case for Suitableness and Specificity. Finally, fine-tuning GP-trans further using gold standard data (GP-hybrid) provides the most suitable and the least generic responses among the 3 models in line with their performance ranking of automatic metrics.

7 Conclusion and Future Work

Counter narrative generation using neural architectures is beginning to be studied for hatred intervention. In this paper, we presented the first attempt of CN generation for Italian, investigating several variations of generation when gold data is limited or not available. Our experiments reveal that with simple data augmentation strategies paired with powerful LMs can bring promising

Model	BLEU _s	BLEU _c	ROUGE1	ROUGE2	ROUGEL	RR	Novelty
GP-trans	0.327	0.048	0.236	0.061	0.196	12.795	0.785
GP-ita	0.417	0.231	0.343	0.187	0.305	12.870	0.561
GP-hybrid	0.460	0.287	0.380	0.234	0.344	11.752	0.522

Table 1: Quantitative results of fine-tuned models. BLEU scores are reported at sentence-level (BLEU_s) and corpus-level (BLEU_c).

GP-trans: *La maggior parte dei pedofili nel Regno Unito sono uomini bianchi, quindi secondo la vostra logica il mondo sarebbe un posto migliore senza uomini bianchi?* [Most pedophiles in the UK are white men, so by your logic would the world be a better place without white men?]

GP-ita: *Non esiste nessuna invasione islamica dell'occidente. La popolazione musulmana in Italia non supera il 5%.* [There is no Islamic invasion of the West. The Muslim population in Italy does not exceed 5%.]

GP-hybrid: *Non si può ritenere una invasione quando il numero di musulmani in Italia è inferiore addirittura al 5%.* [It cannot be considered an invasion when the number of Muslims in Italy is less than even 5%.]

Gold: *Salve, in realtà il numero di musulmani in Italia è pari a poco meno il 5% della popolazione totale: ritiene davvero che si possa parlare di sostituzione?* [Hello, in reality the number of Muslims in Italy is equal to a little less than 5% of the total population: do you really think that we can talk about substitution?]

Table 2: Sample CN generations along with EN translation. GP-trans generation is grammatically correct but focused on the UK/FR scenario. Instead, GP-ita and GP-hybrid can mimic gold arguments with novel and diverse wording.

results for under-resourced languages. As future work we plan to explore methods for improving specificity via the use of diverse types of knowledge (e.g., facts, events and named entities) from external resources. This would also mitigate the problem of the model possibly generating hallucinated content (i.e. non-existent or wrong content), see for example (Maynez et al., 2020; Nie et al., 2019). Finally, we plan to apply this approach to other hate phenomena such as antisemitism, homophobia, and misogyny.

Model	Suitable	Specific	Grammar
GP-trans	2.47	2.20	4.52
GP-ita	2.78	2.32	4.72
GP-hybrid	2.82	2.57	4.40

Table 3: Human evaluation results.

References

- [Benesch et al.2016] Susan Benesch, D Ruths, KP Dillon, HM Saleem, and L Wright. 2016. Considerations for successful counterspeech.
- [Bertoldi et al.2013] Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.
- [Burnap and Williams2015] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.
- [Chung et al.2019] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.
- [de Gibert et al.2018] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.
- [de Jong and Theune2018] Ruud de Jong and Mariët Theune. 2018. Going dutch: Creating simplenlg-nl. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 73–78.
- [Del Vigna12 et al.2017] Fabio Del Vigna12, Andrea Cimino23, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.
- [Ernst et al.2017] Julian Ernst, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer,

- Gary Bente, and Hans-Joachim Roth. 2017. Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.
- [Fersini et al.2018] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.
- [Fortuna and Nunes2018] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.
- [Gatt and Reiter2009] Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93.
- [Gururangan et al.2020] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- [Hua and Wang2018] Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. *arXiv preprint arXiv:1805.10254*.
- [Hua et al.2019] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. *arXiv preprint arXiv:1906.03717*.
- [Kumar et al.2018] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.
- [Le et al.2018] Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.
- [Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- [Mathew et al.2018] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.
- [Mathew et al.2019] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.
- [Mattei et al.2020] Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model.
- [Maynez et al.2020] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- [Mazzei et al.2016] Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th international natural language generation conference*, pages 184–192.
- [Munger2017] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.
- [Nie et al.2019] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy, July. Association for Computational Linguistics.
- [Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- [Qian et al.2019] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.
- [Rach et al.2018] Niklas Rach, Saskia Langhammer, Wolfgang Minker, and Stefan Ultes. 2018. Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*.
- [Rakshit et al.2019] Geetanjali Rakshit, Kevin K Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2019. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*, pages 45–52. Springer.
- [Rashid et al.2019] Ahmad Rashid, Alan Do-Omri, Md Haidar, Qun Liu, Mehdi Rezagholizadeh, et al. 2019. Bilingual-gan: A step towards parallel text generation. *arXiv preprint arXiv:1904.04742*.
- [Schieb and Preuss2016] Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at Fukuoka, Japan*, pages 1–23.

- [Schmidt and Wiegand2017] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10.
- [Silva et al.2016] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*.
- [Silverman et al.2016] Tanya Silverman, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*, pages 1–54.
- [Tekiroğlu et al.2020] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- [Wachsmuth et al.2018] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.
- [Wang and Wan2018] Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.
- [Warner and Hirschberg2012] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.
- [Waseem and Hovy2016] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- [Wright et al.2017] Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Surviving the Legal Jungle: Text Classification of Italian Laws in extremely Noisy conditions

Riccardo Coltrinari

Computer Science dept.
University of Camerino
Camerino (MC)

$\left\{ \begin{array}{l} \text{riccardo.coltrinari} \\ \text{alessandro.antinori} \end{array} \right\} @ \text{studenti.unicam.it}$

Alessandro Antinori

Computer Science dept.
University of Camerino
Camerino (MC)

Fabio Celli

Research and Development
Maggioli S.p.A.
Santarcangelo (RN)

`fabio.celli@maggioli.it`

Abstract

In this paper, we present a method based on Linear Discriminant Analysis for legal text classification of extremely noisy data, such as duplicated documents classified in different classes. The results show that Linear Discriminant Analysis obtains very good performances both in clean and noisy conditions, if used as classifier in ensemble learning and in multi-label text classification.

1 Motivation and Background

We address text categorization of business-oriented legal documents in Italian, but with a custom and overlapping hierarchy of product categories. A typical approach to tackle similar tasks is to exploit resources such as EUROVOC (Daudaravicius, 2012), a multilingual thesaurus consisting of over 6700 hierarchically-organised class descriptors used by many organizations of the European Union (EU) for the classification and retrieval of official documents. Our editorial system has a hierarchy of 23 product categories and more than 20600 labels, manually annotated and customized for different clients in more than 15 years, hence it is not possible to exploit resources like EUROVOC to categorize documents.

In this paper, we propose a fast and efficient method for document classification for noisy data based on Linear Discriminant Analysis, a dimensionality reduction technique that has been employed successfully in many domains, including neuroimaging and medicine. We believe that our contribution will be useful to the NLP community in the context of document categorization as

well as automatic ontology population, in particular when dealing with very noisy data.

The paper is structured as follows: in Section 1.1 we present the related works in the field of text classification and the potential of Linear Discriminant Analysis, in Section 2 we describe the datasets we used, in Section 3 we report and discuss the result of our classification experiments and in Section 4 we draw our conclusions.

1.1 Related Work

There are many applications of NLP in the legal text domain, such as the creation of ontologies for knowledge extraction (Lenci et al., 2009) or legal reasoning (Palmirani et al., 2018), other tasks include dependency parsing (Dell’Orletta et al., 2012), deception detection (Fornaciari et al., 2013) and semantic annotation exploiting external resources like FrameNet (Venturi, 2011). In this domain, the most popular way to perform text categorization is to use ontologies: for example many used EUROVOC to label documents in several languages (Steinberger et al., 2013) with one label for each document, in order to train SVMs (Boella et al., 2013) or deep learning models (Caled et al., 2019), for the prediction of labels at different levels of granularity in the label hierarchy. Another approach is to use the judgments of the Supreme Court as gold standard labels, thus reducing the complexity of the task, and then train machine learning models, such as SVMs, to perform classification (Sulea et al., 2017). It is known that active learning does not reach a good performance in the legal domain (Cardellino et al., 2015), but it is possible to align different resources to perform ontology population or expansion (Cardellino et al., 2017). The state-of-the-art in text classification ranges from 40% to 85% or more, depending on the complexity and size of the dataset, and from the number of document classes (Adhikari et al., 2019). The results of a noise introduction simula-

Copyright ©2020 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tion study revealed that substituting up to 40% of words with random text strings yields to a small decrease in text classification performance, while the substitution of more than 40% of the text yields a dramatic decrease in classification performance (Agarwal et al., 2007).

A similar task, Extreme Multi-Label Text Classification (XMTC), consists in the classification of documents annotated with multiple tags. Recent experiments of XMTC with Convolutional Neural Networks on a dataset of 57k legal documents annotated with multiple concepts from EU-ROVOC, revealed that word embeddings extracted with label-wise attention Networks (Mullenbach et al., 2018) leads to the best overall performance, compared pre-trained word embeddings, Hierarchical word embedding and Max-Pooling Scorers that produce section-based word embeddings (Chalkidis et al., 2019). It has been demonstrated in more than one context that cNNs perform well for text categorization, but also that there is no single algorithm that performed the best across the combination of data sets and training sample sizes (Keeling et al., 2019). The rationale behind the good performance of label-wise attention networks is their ability to maximise the difference of the words/features associated to different labels. A very similar -but faster- approach is Linear Discriminant Analysis (Balakrishnama and Ganapathiraju, 1998), a feature selection and classification technique that has been successfully used for the incremental classification of large streams of data (Pang et al., 2005), to find identity patterns in images before the advent of deep learning (Prince and Elder, 2007) and as feature selection technique for discriminating fMRI response patterns to visual stimuli (Mandelkow et al., 2016).

Linear Discriminant Analysis (henceforth LDA) is a widely accepted dimensionality reduction and classification method, which aims to find a transformation matrix to convert a feature space to a smaller space by maximising the between-class scatter matrix while minimising the within-class scatter matrix (Boroujeni et al., 2018). The criticism towards this technique emphasize the fact that it suffers from the domination of the largest objectives, in particular when close class pairs tend to overlap in a feature subspace, but this can be solved with various optimizations, including eigenvalue decomposition, among others (Li et al., 2017).

2 Data

Our dataset consists of 2030 legal italian documents with an average of 800 words each. We have 23 classes representing products manually annotated over 15 years, every document is categorized in one or more classes. Classes are not balanced, but their distribution is proportional to the whole editorial system, that consists of 443.7k documents. We extracted such a small dataset from the editorial system because we plan to update our models very frequently, using a small portion of documents each time in order to save computational power and time. Figure 1 reports the distribution of the classes in our dataset.

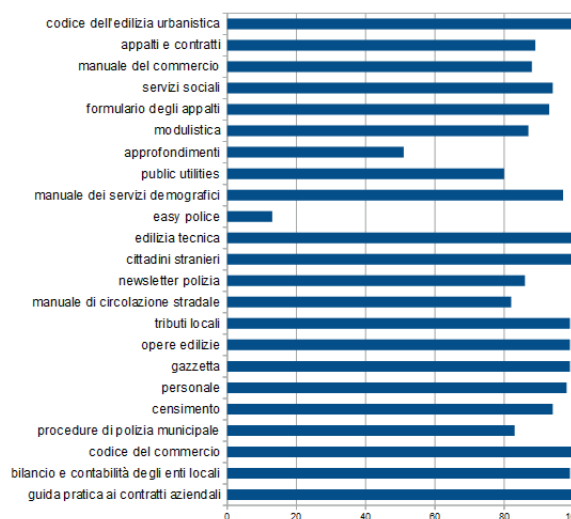


Figure 1: Distribution of the classes in our dataset.

Since documents can fall under more than one class, we have 43% of documents repeated under different classes. We tested the performance of different classifiers under two different conditions: noisy (with repeated documents) and clean (without the repeated documents).

3 Experiments and Discussion

In both cases (noisy and clean) we performed preprocessing on text, deleting punctuation and Italian stopwords. We did not use stemming or lemmatization since their usage has led to a degradation of results. We formalize the task in two ways: a simple multinomial classification, where we train a classifier to predict one class per document, and a multi-label classification, where we produce a score ranking of labels for each document and evaluate if the gold standard label occurs in the first N positions.

features	algorithm	noisy acc(10f-cv)	noisy acc(split)	clean acc(10f-cv)	clean acc(split)
baseline	majority (zeroR)	4.6%	2.9%	8.3%	6.9%
200 glove embeddings	cNN	5.1%	4.6%	10.6%	10.8%
200 glove embeddings	rNN	19.6%	21.7%	33.6%	31.0%
200 glove embeddings	bayesian network	9.6%	9.6%	25.1%	23.8%
200 glove embeddings	naïve bayes	7.6%	9.0%	13.6%	14.9%
200 glove embeddings	SVM	6.4%	3.9%	11.0%	10.2%
200 glove embeddings	random forest	26.4%	27.7%	49.5%	50.4%
200 glove embeddings	LDA	24.8%	25.9%	52.5%	49.0%
4700 words tf-idf	bayesian network	34.4%	32.5%	58.7%	55.6%
4700 words tf-idf	naïve bayes	28.6%	26.9%	51.9%	46.8%
4700 words tf-idf	SVM	38.7%	38.9%	76.9%	74.2%
4700 words tf-idf	random forest	36.8%	36.9%	69.9%	67.3%
200 words selected LDA	bayesian network	34.2%	32.5%	56.2%	56.5%
200 words selected LDA	naïve bayes	29.0%	26.4%	45.6%	48.1%
200 words selected LDA	SVM	37.9%	38.0%	71.0%	70.3%
200 words selected LDA	random forest	37.5%	39.0%	73.5%	71.7%
200 words selected LDA	LDA	35.2%	34.3%	66.9%	64.5%
200 words selected corr	bayesian network	31.5%	30.2%	55.4%	57.3%
200 words selected corr	naïve bayes	22.3%	19.2%	41.8%	46.8%
200 words selected corr	SVM	34.4%	33.6%	69.4%	68.6%
200 words selected corr	random forest	36.1%	36.2%	70.7%	70.3%
200 words selected corr	LDA	33.5%	32.6%	64.1%	60.3%
23 feat LDAclass from 200 w selected LDA	naïve bayes	46.4%	47.6%	67.2%	57.0%
23 feat LDAclass from 200 w selected LDA	SVM	60.0%	58.9%	87.5%	88.3%
23 feat LDAclass from 200 w selected LDA	random forest	52.5%	54.5%	89.3%	88.3%

Table 1: Results of the multinomial text classification with different settings (200 GloVe embeddings, 4700 words tf-idf, 200 words LDA feature selection, 200 words correlation feature selection, 23 LDA predictions as features), algorithms (cNN, rNN, bayesian network, naïve bayes, SVM, random forest and LDA classifier), datasets (noisy, clean) and evaluation methods (10-fold Cross Validation, 70%-30% training test split). The best results for each feature setting are marked in bold.

3.1 Multinomial Classification

We tested different feature settings and algorithms with 10-fold cross validation (10f-cv) and 70%-30% training-testing split in the clean and noisy dataset conditions. Table 1 reports the results in terms of accuracy, that is to say the percentage of documents correctly classified. In both conditions the majority baseline is very low, ranging from 4.6% to 8.3%. First we experimented with pre-trained GloVe word vectors as features (vector size 200). As a matter of fact the GloVe Project provides word vectors of different dimensions for words representation trained on massive web datasets (Pennington et al., 2014). For instance the word vectors we used here have been pre-trained by the GloVe Project from two massive corpora, Wikipedia 2014 and Gigaword 5. As we can see in Table 1 in the GloVe embeddings setting we used the following classification algorithms: cNN (with 2 convolutional layers with ReLU activation, 1 pooling layer and 1 output layer), rNN (with 1 rNN sequence layer, 1 LSTM layer with tanH activation and 1 rNN output layer), bayesian networks, naïve bayes, SVMs, random forest and LDA. In general, Deep Learning algorithms suffer from the small data used for the experiment,

but surprisingly, cNNs performed badly and rNNs worked better, indicating that the sequentiality of text plays an important role. Among the other classification algorithms it turned out that random forest and LDA obtained the best performances, proving that the ability of the algorithm to generalize is crucial. The general low accuracies obtained with these features might indicate that the contexts of our documents represented by word embeddings are not very discriminative. The results increased significantly in the classification with the TF-IDF scores of 4700 words, especially with SVMs as algorithms. This suggests that using more features brings better results without overfitting the data, as shown by the similar accuracies obtained with a 10-fold cross validation and with training-test split. Next we experimented with feature selection, using LDA and Pearsons' correlations to select the best 200 words for the prediction. Results show that, in this feature setting, random forests are the best classification algorithm and that LDA outperforms correlations as feature selection algorithm. Furthermore, as can be seen in the last part of Table 1, we were able to reach state-of-the-art results with an ensemble learning scheme: using LDA as a classifier we transformed

features	algorithm	noisy acc(10f-cv)	noisy acc(split)	clean acc(10f-cv)	clean acc(split)
baseline	majority (zeroR)	4.6%	2.9%	8.3%	6.9%
500 words per label tf-idf selected	scoreranking LDA (1 label)	56.7%	59.9%	52.9%	52.9%
500 words per label tf-idf selected	scoreranking LDA (2 labels)	62.5%	64.2%	63.9%	62.5%
500 words per label tf-idf selected	scoreranking LDA (3 labels)	66.8%	67.5%	68.2%	67.5%
500 words per label tf-idf selected	scoreranking LDA (4 labels)	70.7%	70.3%	73.2%	73.0%
500 words per label tf-idf selected	scoreranking LDA (5 labels)	74.2%	74.4%	76.3%	76.7%
500 words per label tf-idf selected	scoreranking LDA (6 labels)	79.4%	77.0%	79.4%	78.1%
1000 words per label tf-idf selected	scoreranking LDA (1 label)	56.7%	54.7%	53.9%	62.2%
1000 words per label tf-idf selected	scoreranking LDA (2 labels)	63.6%	61.0%	63.8%	67.2%
1000 words per label tf-idf selected	scoreranking LDA (3 labels)	67.9%	65.4%	68.4%	72.9%
1000 words per label tf-idf selected	scoreranking LDA (4 labels)	71.7%	69.2%	72.8%	78.9%
1000 words per label tf-idf selected	scoreranking LDA (5 labels)	75.0%	73.1%	76.6%	82.6%
1000 words per label tf-idf selected	scoreranking LDA (6 labels)	77.9%	77.4%	83.7%	85.2%

Table 2: Results of the text classification with different feature settings (500 or 1000 words per label), number of labels in the ranking evaluated (1 to 6 labels), datasets (noisy, clean) and evaluation methods (10-fold Cross Validation, 70%-30% training test split). The best results for each feature setting are marked in bold.

the initial space of 200 word features, previously selected with LDA, in a space of 23 binary features corresponding to the final classes. On top of that we applied different classification algorithms, finding that SVM is the best performing one in the noisy dataset while random forest obtained the best performance in the clean dataset.

3.2 Multi-Label Classification

The Multi-Label classification task is structured as follows: for each document label in the training set, we create a Bag-of-Words (BoW) from the words of its associated documents, then we use TF-IDF scores to weight every word within the BoW obtaining a word ranking that we use for feature selection, since words with higher values better characterize a particular label. Then we apply LDA classification, but unlike the previous experiment, here the prediction returns a list of all the labels, ordered by the total score achieved, we call score ranking this algorithm. Since the classifier returns a list as an outcome, but the editors (our customers) want to choose one or more label from this list, we have to evaluate if the gold standard label occurs in the returned list, thus we can assign multiple labels to a document and test whether the original one is present or not. In this sense, the Score Ranking classifier is evaluated as a Multi-Class classifier (so the metrics in Table 2 are actually Hit@N metrics where N is the size of the returned list), but the returned list is used by the end users to simulate a Multi-Label functionality, leaving to the editors the choice of the best labels to assign among the ones returned. The result of this experiment, reported in Table 2, shows that the performance with 1 label is in line with the ensem-

ble learning setting of the Multinomial classification, but the score ranking system only works well in the noisy dataset, as the results are very similar in both noisy and clean conditions. The performance increases at an average of +3.9% when keeping more than one label. In general, we observe that using 500 or 1000 words per label yield similar results in our small dataset, but using more words can help to capture more nuances in text, that might be useful in larger sets of documents. We also observe that 1000 words per label increase the results in the clean condition, while 500 words per label are enough in the noisy condition.

4 Conclusion and Future

We experimented with various settings, feature selection methods and classification algorithms, and we found a method to extract good models in extremely noisy conditions, even with documents repeated under different labels. LDA proved to be a valuable classification and feature selection technique, but we obtained the best performances when LDA is combined with other algorithms. The results we obtained with the score ranking classification are in line with the state-of-the-art, but our method is more suitable for small and noisy datasets. In the future we plan to apply the score ranking algorithm on a larger dataset and to use it in a real multi-label environment comparing the results with the state-of-the-art of Extreme Multi-Label Document Classification (Chalkidis et al., 2019). We also plan to make comparisons with the more recent state of the art deep learning techniques and to apply semantic indexing to the documents to check for improvements.

References

- Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.
- Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. 2007. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 3–12. IEEE.
- Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. In *Institute for Signal and information Processing*, volume 18, pages 1–8.
- Guido Boella, Luigi Di Caro, Daniele Rispoli, and Livio Robaldo. 2013. A system for classifying multi-label text into eurovoc. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 239–240.
- Forough Rezaei Boroujeni, Sen Wang, Zhihui Li, Nicholas West, Bela Stantic, Lina Yao, and Guodong Long. 2018. Trace ratio optimization with feature correlation mining for multiclass discriminant analysis. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Danielle Caled, Miguel Won, Bruno Martins, and Mário J Silva. 2019. A hierarchical label network for multi-label eurovoc classification of legislative contents. In *International Conference on Theory and Practice of Digital Libraries*, pages 238–252. Springer.
- Cristian Cardellino, Serena Villata, Laura Alonso Alemany, and Elena Cabrio. 2015. Information extraction with active learning: A case study in legal text. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 483–494. Springer.
- Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Ontology population and alignment for the legal domain: Yago, wikipedia and lkif. In *International Semantic Web Conference: Posters Demos and Industry Tracks*, pages 1–4.
- Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in eu legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87.
- Vidas Daudaravicius. 2012. Automatic multilingual annotation of eu legislation with eurovoc descriptors. In *EEOP2012: Exploring and Exploiting Official Publications Workshop Programme*, page 14.
- Felice Dell’Orletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, and Giulia Venturi. 2012. The splat-2012 shared task on dependency parsing of legal texts. In *Semantic Processing of Legal Texts (SPLaT-2012) Workshop Programme*, page 42.
- Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6. IEEE.
- Robert Keeling, Rishi Chhatwal, Nathaniel Huber-Fliflet, Jianping Zhang, Fusheng Wei, Haozhen Zhao, Shi Ye, and Han Qin. 2019. Empirical comparisons of cnn with other learning algorithms for text classification in legal document review. *arXiv preprint arXiv:1912.09499*.
- Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2009. Ontology learning from italian legal texts. *Law, Ontologies and the Semantic Web*, 188:75–94.
- Zhihui Li, Feiping Nie, Xiaojun Chang, and Yi Yang. 2017. Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2100–2110.
- Hendrik Mandelkow, Jacco A de Zwart, and Jeff H Duyn. 2016. Linear discriminant analysis achieves high classification accuracy for the bold fmri response to naturalistic movie stimuli. *Frontiers in human neuroscience*, 10:128.
- James Mullenbach, Sarah Wiegrefe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.
- Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. 2018. Pronto: Privacy ontology for legal reasoning. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 139–152. Springer.
- Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. 2005. Incremental linear discriminant analysis for classification of data streams. *IEEE transactions on Systems, Man, and Cybernetics, part B (Cybernetics)*, 35(5):905–914.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.
- Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2013. Jrc eurovoc indexer jex-a freely available multi-label categorisation tool. *arXiv preprint arXiv:1309.5223*.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Giulia Venturi. 2011. Semantic annotation of italian legal texts: a framenet-based approach. *Constructions and Frames*, 3(1):46–79.

Clustering verbal Objects: manual and automatic procedures compared

Ilaria Colucci

University of Pavia

Department of Humanities

[colucci.ilaria03@gmail.com]

Elisabetta Jezek

University of Pavia

Department of Humanities

[jezek@unipv.it]

Vít Baisa

Lexical Computing Ltd.

Czech Republic

[vit.baisa@gmail.com]

Abstract

As highlighted by Pustejovsky (1995, 2002), the semantics of each verb is determined by the totality of its complementation patterns. Arguments play in fact a fundamental role in verb meaning and verbal polysemy, thanks to the sense co-composition principle between verb and argument. For this reason, clustering of lexical items filling the Object slot of a verb is believed to bring to surface relevant information about verbal meaning and the verb-Objects relation. The paper presents the results of an experiment comparing the automatic clustering of direct Objects operated by the agglomerative hierarchical algorithm of the Sketch Engine corpus tool with the manual clustering of direct Objects carried out in the T-PAS resource. Cluster analysis is here used to improve the semantic quality of automatic clusters against expert human intuition and as an investigation tool of phenomena intrinsic to semantic selection of verbs and the construction of verb senses in context.

Keywords: Clustering, verbal Objects, Italian, Semantic Types

1 Introduction

Clustering techniques have been used extensively in recent decades in Linguistics and NLP, especially in Word Sense related tasks. As a matter of fact, partitioning data sets on the basis of their similarity at a distributional level clarifies the meaning of lexical elements (Brown et al., 1991). Partitioning verbal arguments, for example, can be beneficial to investigate the sense properties they share but also to explore verbal meaning.

In fact, as highlighted by Pustejovsky (1995, 2002), the semantics of each verb is determined by the totality of its complementation patterns and arguments play a fundamental role in verb meaning and verbal polysemy, thanks to the sense co-composition principle. *Id est*, the process of *bilateral semantic selection* between the verb and its

complement gives rise to a novel sense of the verb in each context of use (*ibidem*).

Clustering lexical items filling the argument positions of a verb is then believed to bring to surface relevant information about verbal meaning and the verb-arguments relation. Clustering them, and especially direct Objects in pro-drop languages such as Italian, allows hence to investigate how to better induce, discriminate and disambiguate verb senses. Because argument fillers share the same semantic nature, they can be grouped and generalized with respect to their content and be associated with semantic types, i.e. empirically identified semantic classes representing selectional properties and preferences of verbs.

Clustering of Objects can therefore be used as a survey tool for the intrinsic phenomena of semantic classes and, at the same time, as an object of investigation to improve the clustering automatic models themselves against human partitioning. This paper presents the results of an experiment comparing manual and automatic clustering of Italian Object fillers to be used in verb-sense identification and, along with it, it describes the linguistic phenomena that emerged from the semantic analysis of non-supervised clusters. The comparison concerns the agglomerative hierarchical clustering algorithm of the Sketch Engine corpus tool¹ (Kilgarriff et al., 2014) and the manual clustering carried out in the T-PAS resource² (Jezek et al., 2014), in which verbal senses are identified in context based on the fillers of the argument positions (see section 1.1) and are annotated with a semantic type (ST; see section 1.2) able to identify them. Thanks to their semantic generalization properties, ST are also believed to represent a useful comparative tool between manual and automatic clustering. After presenting the theoretical background of the research, section 2 will cover

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹ www.sketchengine.eu/

² tpas.fbk.eu

data, method and work pipeline, while clustering evaluation via metrics and linguistic analysis will be presented in section 3.

1.1 Clustering verbal Objects fillers

Clustering is a Data Mining task (Kotu & Deshpande, 2014) in which a grouping process of a set of objects is carried out, obtaining clusters of elements which are similar to each other but dissimilar from the objects of other groups (Xu & Wunsch, 2008). In most implementations, clustering is used with an exploratory function, i.e. it is a technique applied to data sets for which there is no *a priori* knowledge concerning the set membership of the samples (Lavine and Mirjankar, 2006). In these cases, clustering is therefore considered as a non-supervised procedure with the aim of providing an insight into the studied data. However, it can be considered a supervised method and regarded as a classification task when a manually created benchmark (a *ground truth*) is used to assess the output of the clustering (Bishop, 1995). The manually created partition or the manually defined set of classes is used to validate the groupings proposed by the automatic algorithm, through a process defined as external clustering evaluation (Gan, Ma and Wu, 2007). The idea behind this paper is to operate through a procedure very similar to external evaluation in which the manual clustering and the automatic one taken into consideration are mutually compared; but yet here the aim is not to validate the automatic model but more to bring out matches and differences between the partitioning criteria at the basis of the supervised clustering and the unsupervised one.

The supervised clustering under consideration here was performed on the lexical items that fill different argument positions in T-PAS, a resource of predicate-argument structures for Italian obtained from corpora (Ježek et al., 2014). T-PAS contains, for each argument slot, the specification of the semantic class to which the fillers found in that position in the corpus belong. We considered the direct Object clusters, which therefore contain the fillers that occupy that slot in the various occurrences of the corpus. To clarify this, given the following sense for verb *pilotare* (to pilot), the related cluster for the Object position will appear as follows:

(1) *pilotare*

1. [Human] *pilotare* ([Flying Vehicle] | [Water Vehicle])

pilotare_clust1: {macchina (car), moto (motorbike), barca (boat), caccia (fighter aircraft), nave (ship)}

The ST defined for the direct Object slot can thus also be used as a label to semantically identify what is contained in the cluster.

As for automatic clustering, in our comparison we used the built-in clustering function (Baisa et al., 2015) in the Sketch Engine tool (SkE). The model is based on a hierarchical agglomerative algorithm that compute the distributional similarity³ between the Object fillers and groups them in an unsupervised way, starting from a minimum similarity value given to the algorithm (Kilgariff et al., 2014). Clusters creation starts with computing Word Sketches, i.e. automatic, corpus-based summaries of a word's grammatical and collocational behaviour (Kilgariff et al., 2004). The results concerning the direct Object are then grouped through a bottom-up process in which clusters are populated through pairings of words. The inclusion and exclusion criterion is a minimum default value of 0.15⁴ for distributional similarity. The clusters created in Sketch Engine for *pilotare* are the followings, for which, unlike T-PAS, ST labelling is not available:

(2) *pilotare_clust1*: {nave (ship), barca (boat)}
pilotare_clust2: {macchina (car), moto (motorbike)}
pilotare_clust3: {caccia (fighter aircraft)}⁵

The main difference between T-PAS and SkE clustering procedures are the semantic-distributional criteria on which they are based. T-PAS approach can be defined as verb-oriented: Objects are primarily clustered on the basis of their verbal distributional behaviour and ability to activate a given verbal sense as direct objects. Since all fillers occupying a given slot for a given sense share the same relation with the verb, they can be ontologically and semantically generalized with an ST on the basis of their common semantic traits. This generalization allows to make the verbal selectional constraints visible. On the contrary, SkE

³ See Kilgariff et al. (2015) for statistics and technical details on similarity computing.

⁴ We also conducted a similarity value manipulation experiment, which confirmed what discussed in detail in section 3.

⁵ Clusters consist at least of 1 word and up to 1000.

performs noun-based clustering: it takes into account the general distributional behaviour of fillers, not merely the verbal one. In the process of creating sets, each filler behaviour is weighed against the entire reference corpus and with respect to the frequencies of appearance in different contexts. The elements clustered together in SkE are therefore not only similar in their sense and behaviour as direct objects, but also respect to the whole nominal class they belong to.

1.2 T-PAS System of Semantic Types

As mentioned above, in T-PAS argument slots are linked to ST labels, semantic classes able to generalize over the sets of lexical items in argument positions found in the corpus (Ježek et al., 2014). The labels belong to the System of Semantic Types (see Figure 1 for an excerpt), a hierarchical structure of semantic categories achieved by performing the CPA procedure (Hanks, 2004), on the evidence of 1200 Italian verbs (Ježek, 2019), i.e. through the manual analysis of examples in corpora of slots’s fillers and their co-occurrence statistics. They characterize a group of lexical elements with respect to their content, defining also a criterion of similarity and dissimilarity on which T-PAS clusters are created. STs are used here as a reference for the comparison of the two clustering models, for the verification of the clusters internal semantic quality.

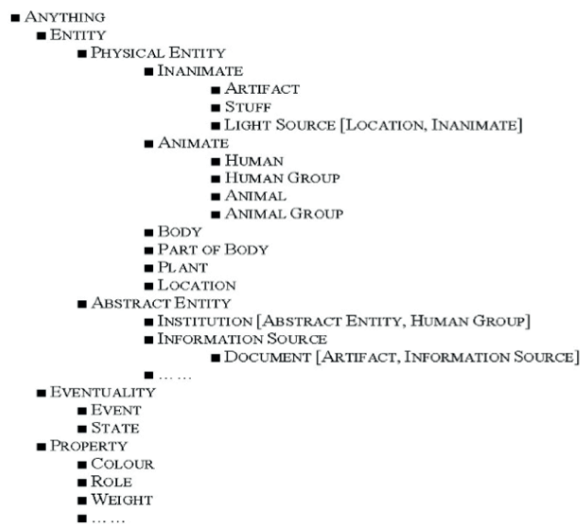


Figure 1: Top-level nodes and a selection of *leafs* from the ST System (Ježek, 2019)

2 Data and method

The research has been developed through a pipeline organized according to the following steps:

1. *Data extraction*: Data for both clusterings are extracted from the web crawled corpus ItWac reduced (Baroni et al., 2009). In this early stage the clusters of Object fillers for each verb included in T-PAS are extracted from the corpus annotated lines, while for Sketch Engine, the clusters are extracted for all verbs present in the ItWac corpus. All lines in the corpus are then scanned and verbal Objects are mapped through the condition: *OBJ* = *post verbal noun* (PostV_N). Since T-PAS does not annotate individual fillers as such but only works at verb and sentence level, this function is also used to retrieve its Objects.

2. *Data intersection*: The obtained clusters are intersected with each other in order to obtain a database in which there are sets for the same verbs and containing the same fillers, to focus on how the two models carried out the partition.

3. *Data filtering*: In this step the database is cleared from:

- verbs with structures recognized as complex and non-compositional, i.e. idiomatic constructions;
- verbs with the ST [Anything] (top node in Fig. 1) in the object slot, as it does not entail selection restrictions within the T-PAS clusters;
- verbs with Object clusters with more than 29 internal elements.

At the end of the filtering process the clusters of the two models are aligned with respect to the STs, i.e. all possible STs signaled in T-PAS for the Object of a verb are treated as a single set of semantic conditions, in order to analyze the internal quality of SkE clusters through them. The aligned structure of the verb *acquire* (to acquire) in Figure 2 is given here as an example.

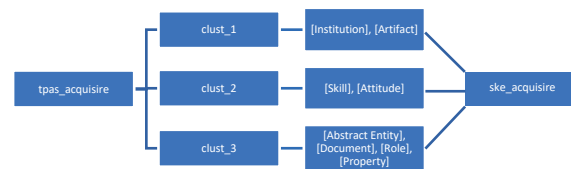


Figure 2: Aligned STs structure of verb *acquire*

The final database comes to a total of 397 verbs and 3938 clusters, including both T-PAS and SkE clusters. We provide an illustrative table (Table 1)

showing the first and last verb among those analyzed and the information on their respective clusters: *abbagliare* (to dazzle) and *votare* (to vote).

Verb	Source	Clusters	Clustered items	Items nr.
abbagliare	T-PAS	1	nemico, viso, utente	3
abbagliare	T-PAS	2	cliente, inquisitore, uomo, visitatore	4
abbagliare	SkE	occhio	nemico, uomo	2
abbagliare	SkE	inquisitore	inquisitore	1
abbagliare	SkE	pilota	visitatore, cliente, utente	3
abbagliare	SkE	viso	viso	1
votare	T-PAS	1	barzelletta, foto, poesia, sito	4
votare	T-PAS	2	riduzione, emendamento, legge	3
votare	SkE	barzelletta	barzelletta	1
votare	SkE	post	poesia, foto	2
votare	SkE	sito	sito	1
votare	SkE	mozione	emendamento	1
votare	SkE	risoluzione	legge	1
votare	SkE	bilancio	riduzione	1

Table 1: Complete set of clusters of the first and last verb of the data set

3 Clustering evaluation

3.1 SkE clustering evaluation

To verify the compatibility between the two clusterings, the similarity between the two partitions has been evaluated through different metrics able to offer an external evaluation of the unsupervised model. To account for both the presence of common pairings, as well as the homogeneity and completeness of the clustering, the following metrics were considered: Fowlkes & Mallow Index (F&M), Adjusted Rand Index (ARI), Homogeneity, Completeness.

F&M, as the geometric mean between precision and recall, was used to verify the similarity between the two models from how many partition pairings are in common. This index also allows to better balance the possible noise or unrelatedness between clustering (Fowlkes & Mallows, 1983). ARI (Hubert & Arabie, 1985) always gives information on the overlapping of the two clusterings in comparison but balances the very large number of clustered elements in T-PAS (Romano et al., 2016). Homogeneity and completeness metrics (Rosenberg & Hirschberg, 2007) are helpful to better investigate the internal content of the SkE clusters. They allow to highlight a possible internal structure, hierarchically and semantically coherent with the taxonomy identified for ST. Ho-

mogeneity evaluates if all automatic clusters created contain only elements that are members of a single class in the manual reference. Completeness, instead, evaluates if all the objects that are members of a given cluster in SkE are elements of the same cluster in T-PAS.

As reported by their respective creators, all metrics have an optimal result range between 0 and 1. The possible results between these two limits can be classified with respect to the greater or lesser proximity to the optimal limit: the results closer to 1 denote greater similarity of output between the two models, the results closer to 0 instead less similarity (Gan, Ma and Wu, 2007).

In this sense, we can define three bands of possibilities, coherently with the approach *the higher the better* generally used in cluster analysis: from 0.01 to 0.399, the clustering compared to the golden standard is highly different, from 0.4 to 0.699 the result and the correspondence is medium-good, while the results above 0.7 and up to 0.999 are the ideal ones, which indicate a marked correspondence between the compared models. However, since metrics such as F&W and ARI have shown the lack of partitions in higher ranges (the first beyond 0.82, the latter beyond 0.7), we choose to consider the whole group of medium good results between 0.4 and 1. The absence of the higher ranges stands for low compatibility between the two models.

Metric	Clusters in the [0.4-1] range
Adjusted Rand Score	11.08%
Fowlkes & Mallow	36.18%
Homogeneity	94.96%
Completeness	41.31%

Table 2: Metrics results

As we see in Table 2, what we find in fact is a situation of only limited correspondence between the two clustering, with a rather low overlap and similarity as indicated by the ARI and the F&M, even with the internal noise balance. At least two reasons may be behind the scarce similarity: the tendency of SkE to create small fine-grained clusters populated by few elements that give more weight to specificity than to generalization capacity; the fact that in T-PAS for a given verb sense the Object slot can be compatible with more STs (see (1)), and such STs can also be hierarchically distant in the general system of labels. This leads to clusters containing fillers able to activate a

given verbal sense but which are quite heterogeneous among themselves and semantically dissimilar, with respect to the rest of the distributional relations between the fillers. An example can be the verb *trasportare* (to transport), which has as T-PAS cluster for the first sense a set of 18 Objects (see (4)); such fillers belong to three different STs: [Inanimate], [Animate] and [Energy]. The latter ST, [Energy], is hierarchically distant to the others since it has a different parent node than [Animate] and [Inanimate], which both pertain to a lower level in the hierarchy.

(3) *trasportare*

1. [Human] | [Vehicle] | [Watercourse] *trasportare* [Inanimate] | [Animate] | [Energy]

(4) *trasportare_clust1*: {acqua (water), alimento (nourishment), animale (animal), arma (weapon), bene (asset/good), bicicletta (bicycle), cadavere (corpse), cibo (food), gas (gas), gommone (inflatable raft), macchina (machine), oggetto (object), peso (weight), student (student), terra (soil), traffico (traffic), viaggiatore (traveller), visitatore (visitor)}

It is clear that a fine-grained algorithm, not able to generalize at a higher level as in T-PAS, will divide fillers labelled with [Animate] or [Inanimate] from those labelled with [Energy]. In fact, SkE for the same verb creates 12 clusters.

As shown in Table 2, the results of the Completeness are in line with what has just been discussed for ARI and F&M: only in 40% of the cases all members of a T-PAS cluster are members of a single SkE cluster. These are generally small or medium sized clusters with only one associated ST or with hierarchically close alternative ST structures. Homogeneity highlights the primary characteristic of SkE clusters and the algorithm: it is preferable to create smaller but internally purer clusters, rather than larger sets with members of other classes. This implies the creation in SkE of semantically specific clusters, that privilege the inter-relation between Object fillers but not the higher semantic level between Object fillers and verb.

From a different perspective, we can say that the noun-oriented criteria of clustering and the verb-oriented ones tend to converge when we consider small clusters, in which the elements belonging to a set in SkE generally belong to the same set in T-PAS.

As for wide clusters, they are particularly rare in SkE and tend to be smaller in size than T-PAS anyway. Their content also seems to be dependent on various factors on which the linguistic analysis has shed light.

3.2 Linguistic analysis of the clusters

To verify the nature of the diversity between the two clusterings measured with the metrics reported in 3.1, a detailed analysis of the lexical-semantic phenomena visible internally to the clusters was carried out considering:

- The consistency, for automatic clusters, with one and only one of the aligned T-PAS STs, i.e. the precision and purity at the semantic level of clusters compared to the generalization of the ST;
- Internal homogeneity, i.e. whether the clusters meet verb-sense oriented or noun-sense oriented criteria and, if the latter, whether the cluster items are linked by syntagmatic relationships and there is some kind of affinity or implication between them. Thus, the types of semantic relations present between the words are identified;

- The overlap between clusters with respect to the ARI, and in relation to cluster size and clustering difficulty depending on several STs possible for the same slot;

- The problem of incorrect mapping as Objects of postverbal Subjects, subjects of inaccusative verbs, structures with *si* particle (e.g. reflexive, impersonal), i.e. the clusters' internal noise.

The research has shown that SkE clusters tend to be small-medium sized, semantically homogeneous, often able to isolate very specific semantic relations. They are generally not consistent *per se* with the verb sense identified by T-PAS but create partitions: a) usually of medium size and consistent with only one parallel ST, b) single element groups that generally belong to a higher level of specificity or to a different semantic domain, and c) groups that are inconsistent with the sense of the verb but cluster words on the basis of the following criteria:

- Belonging to the same domain (e.g. informatics for *distribuire* {software, applicazione});
- Being part of the same ST, but as very specific instances, not separated by the T-PAS hierarchy (e.g. {abbazia, monastero, santuario} for *saccheggiare* and the type [Location]);
- The possibility of a conceptual association or affinity (e.g. {seminario, incontro, seduta} for *organizzare*);

- Purely distributional parameters and undefined semantic relations (e.g. in *gestire* {contenuto, caso});
- A relationship of synonymy or meronymy (e.g. {spinta, propensione} for *frenare* or for *fratturare* {dito, mano, braccio}); antonymy, hyponymy, hyponymy are generally represented by different clusters.

The parameters of consistency, internal homogeneity and overlapping between the models seem to relate to the same factors: first, the size of the clusters, i.e. how many clustered elements are part of the set; second, the structure of STs possible for the Object (see (5)), i.e. if for the same slot only one ST is possible, if several alternatives are available or, also, if a lexical set is signaled in the T-PAS annotation - that is, if among the fillers a set of lexical elements is present that has high frequency or has the typical behaviour of a collocation (e.g. {messaggio | ricordo} in (5)). This is relevant since the computation of SkE starts precisely from the frequency and collocational behaviour of a word.

(5) *cancellare* (sense description: to eliminate, to make inexistent):

1. [Human] | [Inanimate1] | [Abstract Entity1] | [Eventuality1] *cancellare* [Inanimate2] | [Abstract Entity2 {messaggio | ricordo}] | [Eventuality2]

The third relevant factor is hierarchical proximity, i.e. if STs possible for a slot are sisters of the same parent node between the types present in the hierarchy (see (6)).

(6) no proximity: [Animate] vs. [Institution]
in proximity: [Command] vs [Request]

The clusters of SkE, even if not corresponding to those of T-PAS, are rated totally consistent or at 70-80% consistent with one of the aligned STs 62% of the times; in the remaining 38% of cases, there is a significant number of clusters that can be generalized with a ST. Consistency is more difficult to reach if a given sense is annotated in T-PAS with several alternative STs or STs and lexical sets co-presenting; SkE can produce new combinations in which fillers corresponding to different STs are included in the same cluster. Very frequently SkE atomizes the set of fillers in nuclear clusters, made up of only one or two elements that are necessarily consistent with one ST but are not of much help to the study of semantic relations.

As said, hierarchical proximity of STs and the size of the cluster can influence its handling: if the possible STs for the Obj-slot are hierarchically distant and the T-PAS cluster is small, the SkE outcome will tend to be heterogeneous and inconsistent. Consider the verb *affogare* (to drown), which has [Animate] and [Emotion] as possible STs for the same slot but different senses. T-PAS clusters are small, *clust_1* counts 3 elements and *clust_2* counts 7, in which the two possible types are clearly distinguished. One would expect to find the two senses separated in the SkE clusters as well, since [Emotion] belongs to lower levels of the hierarchy and has a different parent node of [Animate]. However, for SkE we find clusters such as:

(7) *affogare_clust3*: {figlio (son), bimbo (child), pensiero (thought)}

If, on the contrary, we consider closer ST, the clustering will be homogeneous, even if not verb-sense-oriented because too fine-grained.

Medium-sized clusters (more than 10 clustered elements) seem to perform quite well both with hierarchically close and distant STs. A useful example can be *smarrire* (to lose) in T-PAS (8), for which SkE presents the clusters in (9):

(8) *smarrire*

1. [Human] *smarrire* [Artifact]
2. [Human] | [Human group] *smarrire* [Concept] | [Property]

(9) *smarrire_clust1*: {borsello (man bag)}
smarrire_clust2: {significato (meaning), ragione (reason), memoria (memory), pensiero (thought)}
smarrire_clust3: {capacità (capacity), consapevolezza (awareness)}
smarrire_clust4: {nozione (notion), certezza (certainty)}
smarrire_clust5: {senno (sense)}
smarrire_clust6: {fiducia (trust), voglia (will)}
smarrire_clust7: {documento (document)}
smarrire_clust8: {cellulare (mobile phone)}

Considering that for *smarrire* T-PAS creates only two clusters (see (10)), the sets in (9) also highlight the general tendency of SkE to create smaller, semantically highly fine-grained clusters.

(10) *smarrire_clust1*: {borsello (man bag), cellulare (mobile phone), documento (document)}

smarrire_clust2: {*consapevolezza* (awareness), *capacità* (capacity), *certezza* (certainty), *fiducia* (trust), *memoria* (memory), *nozione* (notion), *pensiero* (thought), *ragione* (reason), *senno* (sense), *significato* (meaning), *voglia* (will)}

Large clusters are generally difficult to handle because ideally portionable in more distributionally cohesive groups. As regards the problem of internal noise, due to the *PostV_N* relation, we can note that the phenomenon is pervasive and important, since it affects the results of internal coherence and homogeneity. It is, however, a phenomenon that can be curbed with a revision of the extraction function. What emerges from the analysis is a distance in the general structure of the two clustering results but a good compatibility from the internal semantic point of view. T-PAS privileges rather more complex semantic groupings on a level of co-composition between verb and meaning, linked to conceptual operations of generalization. On the contrary, SkE creates complex and homogeneous structures of relations inside data, even if sometimes this implies clusters that are too fragmented and not always optimal also from a noun-oriented perspective. T-PAS seems to pertain to a higher level of granularity respect to SkE, whose clusters can be considered as possible sub-partitions of the STs.

4 Conclusion

The paper presented the statistical and linguistic results of a comparison between SkE unsupervised clustering model and the manual and verb-sense oriented clustering of T-PAS. It highlighted how the noun-oriented model and the verb-oriented one are not overlapping if not partially. The SkE clustering, even if not overlapping, can still be considered as internally compatible with the T-PAS partition, since the homogeneity metric reaches good results. The internal linguistic analysis allowed to identify the semantic quality through the consistency with a semantic type, the internal homogeneity, the adherence with the verb-oriented approach of T-PAS. The reasons that regulate the fragmentation of clusters in SkE, i.e. motivations that follow a fine-grained logic, were then presented. The analysis made possible to shed a light on the semantic compatibility between the two approaches, which seem to pertain to different levels of granularity.

The difference in the partition output and the parallel semantic compatibility allows us to claim

that the SkE automatic clustering is more useful for the internal investigation of STs than to investigate the verb-Object co-composition relation. It would be interesting to conduct further comparisons between other automatic clustering techniques and that of T-PAS, to investigate additional semantic implications of clustering through noun-based and verb-based approaches.

Reference

- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. In *Language resources and evaluation*, 43(3):209-226.
- Baisa, V., El Maarouf, I., Rychlý, P., & Rambousek, A. (2015). Software and Data for Corpus Pattern Analysis. In *RASLAN*, 75-86.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Cambridge UK. Oxford University Press.
- Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1991). Word sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91)*, 264-270.
- Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. In *Journal of the American statistical association*, 78(383):553-569.
- Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.
- Hanks, P. (1996). Contextual dependency and lexical sets. In *International journal of corpus linguistics*, 1(19):75-98.
- Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, 1:87-98.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. In *Journal of classification*, 2(1):193-218.
- Ježek, E. (2019). Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies. In *JOWO*.
- Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., & Popescu, O. (2014). T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. In *Proceedings of LREC*, 890-895.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Kořář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1:7-36.

- Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Ko-
vář, V., Michelfeit, J., Rychlý, P., Suchomel, V.
(2015). *Statistics used in Sketch Engine*. (document
available at: <https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf>)
- Kilgariff, A., Rychly, P., Smrz, P., and Tugwell, D.
(2004). The Sketch Engine. In *Information Technol-
ogy*, 105:116-126.
- Lavine, B. K., & Mirjankar, N. (2006). Clustering and
classification of analytical data. *Encyclopedia of
Analytical Chemistry: Applications, Theory and In-
strumentation*.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cam-
bridge MA. MIT Press.
- Pustejovsky, J. (2002). Syntagmatic processes. In
Cruse, A. D., Hundsnurscher, F., Job, M., Lutzeier,
P. (eds.) *Lexicology: A Handbook on the Nature and
Structure of Words and Vocabularies*. de Gruyter.
- Romano, S., Vinh, N. X., Bailey, J., & Verspoor, K.
(2016). Adjusting for chance clustering comparison
measures. In *The Journal of Machine Learning Re-
search*, 17(1):4635-4666.
- Rosenberg, A., & Hirschberg, J. (2007). V-measure: A
conditional entropy-based external cluster evalua-
tion measure. In *Proceedings of EMNLP-CoNLL
2007*, 410-420.
- Xu, R., & Wunsch, D. (2008). *Clustering*. Hoboken
NJ. John Wiley & Sons.

GePpeTto Carves Italian into a Language Model

Lorenzo De Mattei^{•♦†}, Michele Cafagna[†], Felice Dell’Orletta^{*}, Malvina Nissim[◊], Marco Guerini[‡]

[•]Department of Computer Science, University of Pisa, Italy

[◊]Center for Language and Cognition Groningen, University of Groningen, The Netherlands

^{*}ItaliaNLP Lab, Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa, Italy

[†]Aptus.AI, Pisa, Italy

[‡]Fondazione Bruno Kessler, Trento, Italy

lorenzo.demattei@di.unipi.it, michele@aptus.ai,

felice.dellorletta@ilc.cnr.it, m.nissim@rug.nl, guerini@fbk.eu

Abstract

In the last few years, pre-trained neural architectures have provided impressive improvements across several NLP tasks. Still, generative language models are available mainly for English. We develop GePpeTto, the first generative language model for Italian, built using the GPT-2 architecture. We provide a thorough analysis of GePpeTto’s quality by means of both an automatic and a human-based evaluation. The automatic assessment consists in (i) calculating perplexity across different genres and (ii) a profiling analysis over GePpeTto’s writing characteristics. We find that GePpeTto’s production is a sort of *bonsai* version of human production, with shorter but yet complex sentences. Human evaluation is performed over a sentence completion task, where GePpeTto’s output is judged as natural more often than not, and much closer to the original human texts than to a simpler language model which we take as baseline.

1 Introduction

Language Models (LMs) based on pre-trained architectures such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) have provided impressive improvements across several NLP tasks. While for BERT-based architectures several monolingual models other than English have been developed, language-specific implementations of generative pre-trained transformer based models, such as GPT-2, are not widely available yet. As a contribution to fill this gap, we developed GePpeTto, the first generative language model for Italian, using the original GPT-2 as a blueprint.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The evaluation of generated text is known to be intrinsically difficult (Gatt and Krahmer, 2018); we adopt here an encompassing approach, performing both automatic and human-based evaluations. The automatic assessment consists in two strategies: the first involves calculating perplexity across different language models trained on various datasets representing different genres. This serves to understand how good GePpeTto is as a language model, and how much it captures the various genres. The second one is a profiling analysis where, by means of a series of linguistic features, we capture some of GePpeTto’s writing characteristics, and compare them to those of the data it was trained on. Finally, the human evaluation is performed over a sentence completion task where GePpeTto is evaluated against gold standard sentences as well as a simple Markov-based baseline.

We make the model available to the community:

<https://github.com/LoreDema/GePpeTto>.

2 GePpeTto

GePpeTto was trained using the original settings of GPT-2 on a collection of Italian texts amounting to almost 13GB. Details on data and model’s parameters are provided in the following sections.

2.1 Data

The training set comprises two main sources. The first one is a dump of Italian Wikipedia (November 2019), consisting of 2.8GB of text. The content was extracted using the Wikiextractor tool (Attardi, 2012). The second one is the ItWac corpus (Baroni et al., 2009), which amounts to 11GB of web texts. This collection provides a mix of standard and less standard Italian, on a rather wide chronological span, with older texts than the Wikipedia

dump (the latter stretches only to the late 2000s).

Minimal processing was applied to the texts. All Wikipedia documents were prefixed by the token “Wikipedia” followed by the page’s title words. All ItWac texts were introduced by the token “Links” followed by the webpage address the text was coming from. For all texts in both collections, end of document was marked with the string `<|endoftext|>`, as done for the original GPT-2 training set (Radford et al., 2019).

2.2 Model

GePpeTto was trained on 4 GPUs Tesla T4, using the PyTorch implementation available on huggingface.¹ The model’s size corresponds to GPT-2 small, with 12 layers and 117M parameters (Radford et al., 2019). The vocabulary size is 30,000, the batch size is 32 and the block size is 100. We use 10,000 warmup steps, the adam optimiser, and an initial learning rate of 5e-5. We train the model for 620,000 steps.

2.3 Examples

Through some selected examples, we illustrate here GePpeTto’s generation behaviour. We prompt it with different inputs, and show how the generated text is indeed style-compliant with the prompts. In the first example, we start the prompt with the ‘keyword’ “Wikipedia”, priming GePpeTto to reproduce text in the style and format of Wikipedia pages (recall that the tag “Wikipedia” was proposed to all Wikipedia training data).

prompt: “Wikipedia La Repubblica del Burmini”

“Wikipedia La Repubblica del Burmini \n\nLa Repubblica del Burmini (in francese “La Repubblica del Burginai”), originariamente repubblica di Francia, venne creata nel 1859 insieme al Regno d’Italia, con capitale Bourges, per promuovere l’unificazione dei due territori della Francia. Le città della Francia si unirono nel 1877 ed i cinque stati facenti parte dell’impero erano Francia, Italia e Savoia.”

In the second example, we prompt GePpeTto with the same text “*La Repubblica del Burmini*” but preposing to it the ‘keyword’ “Links” followed by the web address of an Italian newspaper, priming GePpeTto to

¹<https://huggingface.co>.

reproduce a text in the style of a news article. Interestingly the output is in news style but the prompt is interpreted as the header of the news, i.e., ‘place and date’.

prompt: “Links <http://www.corriere.it> La Repubblica del Burmini”

“Links <http://www.corriere.it> La Repubblica del Burmini - 14 gennaio 2003. Nota: Per gli ambiti regionali, le Regioni dovranno determinare in un solo anno le modalità di trasporto, tra gli altri, degli studenti del secondo ciclo, e le modalità per pagare eventuali supplementi a carico del servizio sanitario nazionale. Si tratta, in altre parole, di un decreto legislativo con un semplice riferimento alla Regione Lombardia.

In the following three examples we can see that even without using any specific keywords, GePpeTto’s is easily style-primed by clear prompts. In the first case, we use as input a string prompting forum-like language, in the second one, we use legal terms in the prompt, and in the last one, we prompt GePpeTto with more sophisticated language, which the model appears to readily pick up in its generated text.

prompt: “Piero DM wrote:”

“Piero DM wrote: >Una domanda sul lavoro a tempo pieno dovrebbe far sorridere.....<< Scusate, ma vorrei avere una vostra opinione.”

prompt: “Visto l’articolo 4 del dpcm:”

“Visto l’articolo 4 del dpcm 29 dicembre 1999, con il quale viene disposto l’inquadramento tra il personale interessato alle operazioni e le RSU per l’applicazione del quarto comma del predetto articolo 4, paragrafo 1, lettera a) del presente accordo,”

prompt: “l’intricato esistenzialismo”

“l’intricato esistenzialismo diventa il motore di quella riflessione metafisico-filosofica di cui il primo pensatore si avvale. L’altruismo di fondo appartiene quasi completamente alla filosofia; l’acuto verso la razionalità senza limite.

3 Automatic Evaluation

GePpeTto is trained as a language model for Italian. To assess its closeness to actual Ital-

ian texts, we calculate perplexity on a variety of sources, including a small leave out test set (1%) of GePpeTto’s training corpus (Section 3.1). In addition, we explore GePpeTto’s linguistic profile by comparing its production with human-written texts along a series of linguistic features (Section 3.2).

3.1 Perplexity

As a first evaluation, we are interested in understanding the quality of GePpeTto as a language model in its own training domain. As a second evaluation we want test its performance at zero-shot domain transfer (i.e. language modeling of a different domain). We use perplexity as a measure of language modelling performance. The different domains we consider, and the relative corpora we use, are as follows:

- own domains: Wikipedia and ItWac;
- legal domain: a corpus of Italian laws scraped from EUR-Lex² (tables excluded);
- news: a corpus of articles from the online versions of two newspapers, i.e., *la Repubblica*³ and *Il Giornale*⁴ (De Mattei et al., 2020);
- social media: a corpus of forum comments (Maslennikova et al., 2019).

To compute the perplexity scores (Table 1) we used a random sample of 4M tokens for each corpus. As expected, GePpeTto performs better on its own domains. Although ItWac is four times bigger than Wikipedia, the lower performance on the former might be due to ItWac being open domain with a large diversity of styles, while Wikipedia is more ‘standardised’. Consistently with this hypothesis, we observe a similar trend in ‘out-of-domain’ testing, where GePpeTto performs better on domains with a well coded style, namely legal documents. On domains with less coded styles, such as news and especially forum comments, we observe a performance drop.

If we compare perplexity scores with the original English GPT-2 small model, we see that GePpeTto’s results are slightly worse on the own domain corpora, which could be due to the smaller size of the training set. Out-of-domain perplexity scores are comparable between the two models.

3.2 Linguistic Profiling

For our second evaluation, we used Profiling-UD (Brunato et al., 2020), a tool for the automatic

²<https://eur-lex.europa.eu/>

³<https://www.repubblica.it>

⁴<https://www.ilgiornale.it/>

DOMAIN	PERPLEXITY
Wikipedia	26.4910
ItWac	30.9698
Legal	39.6087
News	48.3468
Social Media	131.3812

Table 1: Perplexity of GePpeTto over several in-domain and out-of-domain corpora.

Feature	Original		GePpeTto	
	μ	std	μ	std
CPT	4.809	0.959	4.750	1.127
TPS	32.302	28.322	20.382	11.127
TPC	12.393	11.504	10.711	8.529
LL _{max}	13.290	13.370	8.922	6.112
LL _{avg}	2.555	1.002	2.373	0.676

Table 2: Main linguistic features considered in our analysis. CPT = chars per token, TPS = token per sentence, TPC = tokens per clause, LL = links length.

analysis of texts that extracts several linguistic features of varying complexity. These features range from raw text properties, such as average length of words and sentences, to lexical, morpho-syntactic, and syntactic properties, such as part-of-speech (POS) distribution and inflectional properties of verbs. More complex aspects of sentence structure are derived from syntactic annotation, and model global and local properties of parsed tree structure, such as the order of subjects/objects with respect to the verb, the distribution of syntactic relations, and the use of subordination.

In our analysis we focus on two macro aspects of GePpeTto’s output, namely lexical complexity and syntactic complexity, and compare them to human productions. To do so, we rely on a selection of Profiling-UD’s features which we use as proxies for the macro-aspects that we consider.

We run the profiling analysis on a sample of both gold and generated texts. For gold, we randomly sample the test set for a total of about 19k sentences. For GePpeTto, we pick the first token from each of the 19k gold sentences, and use it as a prompt to the model. We profile these generated texts.

Lexical complexity. We proxy lexical complexity with the number of characters per word, overall frequency of tokens, also with reference to an ex-

ternal dictionary, and POS distribution.

The number of characters per token (CPT), which indicates whether shorter (usually more common) or longer (usually more complex/specialised) words are used, is completely comparable across the original (4.80, std=0.96) and GePpeTto’s (4.75, std=1.13) language models – see Table 2. This suggests that the complexity of the used vocabulary is not that different.

We compute a reference dictionary of token frequency on ItWac (≈ 1.5 billion tokens), and compare observed token frequency in both gold and generated text to this reference. We observe that in gold sentences, each token has a probability of 0.912 to be in the top 5% of most frequent tokens. In the generated sentences, the probability grows to 0.935, suggesting that GePpeTto is more likely to use more frequent words rather than rarer ones. This observation is in line with previous research which showed that for Nucleus Sampled texts, such as those produced by GPT-2, all tokens come from the top-p%, since the long tail is cut off, while for human produced texts, the probability of all tokens being drawn from the top-p% of the language distribution goes to zero as document length increases (Gehrmann et al., 2019; Zellers et al., 2019).

Regarding POS distribution, we observe that while for most POS tags usage is comparable, for a few others the two language models differ. The latter are, specifically, auxiliaries and proper nouns, which GePpeTto tends to overgenerate in comparison to the original model, and adjectives, which GePpeTto instead uses less than in the original texts. This is seen also for nouns and verbs, but the differences are relatively minimal. Conjunctions are also overall less frequent in GePpeTto. A detailed table will be included in the final version.

Syntactic complexity. At the level of syntax, we proxy complexity by the number of tokens per sentence, and the number of tokens per clause. We also look at the length of a dependency link, that is calculated as the number of words occurring linearly between the syntactic head and its dependent (excluding punctuation dependencies). The value associated with this feature corresponds to the average value extracted for all dependencies in a text. This information is complemented with the feature *Maximum dependency link* corresponding to the longest dependency link for each sentence.

When comparing the number of tokens per sentence (TPS, Table 2), we see that it’s much lower

for GePpeTto’s production rather than for human texts (20.4 tokens per sentence on average for GePpeTto vs 32.3 for gold texts), indicating that GePpeTto generates shorter sentences. Contextually, we also observe that GePpeTto’s generated sentences exhibit less variation in length (smaller STD) than human sentences (larger STD).

The difference in number of tokens at the clause level is relatively smaller, with clauses of length 12.4 in human texts vs 10.7 in GePpeTto (TPC, see Table 2). Considering that a clause is proxied by the presence of a verbal/copular head, it seems that sentences produced by GePpeTto, though shorter, are similar in complexity given the proportional distribution of verbal heads.

The above values taken together might suggest that while complexity at the macro level (sentence length) is higher for natural sentences, at the micro level (clause length) complexity of GePpeTto’s generations and human texts is more similar. While this intuition will require further linguistic analysis, observing the length of syntactic links seems to support it. This feature proxies quite well syntactic complexity, since it indicates how maximally far (and how far on average) a dependent and its head are within a sentence. Both the maximum length and the average length are higher for human texts (LL_{max} and LL_{avg} , see Table 2). However, if we look at them proportionally to sentence length, we find that they are comparable: normalising the longest link by the number of tokens per sentence (LL_{max}/TPS), we obtain similar values for gold (0.411) and for GePpeTto (0.438). This suggests that GePpeTto produces somewhat shorter sentences, but their internal complexity relatively corresponds to the internal complexity of the longer sentences produced by humans.

4 Human evaluation

We also test GePpeTto’s ability to generate Italian texts through a sentence completion task. The automatically generated sentences are presented to human subjects for evaluation on perceived naturalness and compared to gold ones and to a baseline.

While the original (gold) texts represent an upperbound for GePpeTto, we do not actually have a lowerbound against which the quality of GePpeTto can be assessed. To provide a comparison, we train a simple Markov model that would be able to generate text and use it as our baseline. Since the size of a Markov model dra-

matically grows with its vocabulary size, we use 1 million randomly sampled sentences from the same training-set used for GePpeTto. We train a Markov chain generator using the `markovify`⁵ implementation with state size 2, then we generate synthetic texts starting from the last 2 tokens of same prompts used for GePpeTto.

4.1 Tasks

Human subjects are asked to perform two evaluation tasks. One is a comparative ranking task, where subjects are asked to rank three portions of text (produced by gold, GePpeTto, baseline) according to perceived naturalness. The other is a classification task, where subjects are asked to tell, according to their intuition, if a portion of text, seen in isolation, is automatically generated (*yes*, *no*, *can't tell*).

Experimental design. The experiment includes 12 conditions of the stimulus material in a 4x3 design. One level (A) with three conditions is given by {gold, GePpeTto, baseline}. The second level (B) is the prompt+completion combination that results in 4 conditions {5+5, 5+10, 10+5, 10+10}. We use 100 different prompts (randomly selected gold sentences truncated at 5 and 10 tokens). Each of the 100 prompts enters each of the 12 conditions of the 4x3 design, for a total of 12 different stimuli. Basically, each 5 or 10 tokens prompt is completed with 5 or 10 tokens coming either from gold, GePpeTto, or the baseline model. Table 3 shows an example of all the stimuli deriving from the same 5- or 10-token prompt.

Each subject is assigned either to the ranking or to the classification task.

In ranking, we opt for a between subject evaluation set up by assigning each subject to one of the (B) conditions and offer the three versions of (A) to be ranked. For example, one subject is asked to evaluate all the 100 prompts in the 5+5 configuration (dimension B) for the three realisations, i.e., gold, GePpeTto, and baseline (dimension A).

For the classification experiments, we again opt for a between subject evaluation set up, this time by assigning each subject to one of the 12 conditions, randomly picked up for each prompt. In other words, we make sure that each subject is exposed to only one completion per prompt, randomising prompt order. By seeing only one (out of 12) realisation per prompt, each subject sees a

given prompt only once and we can therefore avoid cross-comparison effects of different completions of the same prompt, which could otherwise potentially lead again to an implicit ranking task.

Material. The materials are prepared as follows: we have selected 100 random documents/sentences and have cut them at their 5 first tokens and also their 10 first tokens. Each 5-token and 10-token prompt was given to GePpeTto and baseline so that the models could continue the text.

For each prompt, we obtain one single generated text by the two automatic models and chop them at 5 or at 10 tokens. In other words, each chopped version is derived from the same generated output which is just cut at different lengths.

We cut the sentences (including the original one) to control for the effect of text length. Indeed, we observed in Section 3.2 that GePpeTto generates shorter sentences than humans, which could represent a strong bias in evaluation. In Table 3, we show examples of all the possible stimulus material configurations according to the prompt+completion conditions of level (B).

Instructions and subjects. For both the ranking and classification experiments, subjects were told that they will have to evaluate excerpts of text along a ‘more natural vs. more artificial’ dimension. All stimuli used in both scenarios are the same.

For the ranking scenario, subjects were asked to “*rank the given examples from the most natural to the most artificial*”, where the inputs are three texts (gold, GePpeTto, baseline), all starting with the same prompt, thus the same five or ten tokens.

For the classification scenario, subjects saw instead the portions of text in isolation, and could answer *yes*, *no*, or *can't tell* to the question “*according to your intuition is this sentence written by an artificial intelligence?*”.

A total of 24 unique subjects (12 females) carried out the tasks using Google Forms. Twelve subjects (6 females) were assigned to Task 1 and the others to Task 2. Each subject evaluated 100 cases, and each case was evaluated by three different subjects.

4.2 Results

First, we discuss the results of our human evaluation separately, with observations related to the ranking task and observations related to the classification task. Subsequently, we knit together the two outcomes to draw a wider picture of how humans assess the quality of GePpeTto's output.

⁵<https://github.com/jsvine/markovify>.

5 token prompt: Mentre per quanto riguarda gli	
10 token prompt: Mentre per quanto riguarda gli accordi per la fornitura di	
Gold	
5+5	Mentre per quanto riguarda gli accordi per la fornitura di
5+10	Mentre per quanto riguarda gli accordi per la fornitura di latte, in scadenza questa
10+5	Mentre per quanto riguarda gli accordi per la fornitura di latte, in scadenza questa
10+10	Mentre per quanto riguarda gli accordi per la fornitura di latte, in scadenza questa settimana, Alemanno ha detto
GePpeTto	
5+5	Mentre per quanto riguarda gli emendamenti, fa presente che il
5+10	Mentre per quanto riguarda gli emendamenti, fa presente che il suo gruppo non ha sottoscritto
10+5	Mentre per quanto riguarda gli accordi per la fornitura di beni e servizi, i fatti
10+10	Mentre per quanto riguarda gli accordi per la fornitura di beni e servizi, i fatti in suo possesso hanno come
Markov-based baseline	
5+5	Mentre per quanto riguarda gli aspetti più significativi del mondo
5+10	Mentre per quanto riguarda gli aspetti più significativi del mondo editoriali, con priorità di sviluppo
10+5	Mentre per quanto riguarda gli accordi per la fornitura di biciclette elettriche a 48 bit
10+10	Mentre per quanto riguarda gli accordi per la fornitura di biciclette elettriche a 48 bit (281,5 trilioni di operazioni e

Table 3: Example outputs (stimuli) for different prompt lengths of the same original sentence.

Ranking Overall, results show that the most frequently chosen completion is the gold one, followed by GePpeTto and then the Markov baseline, but the baseline is far more distant from GePpeTto than GePpeTto from gold (Figure 1). If we look at results in more detail (see Table 4), based on the variable that we have considered in the experimental set up, namely length of input and continuation as well as overall sentence length, we observe that the order of preference for gold is 10+10, then 5+10, then 10+5, and lastly 5+5, while for the automatic models the order is 5+5, 10+5, 5+10, and then 10+10, suggesting the following.

First, the shortest the sentence, the hardest it is to discriminate between gold and generated text; indeed, the 5+5 condition is the one that results best for the two models and worst for gold.

Second, when the sentence is the longest (10+10), it is easiest for the subjects to discriminate the gold from the generated sentences. It is also interesting to note that in this condition we observe the largest gap between the two generation models, with GePpeTto getting ranked higher than Markov more than in the other conditions.

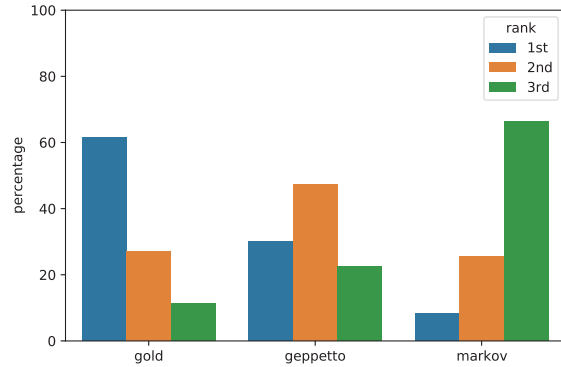


Figure 1: Ranking results for the three models

Third, at equal sentence length (15 tokens) the situation is a bit more fuzzy, but we can observe a slight tendency where it is easier to spot as automatically generated the 5+10 rather than 10+5 cases. This, in combination with the previous observation, seems to imply that the longer the generated text, the easier it is to figure out which texts are automatically produced, which makes sense, since there is more ‘space’ for the models to make mistakes.

model	5+5			5+10			10+5			10+10		
	1 st	2 nd	3 rd	1 st	2 nd	3 rd	1 st	2 nd	3 rd	1 st	2 nd	3 rd
Gold	54	30	16	62	31	7	60	27	13	70	21	9
GePpeTto	34	43	23	30	46	24	33	43	24	23	59	18
Markov	12	27	61	8	23	69	7	30	63	7	20	73

Table 4: Percentages of ranking results according to the various stimulus material conditions.

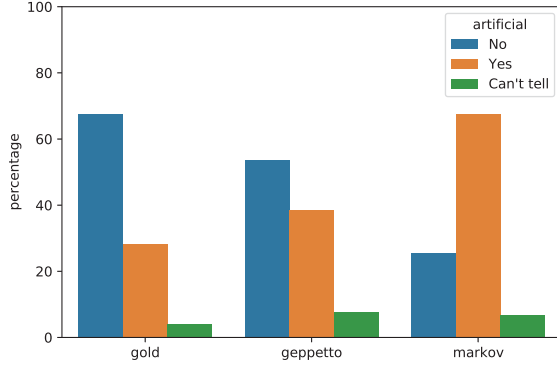


Figure 2: Classification results for the three models

Classification Overall, results show that across all conditions, gold sentences are most often rightly identified as not automatically generated (68% of “no” to the question whether the output was produced by an artificial intelligence), followed by GePpeTto (54%), and lastly by the Markov baseline (26%), indicating, as expected, that the latter produces the least natural outputs. Figure 2 reports the distribution over the various answers. Also in this case the distance between GePpeTto and gold is lower than GePpeTto and the baseline (double in percentage points), indicating that the production of GePpeTto is approaching natural language. It is also interesting to see that the highest percentage of “can’t tell” is recorded for GePpeTto, meaning that for this model it was harder than for baseline and gold to decide whether the text was automatic or not.

Let us look at results in more detail (Table 5), focusing again on length of input and continuation. Regarding continuation, we observe that *+5 conditions are better than *+10 conditions for both automatic models, indicating that the least generated text, the more natural the fragment is perceived.

Regarding input length, we see that for GePpeTto a longer prompt yields better results (10+5 is better than 5+5, and 10+10 is better than

5+10). With 10-token prompts, GePpeTto generates text that is (i) assessed as natural as much as the original text when completed with 5 tokens (62% GePpeTto, 63% original), and (ii) judged as natural 50% of the times when completed with 10 tokens. This seems to suggest that a longer input context is beneficial to GePpeTto when completion size is kept constant. However, we may wonder whether GePpeTto is evaluated as more natural because the generated text is actually better given the more context to start with, or simply because there is more gold text in the stimulus. If it were just for the contribution of a longer gold portion in the stimulus, we should see a similar behaviour for the baseline. Instead, we see that prompt size doesn’t matter for the baseline, at least for the 5 token completion case (33% in both 5+5 and 10+5). In the 10-completions (5+10 and 10+10), the larger amount of gold data in the stimulus probably does alleviate a little the very low naturalness induced by the generated text. While we can tentatively postulate that GePpeTto generates better text when more input is provided, further investigation is required to provide more solid evidence.

Summary of Results. Intersecting the observations from the two experimental setups provides us with a complete picture. In ranking (thus when the models are directly compared), both GePpeTto and the baseline perform best in the 5+5 and 10+5 conditions, suggesting that automatic generation can easily be spotted when compared side by side with human text. In other words, the least generated material, the better.

However, looking at classification, where each textual material is evaluated in isolation, we see that the two models behave in fact very differently. First, there is a much larger proportion of cases produced by GePpeTto that are deemed “natural” (54%) compared to Markov (26%). Second, the margin of uncertainty when judging GePpeTto is higher than for the baseline and

model	5+5			5+10			10+5			10+10		
	yes	no	ct	yes	no	ct	yes	no	ct	yes	no	ct
Gold	26	66	8	27	68	5	32	63	5	28	71	1
GePpeTto	32	55	13	48	46	6	32	62	6	42	50	8
Markov	62	33	5	80	13	7	61	33	6	71	19	10

Table 5: Percentages of classification results according to the various stimulus material conditions. Is the text automatically generated? {yes, no, can't tell (ct)}.

for original text. Lastly, given the same completion size, GePpeTto performs better when its prompt is longer. Whether this is an effect of a larger proportion of gold data in the stimulus or it has to do with providing the model with a larger input context is left to future investigation.

5 Conclusion

GePpeTto is the first GPT-2-based language model for Italian. Through both automatic and manual evaluation we assessed its quality on a variety of texts and in comparison to gold data as well as another statistical generation model. Results show that GePpeTto is able to produce text which is much closer to human quality rather than to the text generated by the other generation model we have used. Linguistic analysis also highlights that GePpeTto's production is quite similar to human production, though in a sort of bonsai version, since its sentences are on average shorter than the original texts, but with similar complexity.

The availability of GePpeTto opens up substantial possibilities. In the same way that GPT-2 is changing the approach to several NLP English tasks, we can expect GePpeTto to serve a similar purpose in Italian language processing.

References

- Giuseppe Attardi. 2012. Wikiextractor. <http://attardi.github.io/wikiextractor>.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).
- Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020. Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.
- Aleksandra Maslennikova, Paolo Labruna, Andrea Cimino, and Felice Dell'Orletta. 2019. Quanti anni hai? Age Identification for Italian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy. CEUR Proceedings 2481.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Phonological Layers of Meaning: A Computational Exploration of Sound Iconicity

Andrea Gregor de Varda
Centre for Mind/Brain Sciences
University of Trento
andregregor.devarda@
studenti.unitn.it

Carlo Strapparava
Fondazione Bruno Kessler (FBK)
strappa@fbk.eu

Abstract

The present paper aims to investigate the nature and the extent of cross-linguistic phonosemantic correspondences within a computational framework. An LSTM-based Recurrent Neural Network is trained to associate the phonetic representation of a word, encoded as a sequence of feature vectors, to its corresponding semantic representation in a multilingual vector space. The processing network is tested, without further training, in a language that does not appear in the training set. The performance of the multilingual model is compared with a monolingual upper bound and a randomized baseline. After the quantitative evaluation of its performance, a qualitative analysis is carried out on the network's most effective predictions, showing an inhomogeneous distribution of phonosemantic information in the lexicon, influenced by semantic, syntactic, and pragmatic factors.

1 Introduction

The idea of a consistent relationship between sound and meaning has held a particular fascination over philosophers and linguists (Plato, 1998). However, in recent times, this charming hypothesis has progressively lost the interest of scholars, especially in the post-Saussurean linguistic tradition, which emphasized the arbitrariness in such relation. The idea that sounds have inherent meanings has recaptured its original attractiveness in the field of cognitive sciences, where the attention has initially focused on the link between sound and shape. A prominent example of these

naturally biased mappings came from Köhler's (1929) finding that, when asked to match two novel shapes with the non-words 'maluma' and 'takete', English-speaking adults tended to label as 'maluma' the curled shape, and as 'takete' the sharp one. This germinal study paved the way to several replications and expansions of its findings, that reproduced Köhler's results in different geo-cultural contexts (Bremner et al., 2013) and at different developmental stages (Maurer et al., 2006). Since then, different studies have tackled the topic of iconicity in language from a broader perspective, showing that adults can associate visually presented characters (Koriat and Levy, 1977) and auditorily presented words (Berlin, 1995) of a foreign language to their meaning, with an accuracy above chance.

Recently, linguistic iconicity has gone from being a marginal – although appealing – matter to being integrated into broader theories of language evolution and acquisition. Indeed, rejecting the assumption of an arbitrary mapping between sound and meaning sensibly reduces the problem space of language emergence, establishing constraints on the consensus of word choice. Furthermore, an iconic relation between a sound and its referent might help with memory consolidation in the process of language acquisition (Sathian and Ramachandran, 2019). Ramachandran and Hubbard (2001) speculate that phenomena as the one reported by Köhler might arise from neural connections among adjacent cortical areas, where the visual features of the referent, the appearance of the speaker's lips and the kinaesthetic features of the articulation are combined. According to their view, such neural connections would have influenced both the phylogenetic evolution and the ontogenetic development of language. Although the previous findings are consistent with this hypothesis, an alternative explanation must be taken into account: the roots of these correspon-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dences could be grounded in the knowledge of language, that allows children and adults to generalize the regularities in sound-to-meaning mappings from their native language to nonsense and foreign words. Under this rationale, phonosemantic relations would be implicitly learned from general recurrences in already known languages. A crucial aspect of this account lies in the fact that it does not posit any preexisting disposition wired in the human brain, moving the *locus* of linguistic iconicity from the mind to language itself. A natural question that arises from this perspective is whether linguistic information alone is sufficient to give rise to the phonaesthetic biases presented in the literature. A computational exploration of the phenomenon under scrutiny is a feasible way to approach the subject. The idea that phones have inherent meanings is relatively understudied within the computational framework, and most of the studies addressing the topic have either focused on a single language (Gutiérrez et al., 2016; Sagi and Otis, 2008; Abramova et al., 2013; Monaghan et al., 2014; Tamariz, 2008) or on a small set of concepts on a massively multilingual scale (Blasi et al., 2016; Wichmann et al., 2010). Surprisingly, no study to our knowledge has tackled the topic through a deep learning methodology, and no cross-linguistic investigation has been performed on a lexicon-wide level. The purpose of the present study is two-fold: first, we wish to explore the idea of a cross-linguistic correspondence between the phonetic and the semantic representation of a word on the whole lexicon, without any theory-driven restriction guiding our choice of the lexical items. Then, we aim to examine whether the meaning that is rooted in the sound that words are made of is homogeneously distributed in the lexicon. Ultimately, these two goals converge toward the research question hinted above, namely, whether linguistic information alone could suffice for the extrapolation of the phonosemantic biases reported in the present section. A possible way to answer this question is to assess the ability of a *tabula rasa* neural network to extend the regularities captured in a set of given languages to a previously unseen one. Although equipped with clear structural priors, neural networks do not conceal biases that resemble those assumed to model the aforementioned phonosemantic correspondences. If a processing network showed the ability to induce cross-linguistic regularities in

sound-to-meaning mappings, this would suggest that linguistic data contain a sufficient amount of information to encode for phonosymbolic biases.

The present study aims to explore the possibility of a certain degree of cross-linguistic correspondence between sound and meaning that is already encoded in language. A Long Short-Term Memory network (LSTM) is trained on four languages to associate the sequence of sounds that compose a word, encoded as phonetic vectors, to its corresponding semantic representation in a multilingual vector space. Then the processing network is tested, without further training, on a language that does not appear in the set of languages on which the training has been performed. The performance of the multilingual model is compared with the results of (a) a monolingual model, trained and tested on different subsets of a single language’s vocabulary, and (b) a baseline model, where the output vectors in the training are randomly shuffled. After the quantitative evaluation of its performance, a qualitative analysis is carried out on the network’s most effective predictions.

2 Methods

In the present study, an LSTM-based Recurrent Neural Network is trained to associate the phonetic to the corresponding semantic representation of a word. The semantic representations consist in 300-dimensional word embeddings in a multilingual vector space, whereas their corresponding phonetic features are expressed as sequences of phonetic vectors in 22 dimensions. The experimental pipeline is summarized in the flowchart in Figure 1.

2.1 Semantic vectors

The semantic representations included in the model, provided by Facebook Research, consist in multilingual word embeddings generated with `fastText` from Wikipedia data (Bojanowski et al., 2017) and aligned in a common vector space through a fully unsupervised methodology (Conneau et al., 2017)¹. The present study is conducted on Italian, German, French, Vietnamese, and Turkish embeddings.

¹Publicly available at <https://github.com/facebookresearch/MUSE>

2.2 Phonetic vectors

For each word in the embedding dataset, we obtained its phonemic transcription with `Epitrans`, a Python library for transliterating orthographic text in the International Phonetic Alphabet (IPA) format. Then, we converted the IPA string into a sequence of feature vectors in 22 dimensions with `PanPhon`, a package that traduces IPA segments into subsegmental articulatory features (Mortensen et al., 2016). It has been shown that phonologically aware models built on the linguistically motivated and information-rich representations yielded by the `Epitrans-PanPhon` pipeline outperform the raw hot-encoding of character-based models in different tasks (Mortensen et al., 2016; Bharadwaj et al., 2016).

2.3 Neural architecture

An LSTM-based Recurrent Neural Network is trained to map the sequences of phonetic feature vectors in input into semantic vectors in output. The model is built with `Keras`, a deep learning framework for Python (Chollet et al., 2015); it includes a single LSTM layer with 172 units, a dropout of 0.2 and a recurrent dropout of 0.2. Cosine similarity is used as both objective function and metric, and the Adam optimization method is employed for training (Kingma and Ba, 2014). We adopted the *tanh* activation function for the output layer since its codomain corresponds to the range $(-1, 1)$, in which the semantic vectors are defined. The hyperparameters are set without tuning.

2.4 Experimental conditions

The experimental conditions are characterized by different combinations of training and testing sets. In the multilingual condition, the model is trained for one epoch on the Italian, German, French, and Vietnamese datasets, and then tested in Turkish. Our unique concern in the language selection was that none of the languages in the training set was typologically close with the language presented in the test set. Turkish has been chosen for the test set since it is not considered to be related to any of the languages presented in the training set, at least within a reasonable time window. Indeed, Turkish is a Turkic language, whereas Italian, German and French are Indoeuropean, and Vietnamese belongs to the Austroasiatic language family. To establish a baseline for the evaluation of the model’s performance, we trained a model randomly shuffling

the output vectors. We will refer to this manipulation as the random condition. In the monolingual condition, which defines the upper bound of the network’s performance, the LSTM is trained and tested on different subsets of the Italian dataset, with a train-test split ratio of 0.2. In order to compensate for the different dimensions of the training set (roughly one fifth of the multilingual sample), the monolingual model is trained for five epochs.

3 Results

Table 1 lists the test results for each of the models described in Section 2.4. The number of lexical items included in the training and in the test set are reported in the $\text{Dim}_{\text{train}}$ and the Dim_{test} columns, respectively. The last column of the table presents the average cosine similarity between the target semantic vector and the model’s prediction for every word in the test set. As reported below, the multilingual model outperforms the random baseline, with a 0.0351 points higher average cosine similarity. As expected, the monolingual performance is stronger than the one achieved by the multilingual model, with a difference of 0.0453 in the metric. The relatively modest magnitude of the difference between the monolingual and the multilingual results should be attributed to the limited size of the training set in the former condition: increasing the number of epochs might have partially compensated for the shortage in the training data, but additional forward and back propagation on the same data might not be as effective as further training on unseen data, especially in terms of generalization. The general pattern of results, with the multilingual performance almost halfway between the monolingual and the random results, is in line with our predictions. The difference between the multilingual and the random condition is consistent with the hypothesis that a certain degree of cross-linguistic correspondence between phonetic and semantic representations is already encoded in language; moreover, it shows that, with sufficient training, this correspondence can be efficiently captured by an LSTM network.

4 Qualitative analysis

As previously mentioned, the LSTM network trained on multilingual data showed the ability to induce cross-linguistic regularities in sound-to-meaning mappings, suggesting that linguistic data

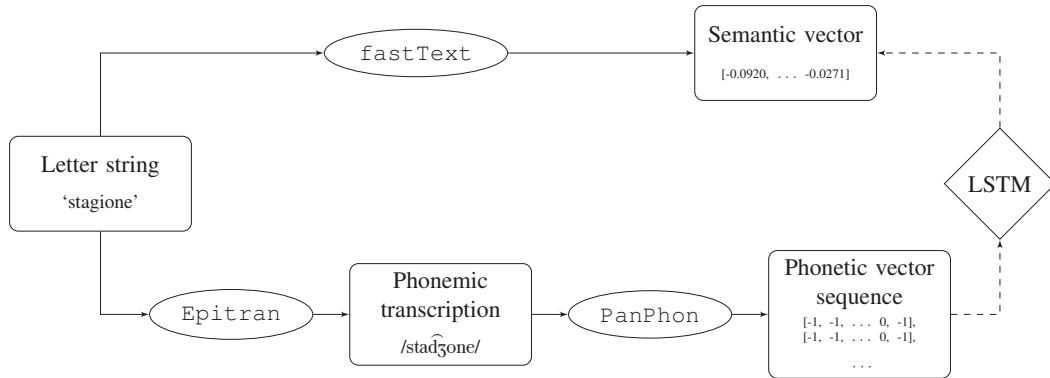


Figure 1: Schematic representation of the experimental pipeline

Model	Dim _{train}	Dim _{test}	Cosine similarity
Multilingual	794,870	199,108	0.4467
Random	794,870	199,108	0.4116
Monolingual	159,621	39,906	0.4920

Table 1: Test results by experimental condition

alone contain the sufficient amount of information to encode for phonosymbolic biases.

In the light of the results presented above, a natural question that arises is whether phonosemantic information is uniformly distributed in the lexicon, or some semantic areas tend to incorporate stronger correspondences with their phonetic counterparts. We hypothesized that some areas of the semantic space might show a more consistent mapping with their phonological realization, but without any clear *a priori* expectation on the regions that could reveal higher phonosemantic transparency, we addressed this problem through a data-driven qualitative analysis. We extracted from the test results of the multilingual condition 39,821 items (20% of the total), selecting the words that the network had predicted with the higher precision – that is, the words whose vector prediction had the higher cosine similarity with respect to the target. Then, we restricted the analysis by excluding the items with low frequency. We conjectured that it would be unlikely for rare and unfamiliar terms to convey phonosemantic relations without being etymologically related to other languages. For instance, across different disciplines, the technical jargon – whose instances are typically infrequent in corpora – is commonly derived from Greek and Latin roots. We employed the Twitter-based Turkish frequency

estimates from the Worldlex dataset², that has been shown to outperform traditional frequency estimates in predicting lexical decision reaction times, thus exhibiting a higher cognitive validity (Gimenes and New, 2016). From the previously extracted items, we excluded those that were not in the list of the 20,000 most frequent words (that is, the 1.21% of the words with higher frequency). The resulting items were translated into English with Googletrans, a Python library that implements Google Translate API. The results of the analysis are reported in Table 2, where the items that satisfy the aforementioned constraints (from now on, the *quality subset*) are grouped into four intuitive categories according to their meaning and their grammatical function.

The most represented categories of words in this subset of efficiently predicted items are proper names and lexical borrowings, with the former generally associated with a higher cosine similarity between target and prediction. They are not reported in Table 2, since their detailed analysis is not relevant for the purposes of the study. However, the predominance of proper names over lexical borrowings is compatible with one of the basic postulates of model-theoretic semantics. It is generally assumed that proper names, unlike definite descriptions and generalized quantifiers, directly refer to entities in the world (Delfitto and Zamparelli, 2009); hence, they are expected to hold their exact meaning across languages.

The cross-linguistic consistencies in proper names and lexical borrowings are clearly due to contact between languages. Other word categories strongly associated with the phonosemantic fea-

²Publicly available at <http://worldlex.lexique.org>

Internal states	istediğimde ('I want'), düşüncelerimi ('my thoughts'), isteyenlere ('those who want'), düşünsenize ('imagine'), düşüncem (‘I thought’), açıkçası ('frankly'), aşkınsın ('you are in love'), kendimde (‘in myself’)
Function words	vee ('and'), kendileri ('themselves'), onların ('they'), gerektiğinde (‘when’), mıydın ('did you')
Interjections	hee ('ooh'), boku ('shit'), himm (‘uhm’)
Other	yaklaşım ('approach'), poğaça (‘pastry’), demis ('said'), gerçekmiş (‘real’), tabii ki ('of course'), uygulama- lamaları ('applications'), gani (‘abundant’)

Table 2: Intuitive clustering of the model’s best predictions

tures detected by the network are undoubtedly more relevant in revealing lexical clusters with privileged sound-to-meaning mappings. For instance, a conspicuous portion of items in the quality subset is semantically linked to different internal states, with a predominance of concepts related to mental processes. The quality subset comprises also various function words (conjunctions, pronouns, and one auxiliary verb). This result is particularly informative since function words, being a closed-class category, are not as numerous as content words; therefore, their number of instances in the training set was most likely limited. An additional cluster in the quality set comprises three interjections, including one imprecation. Interjections express spontaneous feelings or reactions (Bloomfield, 1984) and can be closely related to their natural manifestation (Wharton, 2003); hence, it is not surprising to find a more transparent link between their phonoarticulatory expression and their meaning. Moreover, this result is consistent with the findings of Dingemanse et al. (2013), that show that the interjection “Huh?” is a universal, found in roughly the same form and function in spoken languages across the globe.

The present findings suggest that phonosemantic information is not uniformly distributed in the lexicon: the consistency of the mapping between sound and meaning seems to be influenced by semantic, syntactic, and pragmatic factors. Indeed, the semantic neighbourhood linked to inter-

nal states shows a privileged relationship between sound and meaning, whereas on the syntactic side function words seem to be favoured, if their absolute prevalence in the lexicon is taken into account. Moreover, interjections, which are characterized by a strong pragmatic valence, stand among the items predicted with the highest precision by the model.

5 Limitations and further directions

From a methodological standpoint, the reliability of the present results could benefit from the exclusion of lexical borrowings and proper names from the training and the test sets. Excluding etymologically related terms could further improve the reliability of the results, but at the costs of raising the difficulty of assessing the words’ relatedness in different languages, with the subsequent need of a proper metric.

Another confound that we wish to address in future research is the role played by morphological factors in aiding the cross-linguistic feature extraction performed by the network. FastText vectors exploit information related to subword character strings, and might therefore encode regularities pertaining to recurrent morphemes in the non-isolating languages in our dataset (Italian, German, French, and Turkish). We acknowledge that the network might have captured the recurrences encoded in the semantic vectors comprising the training set and their relationships with the corresponding phonetic feature vectors; indeed, we believe that this regularities might have played a relevant role in the monolingual condition, where the model might have learnt that morphologically related words (i.e. in this context, words that are similar at the character- and phoneme-level) tend to be associated with close subregions of the semantic space. Nonetheless, we do not see how this information could have altered significantly the performance in the multilingual condition. That said, we leave for future research an assessment of the algorithm’s performance on semantic vectors which lack access to subword-related information, such as word2vec (Mikolov et al., 2013), and in languages with opaque orthography (e.g. English and French) and non-concatenative morphology (e.g. Chinese)³.

³We gratefully thank an anonymous reviewer for drawing our attention to this matter and suggesting the mentioned options to address this confound.

As for all the studies that employ artificial neural networks to draw conclusions on human cognition, it is mandatory to clarify some limitations on the extent of the inferences that can legitimately follow the presented results. The finding that a neural network can succeed in a task without the structural priors postulated in the human mind does not necessarily imply that these priors are not actually encoded in the brain: the assumption of a functional equivalence between artificial and biological processes needs to be independently motivated. Moreover, it should be noticed that the participants of the behavioural studies presented in Section 1 were not necessarily polyglots, whereas the promising cross-linguistic performances described in the results have been obtained with a multilingual model. In addition to these intrinsic methodological limitations, an account that does not assume any prior specification for the linguistically encoded phonosymbolic mappings would leave an open question concerning their origin. Hence, the present study does not claim to reject the multi-sensory integration hypothesis presented in the Introduction. Its purpose is simply to show that, in principle, linguistic information alone could suffice for a generalization in sound-to-meaning mappings.

References

- Ekaterina Abramova, Raquel Fernández, and Federico Sangati. 2013. Automatic labeling of phonesthetic senses.
- Brent Berlin. 1995. *Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature*, page 76–93. Cambridge University Press.
- Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.
- Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.
- Leonard Bloomfield. 1984. *Language*. University of Chicago Press. Google-Books-ID: 87BCD-VsmFE4C.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.
- Andrew J. Bremner, Serge Caparos, Jules Davidoff, Jan de Fockert, Karina J. Linnell, and Charles Spence. 2013. “Bouba” and “Kiki” in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, 126(2):165 – 172.
- François Chollet et al. 2015. Keras. <https://keras.io>.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data.
- Denis Delfitto and Roberto Zamparelli. 2009. *Le strutture del significato*. Itinerari Linguistica. Mulino, Bologna. OCLC: 695640183.
- Mark Dingemanse, Francisco Torreira, and N. J. Enfield. 2013. Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE*, 8(11).
- Manuel Gimenes and Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48(3):963–972.
- E. Dario Gutiérrez, Roger Levy, and Benjamin Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2379–2388, Berlin, Germany. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.
- Wolfgang Köhler. 1929. *Gestalt Psychology*. Liveright.
- Asher Koriat and Ilia Levy. 1977. The symbolic implications of vowels and of their orthographic representations in two natural languages. *Journal of Psycholinguistic Research*, 6(2):93–103.
- Daphne Maurer, Thanujeni Pathman, and Catherine J. Mondloch. 2006. The shape of boubas:

- sound–shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–322.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
- Padraic Monaghan, Richard Shillcock, Morten Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369.
- David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.
- Plato. 1998. *Cratylus*. Hackett Publishing Company.
- V. S. Ramachandran and E. M. Hubbard. 2001. Synaesthesia—a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.
- Eyal Sagi and Katya Otis. 2008. Semantic glimmers: Phonaesthemes facilitate access to sentence meaning.
- K. Sathian and V. S. Ramachandran. 2019. *Multisensory perception: from laboratory to clinic*. Elsevier.
- M. Tamariz. 2008. Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3:259–278.
- Tim Wharton. 2003. Interjections, language, and the ‘showing/saying’ continuum. *Pragmatics & Cognition*, 11:39–91.
- Søren Wichmann, Eric Holman, and Cecil Brown. 2010. Sound symbolism in basic vocabulary. *Entropy*, 12.

Natural Language Generation in Dialogue Systems for Customer Care

Mirko Di Lascio[♡], Manuela Sanguinetti^{♡◇}, Luca Anselma[♡], Dario Mana[♣],
Alessandro Mazzei[♡], Viviana Patti[♡], Rossana Simeoni[♣]

[♡]Dipartimento di Informatica, Università degli Studi di Torino, Italy

[◇]Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Italy

[♣]TIM, Torino, Italy

[♡]{first.last}@unito.it, [◇]{first.last}@unica.it, [♣]{first.last}@telecomitalia.it

Abstract

English. In this paper we discuss the role of natural language generation (NLG) in modern dialogue systems (DSs). In particular, we will study the role that a linguistically sound NLG architecture can have in a DS. Using real examples from a new corpus of dialogue in customer-care domain, we will study how the non-linguistic contextual data can be exploited by using NLG.

1 Introduction

In this paper we present the first results of an ongoing project on the design of a dialogue system for customer care in the telco field. In most of the dialogue systems (DSs), the generation side of the communication is quite limited to the use of *templates* (Van Deemter et al., 2005). Templates are pre-compiled sentences with empty *slots* that can be filled with appropriate *fillers*. Most of commercial DSs, following the classical cascade architecture $NLUnderstanding \leftrightarrow DialogueManager \leftrightarrow NLGeneration$ (McTear et al., 2016), use machine learning-based Natural Language Understanding (NLU) techniques to identify important concepts (e.g., *intent* and *entities* in (Google, 2020)) that will be used by the dialogue manager (i) to update the state of the system and (ii) to produce the next dialogue act (Bobrow et al., 1977; Traum and Larsson, 2003), possibly filling the slots in the generation templates.

This classical, and quite common, information flow/architecture for dialogue processing has, as a working hypothesis, the assumption that most of *necessary* information is provided by the

user’s utterance: we call this information *linguistic channel* (L-channel). However, especially in the customer-care domain, this assumption is only partially true. For instance, in the sentence “*Scusami ma vorrei sapere come mai mi vengono fatti certi addebiti?*” (“Excuse me, I’d like to know why I’m charged certain fees?”), even a very advanced NLU module can produce only a vague information about the user’s request to the DialogueManager. Indeed, in order to provide good enough responses, the DialogueManager resorts to other two sources of information: the *domain context channel* (DC-channel) and the *user model channel* (UM-channel). The DC-channel is fundamental to produce the *content* of the answer, while the UM-channel is necessary to give also the *correct form*.

It is worth noting that both channels, that are often neglected in the design of commercial DSs for customer-care domain, have central roles in the design of (linguistically sound) natural language generation (NLG) systems (Reiter and Dale, 2000). In particular, considering the standard architecture for data-to-text NLG systems (Reiter, 2007; Gatt and Krahmer, 2018), the analysis of the DC-channel exactly corresponds to the *content selection* task and the UM-channel influences both the sentence planning and sentence realization phases. In other words, the central claims of this paper are that in commercial DSs for customer care: (1) *L-channel is often not informative enough and one needs to use the DC-channel and the UM-channel for producing a sufficiently good answer*, (2) *DC-channel and UM-channel can be exploited by using standard symbolic¹ NLG techniques and methods*. The remainder of the paper supports both of these claims while presenting our ongoing project on the development of a rule-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹The well-known problem of *hallucinations* in neural networks deters their use in real-world NLG (Rohrbach et al., 2018).

based NLG prototype to be used in a customer care domain. Section 2 presents the corpus developed in the first stage of this project, consisting of real dialogues containing explanation requests in telco customer-care domain. Section 3 presents an NLG architecture for managing the L-DC-UM channels that can be adopted in a DS for customer care. Finally, Section 4 concludes the paper with few remarks on the current state of the project and on future work.


2 A Dialogue Corpus for Customer-care Domain


This study builds upon the analysis of a corpus of dialogues between customers and a DS for customer service developed by an Italian telecommunications company. The dialogues, which take place by means of a textual chat, mainly deal with requests for commercial assistance, both on land-line and mobile phones. For the purpose of this study, the corpus was extracted by selecting, from a sample of dialogues held over 24 hours, a reduced subset that included requests for explanations from customers. The selection criteria were conceived so as to include all the dialogues where at least one message from the user contained a clearly stated request for explanation. The kind of requests identified in this collection basically reflects the problems typically encountered with a telecom service provider, such as undue or unfamiliar charges in the bill or in the phone credit (about 52% of the overall number of requests in this dataset).

The resulting corpus consists of 142 dialogues, with an average of 11 turns per dialogue, and an average length of 9 tokens in customer messages and 38 tokens in the bot messages. Such difference in the message length is due to the way the assistant's responses are currently structured, in that they usually include detailed information on invoice items or options available, while, on the other hand, customer's messages are most often quite concise. Also, the relatively high number of turns per dialogue might be explained with the high occurrence in the corpus of repeated or rephrased messages, both by the chatbot and by the customer, due to recurring misunderstandings on both sides.

As a matter of fact, the presence of such phenomena in the corpus, along with the overall goals set forth for the development of the NLG module


(a)

 **Vuoi vedere il dettaglio della fattura?**
(Do you want to see the invoice details?)


 **La fattura di gennaio 2020 non è arrivata**
(I haven't received the invoice of January 2020 yet)

Ignoring question/feedback - Neutral


Ecco il dettaglio delle voci di spesa presenti nella fattura InvoiceNumber del mese di gennaio per la linea PhoneNumber: InvoiceDetails.

 **Vuoi avere il dettaglio di una specifica voce di spesa presente in fattura?**
(Here is the detail of the items on the invoice InvoiceNumber of January for the PhoneNumber: InvoiceDetails. Do you want the detail of a specific item in the invoice?)


Straight wrong response

 **Non mi interessa questa fattura, mi serve gennaio 2020**
(I don't need this invoice, I need the one of January 2020)

Repetition - Somewhat frustrated


 **Vuoi cambiare argomento?**
(Do you want to change topic?)


Topic change

 **Ciao**
(Whatever)

Non-cooperativity - Somewhat frustrated

(b)


 **Scusami, non ho capito bene la tua richiesta...**
(I'm sorry, I didn't get your question... Could you rephrase it in a simpler way?)

 **No non hai capito niente**
(No you got it all wrong)

Non cooperativity - Somewhat angry

Mi spiace non aver compreso correttamente.

Ho la possibilità di farti chattare con un operatore umano più esperto. Vuoi?

 (I'm sorry I haven't understood that correctly. I can put you in contact with a human representative. Is this what you want?)

Empathy

Figure 1: Excerpts from the annotated dataset. Annotation of errors is highlighted in red, that of customer's emotions in blue and the agent's empathy in orange.

in this project, led us to the design of an annotation process that involved different dimensions, such as errors in conversation and emotions. By er-

ror, in this context, we mean any event that might have a negative impact on the flow of the interaction, and more generally on its quality, potentially resulting in breakdowns (i.e. whenever one party leaves the conversation without completing the given task (Martinovsky and Traum, 2003)). The error tagset used in this corpus is partially inspired by three of the popular Gricean maxims, i.e. those of *Quantity*, *Relation* and *Manner* (Grice, 1989) (each one including further sub-types, not described here), and it has been conceived so as to include error categories that may apply to both conversation parties. The second dimension, instead, is meant to include, on the one hand, customers’ emotions (as perceived in their messages), and, on the other hand, the chatbot’s empathic responses (if any). In particular, as regards customers’ emotions, besides two generic labels for neutral and positive messages, we mostly focused on negative emotions, especially anger and frustration, also introducing for these ones two finer-grained labels that define their lower or higher intensity. While a full description of the annotation scheme is beyond the scope of this paper, Figure 1 shows two brief examples of how we applied this scheme to the sample dataset². An overview of the scheme with a discussion on the main findings and annotation issues can be found in Sanguinetti et al. (2020).

Due to privacy concerns and the related anonymization issues that may arise (as further discussed in Section 4), the corpus cannot yet be publicly released. However, in an attempt to provide a qualitative analysis of the annotated data, we collected some basic statistics on the distribution of errors and emotions labeled in this sample set. Overall, we report an amount of 326 errors (about 21% of the total number of turns) from both parties; among them, the error class that includes violations of the maxim of Relevance is by far the most frequent one (65% of the errors). Such violations may take different forms, also depending on whether they come from the customer or the chatbot. As regards the customer, errors of such kind typically take place when the user does not take into account the previous message from the chatbot, thus providing irrelevant responses that do not allow to move forward with the conversation and make any progress; these cases cover

approximately 21% of customers’ errors. On the chatbot side, the most frequent error type is represented by those cases in which the agent misinterprets a previous customer’s message and proposes to move on to another topic rather than providing a proper response (30% of cases). As for the second annotation dimension, i.e. the one regarding customers’ emotions, most of the messages have a neutral tone (about 86% of user turns), but, among non-neutral messages, the two main negative emotions defined in this scheme, namely anger and frustration, are the ones most frequently encountered in user messages (both with a frequency of 41%), while the cases of messages with a positive emotion constitute less than 1%, and usually translate into some form of gratitude, appreciation, or simple politeness.

All these dimensions are functional to a further development of the NLG module, in that they provide, through different perspectives, useful signals of how, and at which point in the conversation, the template response currently used by the chatbot might be improved using the NLG module. Broadly speaking, framing the error taxonomy within the Grice’s cooperative principle provides a useful support for the generation module to understand, in case an error is reported, how to structure the chatbot response so as to improve the interaction quality in terms of informativeness and relevance (as also discussed in Section 3).

3 Balancing information sources in NLG for DS

In this Section, we illustrate a DS architecture that explicitly accounts for the L-DC-UM information channels. In particular, we point out that DC and UM channels can be managed by using standard NLG methods.

A commonly adopted architecture for NLG in data-to-text systems is a pipeline composed of four modules: data analyzer, text planner, sentence planner and surface realizer (Reiter, 2007; Pauws et al., 2019). Each module tackles a specific issue: (1) the data analyzer determines what can be said, i.e. a domain-specific analysis of input data; (2) the text planner determines what to say, i.e. which information will be communicated; (3) the sentence planner determines how to communicate, with particular attention to the design of the features related to the given content and language (e.g. lexical choices, verb tense, etc.); (4)

²For further details on the scheme and the definition of all tags, the annotation guidelines are available in this document: <https://cutt.ly/cdMcnYM>

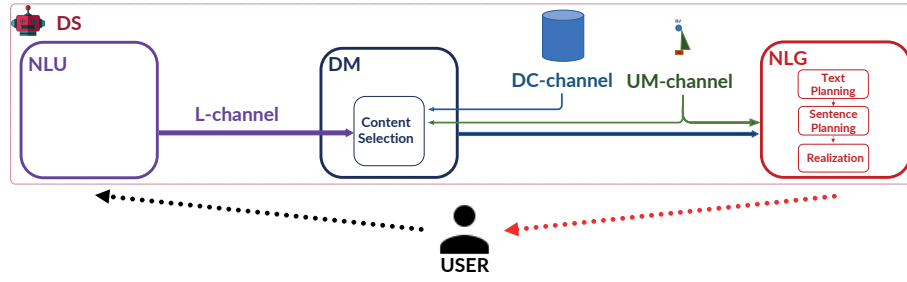


Figure 2: A dialogue system architecture accounting for L-DC-UM channels.

the surface realizer produces the sentences by using the results of the previous modules and considering language-specific constraints as well. Note that by definition NLG does not account for linguistic input (that is, L-channel), all the modules account for the context of the communication. In other words, data analysis and text planning explicitly process the information about the input data (the DC-channel), and text planning and sentence planning process the information about the audience (the UM-channel). Moreover, by using the nomenclature defined in (Reiter and Dale, 2000), the specific task of *content selection* decides *what to say*, that is the atomic nucleus of information that will be communicated.

In our project, we adopt a complete NLG architecture in the design of the DS (Figure 2). In Figure 2, we show the contributions of the L-DC-UM channels in the interaction flow. It is worth noting that we assigned the *content selection* task to the DM module rather than to the text planning of the NLG module. Indeed, the content selection task is crucially the point where all the three information channels need to be merged in order to decide the content of the DS answer to the user question.

In order to understand the contribution of the three information channels to the final message construction, we describe below the main steps of the module design using the following customer’s message, retrieved from the corpus, as an example:

Scusami ma vorrei sapere come mai mi vengono fatti alcuni addebiti?. (“Excuse me, I’d like to know why I’m charged certain fees?”)

Here, the customer requests for an explanation about some (unspecified) charges on her/his bill, making the whole message not informative enough. In this case, the DS can deduce from the L-channel only a generic request of information on transactions. However, using the architecture

shown in Figure 2, a more informative answer can be produced considering the UM-channel and the DC-channel.

As a working hypothesis, we assume that the user model consists uniquely in the age of the user. By assuming that the user is 18 years old, we can say that the DS should use an informal register, i.e. the Italian second person singular (*tu*) rather than the more formal third person singular (*lei*). It is worth noting that the current accounting of the user model is too simple and there is room for improvement both in the formalization of the model, and in the effect of the user model on the generated text. Taking into account the classification of the user model acquisition given by (Reiter et al., 2003), it is interesting to note that the dialogic nature of the system allow for the possibility to explicitly ask users about their knowledge and preferences on the specific domain.

Moreover, we assume that the DC-channel consists of all the transactions of the last 7 months, for example: T1, with an amount of 9.99€ (M1-M7); T2 with an amount of 2€ (M5-M7, appearing twice in M7); and T3 with an amount of 1.59€ (M7) (see Table 1).

	M1	M2	M3	M4	M5	M6	M7
T1	9.99	9.99	9.99	9.99	9.99	9.99	9.99
T2	0	0	0	0	2	2	2, 2
T3	0	0	0	0	0	0	1.59

Table 1: A possible transactions history.

Looking at the data in Table 1, different forms of automatic reasoning could be applied in order to evaluate the relevance of each singular transaction of the user. At this stage of the project, we aim to adapt the theory of importance-effect from (Biran and McKeown, 2017) to our specific domain, where the relevant information is in the form of relational database entries. The idea is to

consider the time evolution of a specific transaction category, giving more emphasis to information contents that can be classified as *exceptional evidences*. Informally, we can say that the transactions T2 and T3 have a more irregular evolution in time with respect to T1, therefore they should be mentioned with more emphasis in the final message.

The current implementation of the DS is based on a trivial NLU (regular-expressions), a symbolic sentence planner and realizer (for Italian) (Anselma and Mazzei, 2018; Mazzei et al., 2016). By considering all the three L-UM-DC channels, the answer generated by the DS is:

Il totale degli addebiti è €15,58. Hai pagato €4,00 (2×€2,00) per l'Offerta Base Mobile e €1,59 per l'Opzione ChiChiama e RiChiama. Infine, hai pagato il rinnovo dell'offerta 20 GB mobile. (“The total charge is €15.58. You have been charged €4.00(2×€2.00) for the Mobile Base Offer and €1.59 for the Who’sCalling and CallNow options. Finally, you have been charged for the renewal of the 20 GB mobile offer.”)

4 Conclusion and Future Work

In this paper we have discussed the main features of the design of a DS system for telco customer care. In particular, we outlined the peculiarities of this domain, describing the construction of a specifically-designed dialogue corpus and discussing a possible integration of standard DS and NLG architectures in order to manage these peculiarities. This is an ongoing project and we are considering various enhancements: (1) we will integrate emoji prediction capabilities into the proposed architecture in order to allow the DS to automatically attach an appropriate emoji at the end of the generated response, relying on previous work for Italian (Ronzano et al., 2018); we would also take into account the current user emotions, while generating an appropriate emoji – it may be the case that an emoji that is adequate when the conversation is characterized by a neutral tone, suddenly becomes inappropriate if the user is frustrated or angry (Pamungkas, 2019; Cercas Curry and Rieser, 2019); (2) we would like to enhance the system so as to adapt the generated responses to other aspects of the users, such as their mental models, levels of domain expertise, and personality traits; (3) we want to evaluate the DS following the user-based comparative schema adopted in

(Demberg et al., 2011).

Finally, we add some closing remarks on the corpus availability and its anonymization. The publication of a dataset of conversations between customers and a company virtual assistant is a great opportunity for the company and for its surrounding communities of academics, designers, and developers. However, it entails a number of obstacles to overcome. Rules and laws by regulating bodies must be strictly followed – see, for example, the GDPR regulation³. This means, first of all, including within the to-be-published dataset only those conversations made by customers who have given their consent to this type of treatment of their data. Moreover, it is mandatory to obscure both personal and sensitive customer data. Such obfuscation activities are particularly difficult in the world of chatbots, where customers are free to input unrestricted text in the conversations. Regular expressions can be used in order to recognize the pieces of data to be obscured, such as email addresses, telephone numbers, social security numbers, bank account identifiers, dates of birth, etc. More sophisticated techniques needed be adopted to identify and obscure, within the text entered by customers, names, surnames, home and work addresses. Even more complex and open is the problem of anonymizing sensitive customer data. For example, consider the case of a disabled customer who reveals his/her sanitary condition to the virtual assistant, in order to obtain a legitimate better treatment from the company: the text revealing the health condition of the customer must be obscured. Other relevant sensitive data include racial or ethnic origins, religious or philosophical beliefs, political opinions, etc. Some of these techniques, used for identifying certain types of data to be obscured, have a certain degree of precision that may even be far, given the current state of the art, from what a trained human analyst could do. Therefore, it is also necessary to consider the need for the dataset being published to be reviewed and edited by specialized personnel before the actual publication. With this in mind, the techniques of data recognition mentioned above - regular expressions, Named Entity Recognition, etc. - could also be exploited to develop tools that can speed up the task of completing and verifying the accurate anonymization of the dataset.

³<https://eur-lex.europa.eu/eli/reg/2016/679/oj>

Acknowledgements

The work of Mirko Di Lascio, Alessandro Mazzei, Manuela Sanguinetti e Viviana Patti has been partially funded by TIM s.p.a. (*Studi e Ricerche su Sistemi Conversazionali Intelligenti*, CENF_CT_RIC_19_01).

References

- Luca Anselma and Alessandro Mazzei. 2018. Designing and testing the messages produced by a virtual dietitian. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 244–253.
- Or Biran and Kathleen McKeown. 2017. Human-centric justification of machine learning predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1461–1467.
- Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. *Artif. Intell.*, 8(2):155–173, April.
- Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden, September. Association for Computational Linguistics.
- Vera Demberg, Andi Winterboer, and Johanna D. Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.
- Google. 2020. Dialogflow documentation. <https://dialogflow.com>. Online; accessed 2020-08-10 11:24:07 +0200.
- Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts.
- Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Dialogue Systems*.
- Alessandro Mazzei, Cristina Battaglini, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK, September 5-8. Association for Computational Linguistics.
- Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*. Springer Publishing Company, Incorporated, 1st edition.
- Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey. *CoRR*, abs/1906.09774.
- Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*, pages 119–145. Springer.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.
- Ehud Reiter, Somayajulu Sripada, and Sandra Williams. 2003. Acquiring and using limited user models in NLG. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.
- Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proc. of the 11th European Workshop on Natural Language Generation, ENLG '07*, pages 97–104, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, Nov. Association for Computational Linguistics.
- Francesco Ronzano, Francesco Barbieri, Endang Wahyu Pamungkas, Viviana Patti, and Francesca Chiusaroli. 2018. Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating Errors and Emotions in Human-Chatbot Interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop (LAW@COLING 2020)*. Association for Computational Linguistics.
- David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*, pages 325–353. Springer.
- Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24, March.

Cross-Language Transformer Adaptation for Frequently Asked Questions

Luca Di Liello[†], Daniele Bonadiman^{‡,*},
Cristina Giannone[†], Andrea Favalli[†], Raniero Romagnoli[†], Alessandro Moschitti[‡]

[‡]DISI, University of Trento, Italy

[†]Almawave Srl., Italy

{luca.diliello, d.bonadiman, alessandro.moschitti}@unitn.it
{c.giannone, a.favalli, r.romagnoli}@almawave.it

Abstract

Transfer learning has been proven to be effective, especially when data for the target domain/task is scarce. Sometimes data for a similar task is only available in another language because it may be very specific. In this paper, we explore the use of machine-translated data to transfer models on a related domain. Specifically, we transfer models from the question duplication task (QDT) to similar FAQ selection tasks. The source domain is the well-known English Quora dataset, while the target domain is a collection of small Italian datasets for real case scenarios consisting of FAQ groups retrieved by pivoting on common answers. Our results show great improvements in the zero-shot learning setting and modest improvements using the standard transfer approach for direct in-domain adaptation¹.

1 Introduction

Frequently Asked Question (FAQ) websites are an essential service for user’s self-assistance. FAQ websites typically present a list of questions, each associated with an answer. When searching for information, users have to go through the FAQs to determine whether there is a similar question providing a solution to their problem. However, this process does not scale well when the number of FAQs increases since too many questions may be presented to the user, and a simple search by the query may not retrieve the desired results. Additionally, in the last decade, users started looking for information using smartphones and voice assistants, such as Alexa, Google Assistant, or Siri.

^{*}work done prior to joining Amazon

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

By design, voice assistants provide users with a different information access paradigm: the FAQ websites’ navigation service is substituted by natural language dialogues, which satisfy the users’ information need in few interactions. To achieve this goal, FAQ retrieval systems need to understand the question and present the user only with a set of strong candidates. One possible solution offered by personal assistants is constituted by (i) a FAQ retrieval system (Caputo et al., 2016) for efficiently finding relevant questions, and (ii) accurate neural models to select the most probable FAQ.

One of the major obstacles for building such a system is the availability of training data for the selection model. FAQ systems are domain-specific in nature since they aim to provide users with information about specific websites or services. Moreover, the industrial setting does not always allow for creating a large corpus of questions for any specific domain, as the customers (FAQ’s owners) typically cannot provide such data. There are many reasons: (i) they are not familiar with the process of training data creation, as it is not part of their business; (ii) the topic of the FAQ system does not require more than tens of question/solution pairs; (iii) it is not easy to generate a dataset for question-question similarity from a question-answer system.

A traditional approach to alleviating such a problem is to use transfer learning (TL), i.e., data from other domains/tasks is used to train a model on the target task. TL research has been boosted by the availability of pre-trained transformer-based models (Vaswani et al., 2017; Devlin et al., 2018), which capture general-purpose language models. In this paper, we approach the problem of FAQ selection, fine-tuning pre-trained language models on the Question Duplication Task (QDT) from Quora². This task aims to identify whether

²<https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs>

Task	Question 1	Question 2	Label
QDT	How many months does it take to gain knowledge in developing Android apps from scratch?	How much time does it take to learn Android app development from scratch?	True
QDT	How do I prepare for software interviews?	What are the best ways to prepare for software interviews?	True
QDT	Why did Harry become a horcrux?	What is a Horcrux?	False
QDT	What is journalism for you?	What is journalism?	False
FAQ	Can medicines be sold on Amazon?	What items can't I sell on Amazon?	True
FAQ	I forgot my username	Why won't the page load?	True
FAQ	Is it possible to change my personal information after I have registered?	Is it possible to change the password?	False
FAQ	Can I have food brought from home during the flight?	What is included in the price I pay?	False

Table 1: Some examples of QDT and FAQ pairs. Notice that in the first block question are paraphrase of each other. The second block contains instead questions that only share a common answer.

two questions are duplicated or not, i.e., semantically equivalent or not. (Androutsopoulos and Malakasiotis, 2010).

Although the FAQ selection task shares some commonalities with QDT one, they are different. A FAQ task can indeed be solved by ranking all the FAQs in the collection using a system that computes the semantic similarity score between two questions, i.e., a Paraphrase Identification model. However, there are still some crucial differences. While QDT requires to infer if two questions are semantically equivalent, FAQ selection seeks questions that share the same intent and, at the same time, that they share the same answer. Moreover, the FAQ selection strongly depends on the domain in which the retrieval system is applied. For example, if a website responds to every technical complaint with “contact us”, there will be many positive pairs that will not share any real answer. Every portal in which a FAQ similarity system is needed, e.g., online services and e-commerce, requires a different level of details depending on the service type and its complexity. Table 1 provides some examples taken from QDT and FAQ datasets to underline the difference better.

One of the largest corpora for the fine-tuning of QDT is the well-known Quora dataset, sourced from the homonymous community question answering website. The dataset is constituted by question pairs, labeled as being duplicates or not. However, the Quora dataset is only available in the English language, preventing its use for building Italian systems.

In this paper, we propose to adapt Transformer architectures to the task of FAQ selection using machine translation. We first translated the Quora dataset to Italian, and then we trained a state-of-

the-art QDT model for Italian. Finally, we tested the adapted QDT model to two FAQ datasets showing significant improvement on the zero-shot learning baselines (i.e., using no target domain training data). Moreover, we show that fine-tuning the adapted model on small target data provides a consistent improvement over models not exploiting our transfer learning approach. It should be noted that our techniques can be seen as an extension of the Transfer and Adapt (TANDA) (Garg et al., 2019), but with the difference that transfer is carried out on a similar approximate task using translated data, i.e., Approximated machine Translated TANDA (ATTANDA).

The rest of the paper is organized as follows: Section 2 describes similar approaches to do Cross-Lingual Transfer Learning, Section 3 provides an overview of the available datasets and Section 4 describes the methodology we developed. Finally, Section 5 summarizes the main results and Section 6 draws the conclusions of this work.

2 Related Works

The current state of the art for QDT makes use of pre-trained transformer-based frameworks, e.g., BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) or XLNet (Yang et al., 2019). These models have millions of parameters that are trained in a two-step approach. First, they are trained as language models using various losses (e.g., masked language modeling or sentence order prediction loss) on a large corpus in an unsupervised way and then are fine-tuned on the target labeled dataset.

In Transfer Learning, a model is *transferred* (i.e., trained) on data coming from a high-resource task and is then adapted to another, usually more specific. All the Transformers-based models can

be seen as Transfer Learning models: they are first trained on large corpora of unlabelled data and then are specialized in a downstream domain. Nonetheless, there are scenarios where data about similar tasks can further improve already-great models.

Cross-lingual transfer-learning (CLTL) is an extension in which data from a high-resource language is used to solve a low-resource language task. This technique is sometimes used in combination with Cross-Lingual Word Embeddings alignment. The actual trend is to align word embeddings to focus only on shared language-independent features and then apply Transfer Learning techniques (Lange et al., 2020; Keung et al., 2020). However, solving a task using data coming from a similar one has different requirements.

A similar approach to ours has been explored by (Schuster et al., 2019), in which they used multilingual data to improve the performance of low-resource languages. However, even if they used translated data, they did not explore applying the transferred model to an affine task. Another approach (Do and Gaspers, 2019) filters high-quality samples from a high-resource language dataset to train the model in reduced time. Authors claim a significant improvement in the target language and task, even using only a small amount of computing.

In (Joty et al., 2017), the authors improve the performance in question-question similarity by using an adversarial approach. Thanks to adversarial training, they extract language-independent features from a trained model with supervision on a high-resource language and adapted to a low-resource one for testing. Results show important improvements in the target language, even in the zero-shot setting.

Also, in (Wang et al., 2020), a complete overview of the common approaches for cross-lingual transfer learning (CLTL) is proposed. Authors start by comparing (i) *joint training*, in which a model is trained on multilingual data using both a monolingual and a cross-lingual loss, and (ii) *CLWE alignment* before training, in which language embeddings are mapped to a shared space before fine-tuning. They find out that both methods perform well and that there is not an overall winner. Finally, they show that training with both approaches outperforms previous state-of-the-art

methods.

3 Datasets

3.1 Quora Question Pairs

The Quora dataset is a collection of question pairs for QDT. It contains many semantically equivalent questions that people asked more than once, for example, "What is the most populous state in the USA?" and "Which state in the United States has the most people?". Human experts have assigned labels; therefore, it is not free from subjective decisions and questionable labels. The dataset contains about 404K question pairs, 37% with a positive label, and 63% with a negative one. However, this dataset is not error-free: many ids are used more than once (14K), and many questions are referred by more than a single id (76K).

3.2 FAQ: RDC and LCN

RDC and LCN are two real-world datasets of FAQ retrieval. They were designed to build a QA component of conversational agent systems in Italian, targeting specific domains. Neither dataset is ready for FAQ retrieval out of the box, so we needed to group questions differently. Given that many questions share a common answer in RDC, we created several examples for the FAQ selection task by clustering questions with respect to the answers. For RDC, since the answers were simply the name of the category in which an answer could be found, we pivoted on the categories to create the clusters.

To build the examples, we first built clusters of equivalent questions, using their similarity gold standard labels, or rather the answers or the categories. LCN consists of 388 questions, which we grouped in 24 clusters of different sizes. The smallest contains only two elements, while the largest contains 50 elements. RDC contains 369 entries, which we grouped in 30 clusters with a minimum and maximum size of 1 and 37, respectively.³

Tests will show that LCN is the hardest dataset. The reason is that clustering has not been applied by pivoting on the answers but the same category instead (answers were not available). Then, each cluster contains questions that do share a precise answer but rather the same category.

³There is an Italian FAQ dataset called QA4FAQ, but it is not suitable for question similarity since annotations for the dataset are not available. <http://qa4faq.github.io>

The transformation of a set of clusters in a training or test set was done with the following algorithm: for N times, an element from each cluster was chosen, called champion, and was temporarily removed from its cluster. Each champion was then paired with a random element from every cluster, assigning positive labels when the two shared belonging to the same cluster. We found that $N = 5$ was a reasonable number of rounds since more would have led to information repetition.

Moreover, there was a need to create both small training and test sets to measure models' performance when fine-tuned on the FAQ domain. We could not divide the dataset described before since training and test sets would have had many common sentences. To accomplish a perfect separation, 70% of the clusters were used to create a train set while the remaining 30% were used for the test set.

3.3 FAQ: ItaFAQ

We built a small FAQ dataset in Italian by scraping popular websites. Then, we asked 10 different people with different backgrounds and levels of education to create additional questions similar to those automatically collected. The specific request was to create questions that would have had the same or a similar answer. The dataset is released as open-source and is available for download⁴. This dataset can be useful to test an information retrieval system. However, it is easier to solve than the previously described RDC and LCN. The main reasons are that (i) humans tend to create partially related new questions, and that (ii) general FAQ dataset about well-known companies and topics are easier to process than strong domain-specific data.

4 ATTANDA Approach

4.1 Machine Translation of Quora

There are no medium or large-size Italian datasets for QDT or FAQ retrieval; thus, we applied machine translation. We used *Microsoft Azure Cognitive Services* to translate Quora Question Pairs into Italian. Since the original Quora dataset had some questions repeated on different entries, we followed the approach in (Haponchik et al., 2018; Bonadiman et al., 2019) and grouped all the questions in clusters by mean of the transitive property:

⁴The dataset can be downloaded at <https://github.com/lucadiliello/italian-faq-dataset>

if a and b are the two questions of a pair with a positive label and C_i is a cluster, $a \in C_i \leftrightarrow b \in C_i$. Moreover, if there is a tuple (a, b) with a positive label and $a \in C_i, b \in C_j$, then C_i and C_j are merged in $C_k = C_i \cup C_j$.

After that, we translated all the questions of the clusters with at least two members. This allowed us to effectively reduce machine translation costs because we avoided translating questions that would have appeared only in negative pairs (millions of negative pairs can be easily generated by randomly picking questions from different clusters). We built the transfer dataset by labeling (i) all pairs of questions in the same cluster as positive examples; and (ii) a random number of pairs with members from different clusters as negative examples. We limited the number of the latter to be equal to the number of positive examples.

4.2 Transformer architectures

To reach the highest performance, we developed our models on the actual state of the art for QA. We took into consideration:

- **Multilingual BERT** (mBERT), a BERT model trained on the 104 largest Wikipedia, in terms of the number of articles. The model contains 177M⁵ parameters and has 12 transformer layers (Devlin et al., 2018);
- **Italian BERT**⁶, a BERT model trained only on Italian text. The version we used was trained over the concatenation of the OSCAR corpus and the Italian OPUS corpus, for a total of 81GB of text. This model features a total of 110M parameters on 12 layers;
- **GilBERTo**⁷, a RoBERTa model trained over 71GB of lowercase Italian text extracted from the OSCAR corpus. The authors state that this model applies masking to whole words (WWM), as in (Martin et al., 2020), instead of masking at the sub-words level, as in the original BERT. This model has a total of 111M parameters.

⁵mBERT has a bigger size since its vocabulary is considerably larger than monolingual models.

⁶Italian BERT models and code are available at <https://github.com/dbmdz/berts>

⁷GilBERTo models and code are available at <https://github.com/idb-ita/GilBERTo>

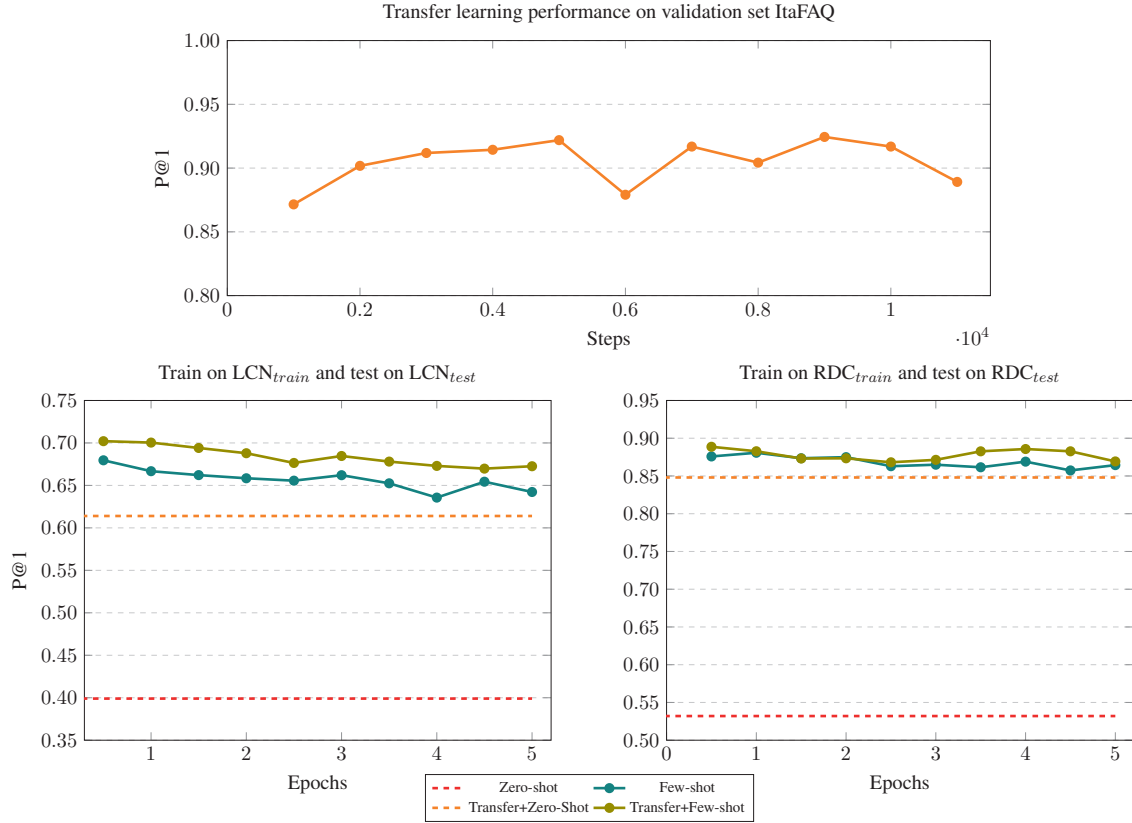


Figure 1: **Above:** Trend of the P@1 when transferring on Quora and validating on ItaFAQ. **Below:** Comparison of *Zero-Shot* (no training at all on Quora nor RDC_{train} or LCN_{train}), *Few-Shot* learning (only in-domain adaptation on RDC_{train} or LCN_{train}), *Transfer+Zero-Shot* learning (transfer on Quora and test directly) and *Transfer+Few-Shot* learning (transfer on Quora and in-domain adaptation on RDC_{train} or LCN_{train}). The results are the average of 8 runs with different seed and dataset splits. All runs used the same hyper-parameters: batch size of 32 and Adam optimizer with a learning rate of $1e-05$. Notice the different scales on the y-axis.

4.3 Cross-Domain training

We aim at exploiting data similar to the target task, which may also come from a different language, to train models for our FAQ target task. Our approach can be seen as an extension of TANDA by (Garg et al., 2019), which consists in two-step fine-tuning. First, they transfer the model on a general QA task with a huge dataset, and then they adapt the model to a smaller and specific QA benchmark such as WikiQA. They showed that a transfer step could improve the final performance if the source and target tasks are similar. We extend this idea by creating our transfer dataset utilizing machine translation, as described before. We call our approach ATTANDA (Approximated machine-Translated TANDA).

5 Results

This section shows the results of testing different models on the FAQ retrieval task. We use Precision at 1 (P@1), which is equal to accuracy, as we mainly need to measure if the returned FAQ is correct. LCN_{train} , LCN_{test} , RDC_{train} and RDC_{test}

are the names of the splits of LCN and RDC derived by dividing the set of clusters.

We start by comparing the available, transformer-based models. Table 2 shows that Italian BERT is better than the other models in most tests. This comes not as a surprise since it is specialized in the Italian language, it takes into consideration the case sensitivity of the input text, and it is trained on the most extensive corpus. GiBERTo also performs well, but RoBERTa’s improvement is insufficient to overcome the smaller training set and the case-insensitive tokenizer.

Once we established that the best pre-training model is Italian BERT, since it shows the highest scores in 3 comparisons out of 4, we tested different transfer methods on LCN and RDC splits. We compare the performance of Italian BERT in two scenarios: (i) the model is directly fine-tuned on the target domain, and (ii) the model is first transferred on Quora and then fine-tuned on the target domain (ATTANDA). We also report the results of the model without in-domain fine-tuning

Models	Dataset		Results	
	Train	Test	MRR	P@1
mBERT	-	LCN _{test}	45.4	25.5
IT BERT	-	LCN _{test}	56.6	39.9
GilBERTo	-	LCN _{test}	47.0	29.1
mBERT	-	RDC _{test}	59.4	41.0
IT BERT	-	RDC _{test}	65.1	53.2
GilBERTo	-	RDC _{test}	67.9	56.1
mBERT	Quora	LCN _{test}	64.3	49.2
IT BERT	Quora	LCN _{test}	75.1	61.4
GilBERTo	Quora	LCN _{test}	72.4	58.2
mBERT	Quora	RDC _{test}	88.4	81.1
IT BERT	Quora	RDC _{test}	91.1	84.8
GilBERTo	Quora	RDC _{test}	89.7	83.3

Table 2: Comparison of different transformers-based models. Each model in the bottom half of the table has been trained on Quora with the same hyper-parameters (batch size of 64 and Adam optimizer with a learning rate of $1e-05$) for a single epoch. Reported metrics are the average over 8 runs with different seeds and splits.

(zero-shot tests), taken from table 2.

Figure 1 reports the P@1 while training for the first five epochs on the test sets. This does not affect consistency of results since we do a comparison on the whole fine-tuning phase. All the plots show that transferring the model first on Quora gives an increase in P@1, especially in the early steps. Also, in this setting, training for more than two epochs did not provide further improvement, which could lead to over-fitting. This is intuitive as the training and test splits are small and also contain repeated information. There is no clear reason not to perform a transfer step since the resulting performance is at least equal and the computational effort to train for a single epoch on Quora is negligible.

6 Conclusion

We explored transfer learning in a typical industrial scenario where only small (or no) data is available in the target language. We showed that it is possible to use machine translated data to improve a strictly related task’s performance. We suspect that if the tasks had been more similar, for example, Question Answering and FAQ, the performance gain would have been even better. However, this was a real-world scenario where the target datasets were used for production in real websites, and size and quality were not large. In this setting, applying a transfer phase can improve the retrieval of similar questions, and the transfer step is a low-cost operation compared to the pre-training.

References

- Ion Androutsopoulos and Prodrornos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.
- Daniele Bonadiman, Anjishnu Kumar, and Arpit Mittal. 2019. Large scale question paraphrase retrieval with smoothed deep metric learning. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 68–75.
- Annalina Caputo, Marco de Gemmis, Pasquale Lops, Francesco Lovecchio, Vito Manzari, and Acquadotto Pugliese AQP Spa. 2016. Overview of the evalita 2016 question answering for frequently asked questions (qa4faq) task. In *of the Final Workshop 7 December 2016, Naples*, page 124.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Quynh Do and Judith Gaspers. 2019. Cross-lingual transfer learning with data selection for large-scale spoken language understanding. pages 1455–1460, 01.
- Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.
- Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237, Vancouver, Canada, August. Association for Computational Linguistics.
- Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2020. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner.
- Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*, pages arXiv–1907.

- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.
- Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

How good are humans at Native Language Identification?

A case study on Italian L2 writings

Elisa Di Nuovo

Elisa Corino

Dipartimento di Lingue e Letterature

Straniere e Culture Moderne

University of Turin

elisa.{dinuovo, corino}@unito.it

Cristina Bosco

Dipartimento di Informatica

University of Turin

bosco@di.unito.it

Abstract

In this paper we present a pilot study on human performance for the Native Language Identification task. We performed two tests aimed at exploring the human baseline for the task in which test takers had to identify the writers' L1 relying only on scripts written in Italian by English, French, German and Spanish native speakers. Then, we conducted an error analysis considering the language background of both test takers and text writers.

1 Introduction

Native Language Identification (NLI) is a task usually performed by machines consisting in identifying the mother tongue (henceforth L1) of a person based only on their writings in another language (e.g. L2 or L3¹). To date, the majority of the existing studies have focused on English as L2, that is English used by people who are acquiring English as a second or foreign language (Tomokiyo and Jones, 2001; Koppel et al., 2005; Malmasi et al., 2015; Kulmizev et al., 2017; Markov et al., 2017; Cimino and Dell'Orletta, 2017, among others). Three editions of the NLI shared task had been organized (Tetreault et al., 2013; Schuller et al., 2016; Malmasi et al., 2017) in which systems had to correctly identify the L1 among 11 L1s.

The basic assumption of this task is that when we learn a new language (henceforth Target Language, TL), our L1 interferes within the learning process introducing in the TL productions clues that can be automatically detected. Nevertheless, another issue to be investigated within this task

is the interference in the TL learner's productions also of other languages previously learned as L2 or L3. In fact, L1 may not be the only language playing a role in the acquisition of a TL, since "bi- or multilingualism is as frequent in the population of the world as pure monolingualism, perhaps even more frequent" (Hammarberg, 2001, p. 21). This issue is especially relevant performing the NLI task in languages other than English. For instance, when someone learns Italian, it is likely their L3, since English is the language taught worldwide as L2 (with more than 100 million learners, as stated in the British Council annual report 2018-19).

In this paper, we investigate the human performance for NLI applied on productions of learners Italian, thus focusing not only on the issues related to *second language acquisition* (Ellis, 2015; Slabakova, 2016), but also to *third language acquisition* (Cenoz et al., 2001; Picoral, 2020). We asked human Test Takers (TTs) to perform a simplified NLI task on learner Italian scripts extracted from VALICO (Corino and Marengo, 2017), selecting only four L1s²—i.e. English (EN), French (FR), German (DE) and Spanish (ES). This simplified task will be challenging since all the selected languages are Indo-European languages sharing typological similarities. Moreover, we performed an error analysis of the test results considering learners' L1 and L2(s) and learning stage (i.e. year of study), in addition to text features and TTs' language background. Test results could be useful for the improvement of the error annotation scheme introduced in Di Nuovo et al. (2019).

Our research questions therefore are:

1. How good are humans in performing the NLI task?
2. What features share the most problematic texts?

We try to answer to these questions in this paper organized as follows: in Section 2 we briefly de-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹L3 is any language acquired in addition to someone's L1 and another or more L2s.

²In evaluation campaigns, NLI systems are trained and tested on 11 L1s.

scribe previous work on the subject; in Section 3 we describe the tests performed and discuss the results; in Section 3.2 we conduct the error analysis; and in Section 4 we conclude the paper.

2 Related work

To the best of our knowledge, the first study assessing the human ability in identifying members of the same native language group is that of Ioup (1984). She demonstrated that native speakers were able to identify these groups only when relying on phonological cues, supporting the assumption that “syntactic errors in L2 acquisition cannot be accounted for as a result of the transfer of L1 patterns” (Broselow and Newmeyer, 1988, p. 197).

Odlin (1996) proved instead that readers who know the observed L1s (Korean and Spanish) can distinguish certain syntactic patterns in L2 English texts. Nowadays, his hypothesis is supported by the improved accuracy of machine learning systems using syntactic information (Malmasi et al., 2013; Markov et al., 2017, among others), such as PoS tags, Context-free Grammar Production Rules, Tree Substitution Grammar fragments, Grammatical Dependencies (see Malmasi (2016) for more details).

A more recent study simplified the NLI task to be performed by humans (Malmasi et al., 2015). The authors selected from TOEFL 11 (Blanchard et al., 2013)³ 30 texts (6 per each of the 5 languages included: Arabic, Chinese, German, Hindi, and Spanish), 15 considered *easy* and 15 *hard*, according to the ease with which most systems predicted the L1. They chose ten raters, all professors or researchers who might have had exposure to the 5 selected L1s. The average accuracy achieved by raters (37.3%, top rater 53.3% and lowest 30%) shows how difficult the task can be for humans. Their approach does not give attention to text features and writer’s language background that, in addition to experts’ knowledge of the 5 L1s involved, could have had an impact on the task performance.

Two NLI experiments on English L2 performed by humans are also reported by Jarvis et al. (2019). In both cases, humans were asked to identify in L2 English texts writers of their same L1. In the first experiment, six Finnish speakers were asked to identify the author of the L2 English text as being a Finnish or Swedish native

speaker. In the second experiment, in addition to the six Finnish raters, participated ten Spanish-speaker raters and all had to identify if the writers’ L1 shared their same L1 (Spanish or Finnish). The features that lead the raters to their decision were used to discuss the results achieved (in the second experiment, over 80% accuracy for the top Finnish raters and over 90% for the top Spanish raters). It is important to note that the accuracy achieved in Malmasi et al. (2015) and Jarvis et al. (2019) is not comparable since the experiment settings are completely different.

In this paper we describe experiments which are more similar to that carried out by Malmasi et al. (2015). However, differently from the papers described, we focus on data extracted from the VALICO corpus and in Italian, a language for which there are no previous studies about human NLI (see Malmasi and Dras (2017) for a multilingual approach to NLI in which they developed also a system for L2 Italian, using precisely texts collected from VALICO).

3 Test Description

We asked our TTs to perform two tests. The first (Test 1) is a simplified 4-class NLI test and the second (Test 2) a sort of guided Logistic Regression.

To Test 1 participated 25 TTs, 11 of them to Test 2.⁴ They are all Italian native speakers and were selected according to their knowledge of Italian as L2, foreign languages or linguistics. Our average TT has a master’s degree in a field related to Linguistics or Foreign Languages and speak on average two additional languages among the selected L1s.

The selected L1s—FR, DE, ES and EN—are the most commonly taught in Italy, so theoretically it would have been easier to find human experts knowing them⁵. In addition, these four languages represent two different families: while FR and ES belong to Romance languages like Italian, EN and DE belong to Germanic languages. For this reason we believe the task, although simplified (because constrained to only four languages), challenging

⁴We would like to stress the difficulty we faced in finding the TTs, not only due to the skills required, but also to the time and concentration required to perform these not at all short tests. We want also to point out that Test 1 is the experiment, to date, featuring the highest number of TTs, which are 10 in Malmasi et al. (2015) and 16 in Jarvis et al. (2019).

⁵However, we had difficulties in finding human experts knowing all the four languages. Only three out of twenty five know all the four considered languages.

³TOEFL 11 is the dataset used in NLI shared tasks.

enough. Furthermore, we expect different transfer patterns from the speakers of the two families.

Test 1 is a simplified NLI task, namely a multiple-choice test made of 48 questions. Each question contains a short text written in Italian by a non-native speaker and the TT had to identify the writer’s L1 choosing it between EN, FR, DE and ES. Texts were randomly selected from VALICOUD (Di Nuovo et al., 2019) with an almost balanced distribution with respect to the L1: 11 EN, 14 FR, 10 DE and 13 ES—TTs were not aware about this distribution. Their length ranged from 58 to 484 words (mean length 136.19 words, standard deviation (SD) of 64.73 words). There were no statistically significant differences in length between the L1 groups as determined by one-way ANOVA ($F = 2.24$, $p = .09$).

Test 2 consisted of 24 texts drawn from Test 1 according to the difficulty human TTs had in identifying the correct L1. The TTs were asked to assign a percentage to each of the four L1s involved, performing a sort of guided Logistic Regression: the higher the percentage assigned to a language, the higher chance of being the writer’s L1.

Table 1 resumes the information about the number of texts written by EN, FR, DE, ES native speakers (# Text/L1), and the percentage of TTs knowing that L1 (TT/L1) for both Test 1 and 2.

—	L1	# Text/L1	TT/L1
Test 1	EN	11	100%
	FR	14	72%
	DE	10	28%
	ES	13	56%
Test 2	EN	5	100%
	FR	10	82%
	DE	7	36%
	ES	2	73%

Table 1: Test 1 (25 TTs) and 2 (11 TTs) in figures.

3.1 Test Results and Discussion

The best result on Test 1 was achieved by TT1, correctly identifying the L1 of 26 out of 48 texts (54.2% accuracy), the lowest by TT25, correctly predicting the L1 of only 10 texts (20.8% accuracy). TT1 and TT25 are both PhD students in Digital Humanities: TT1 speaks EN and ES, while TT25 EN and DE. Table 2 shows score, accuracy (Acc.) and TTs’ teaching experience (TTE)⁶.

⁶As teachers of Italian as L2, EN, ES, FR or DE

These results suggest that, in our sample, teaching experience as well as knowing all the four L1s involved (as TT2, TT15 and TT23 did) were not discriminant factors. In addition, we checked if TTs speaking EN, FR, DE or ES could improve the identification of that L1(s) they know, but we did not find significant difference. Conversely to what stated by Jarvis et al. (2019, p. 222), TT18—knowing FR and Italian as L1 and EN as L2—did not identify more FR texts than other TTs knowing FR as L2.

TT	Score	Acc.	TTE
TT1	26	54.2%	yes
TT2, TT3, TT4, TT5	25	52.1%	3 yes, 1 no
TT6, TT7, TT8	24	50.0%	2 yes, 1 no
TT9	23	47.9%	no
TT10, TT11, TT12	22	45.8%	3 yes
TT13, TT14, TT15	21	43.7%	1 yes, 2 no
TT16	20	41.7%	yes
TT17	19	39.6%	yes
TT18	18	37.5%	yes
TT19	17	35.4%	yes
TT20	15	31.2%	yes
TT21, TT22, TT23, TT24	13	27.1%	1 yes, 3 no
TT25	10	20.8%	yes
Mean	20	SD (σ)	4.7

Table 2: Test 1 results - TT, score and accuracy achieved and TTE.

We classified the texts into 2 major categories, *easy* and *hard*, according to the percentage above or below 50% of correct answers assigned by the TTs, respectively. In total we have 17 easy texts (10 ES, 3 FR, 3 EN, 1 DE) and 31 hard texts (3 ES, 11 FR, 8 EN, 9 DE). Almost the totality of ES texts were identified by more than 50% of TTs, but we will comment on this in the next sections.

We further divided these two categories into sub-categories. Easy texts are divided into *Clear-cut* texts, in which authors’ L1 has been easily identified by almost all the TTs, and *Confusing* texts, identified by the majority of the TTs but a number leaned towards the same wrong class. Hard texts are further divided as *Scattered*, *Wrong Scattered*, and *Wrong Clear-cut*. In *Scattered* texts, votes are spread across two or three L1s, but the L1 receiving the majority of votes is the correct one. In *Wrong Scattered*, votes are spread across two or three L1s, but this time, the L1 receiving the majority of votes is a wrong one. Finally, in *Wrong Clear-cut* texts the L1 receiving more than 50% of the votes is an incorrect one.

Figure 1 shows the texts divided into these five categories. In the x axis we have the four possible

L1s and in y axis the text identification number. It is interesting to notice the similarity of Clear-cut and Wrong Clear-cut categories. In both categories, more than 50% of TTs opted for the same L1. The only difference relies in the fact that in the former it is the correct prediction, in the latter a wrong prediction.

Our hypothesis is that texts in the same subcategory share similar features. All Clear-cut texts contain spelling errors, literal translations or calques that clearly and explicitly lead to one of the four L1s (e.g. *cuando* for ES, *bizarre* for FR, *piance* for DE). Confusing texts had some ambiguous clues (e.g. *qualquosa* which can be written by ES or FR speakers, but for different reasons) which may cast doubt on at least 2 L1s. The main clues in Scattered texts were at a grammatical level (e.g. -ing form used instead of relative clauses, wrong agreement in gender and number), so TTs had to pay more attention. In Wrong Scattered and Wrong Clear-cut texts there were shallow clues, such as misspellings and loanwords—as in Clear-cut texts—for example *basura*, ES word for ‘garbage’) which might be indicative of negative transfer, but—differently from Clear-cut texts—these clues were not due to the L1 (in our example EN) but to other known L2s (in this case ES). However, in most of the cases, L2 transfers were in conjunction with L1 clues (such as the use of *ritornare* instead of *restituire* in “rotornare la borsa a Maria”, literal for ‘to return the bag to Maria’, or also *nel piano* instead of *per terra*, probably a translation error due to the polysemy of ‘floor’) that our TTs did not notice or considered less relevant clues. In order to clarify TTs recognition of the clues, we created another test, featuring the same 48 texts, but this time we told the TTs to highlight the clues that they used to make their prediction. We cannot provide information about this because we are still collecting the answers.

However, besides the clues, we wanted to capture also TTs’ uncertainty. Thus, in Test 2, we asked our TTs to assign a percentage to each L1 of the 24 most challenging texts. Figure 2 shows the average results for each L1 per text (correct answer in bold) divided into the subcategories of hard texts: Scattered (S), Wrong Scattered (WS), Wrong Clear-Cut (WCC).

Broadly speaking, there was high uncertainty among the TTs, not always in line with our hypotheses. Even when TTs were particularly sure

about one of the four L1s (e.g. assigning 99% to a L1 and 0% to the others), it was not always the correct L1, nor was explainable by writer’s L2 knowledge. For example in T14, T21, T48, most of TTs thought that ES was the correct L1. Although these texts showed numerous ES-like transfers (also syntactic ones, e.g. ES personal ‘a’ in *aiutare a la donna*, meaning ‘to help the woman’), the writers were FR native speakers, and only one of them claimed to know ES as L2. Still, in the majority of texts, the correct L1 received one of the two highest percentages, suggesting that L1 cues are present and that the TTs correctly interpreted them. It is also interesting to notice that also the analysis conducted by Malmasi (2016, p. 84) about the Oracle classifier suggests that the correct L1 is often in the top two predictions.

3.2 Error Analysis

We calculated recall, precision, accuracy and F1 aggregating all the TTs’ Test 1 answers per class; results—in terms of precision (Pre.), recall (Rec.), accuracy (Acc.) and F1—are shown in Table 3. As known, accuracy is influenced by the class distribution, hence F1 is a better metric in this case.

Overall, it was a challenging task as expected. On the one hand, ES proved to be easier to identify than the other three L1s (F1 score 56% compared to FR 38%). However, since all texts belonged to different proficiency levels not balanced across L1s, we cannot say if it is due to easily recognizable L1 patterns or to different interlanguage stages. On the other hand, DE proved to be the hardest L1 to identify. This could be motivated by the fact that only 28% of the TTs that participated to Test 1 have studied DE as L2. It is interesting to notice that DE is the most confused L1 to identify also in the experiment carried out by Malmasi et al. (2015) as clearly stated in Malmasi’s PhD thesis (2016, p. 88).

Class	Pre.	Rec.	Acc.	F1
EN	0.34	0.38	0.69	0.36
FR	0.41	0.36	0.66	0.38
DE	0.35	0.30	0.74	0.32
ES	0.52	0.60	0.74	0.56

Table 3: Test 1 aggregated results per class.

Linking these results with TTs’ language knowledge and writers’ L2 background, we can speculate that TTs’ language knowledge in itself

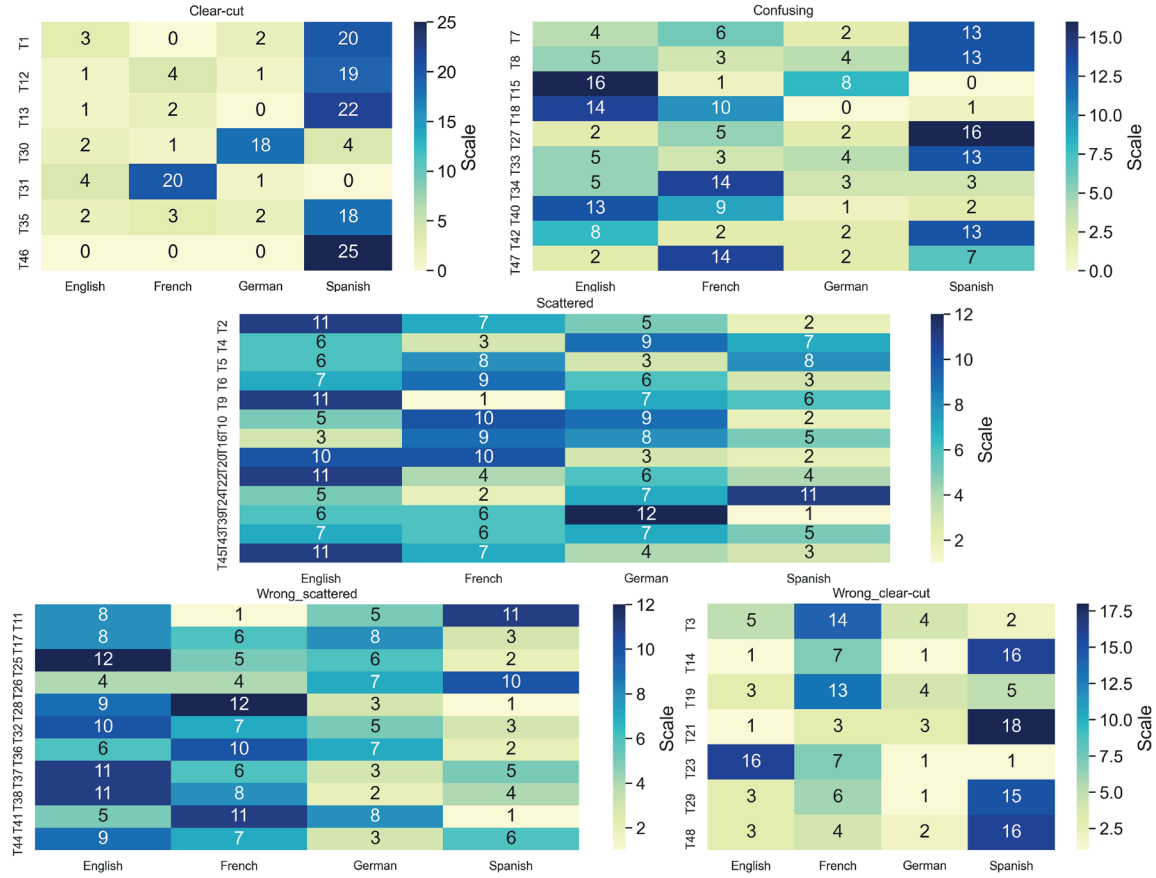


Figure 1: The five text categories stemmed from Test 1.

is not enough to improve L1 identification. In fact, all the TTs know EN as L2, but ES was the language identified correctly most of the time. This might be due to the fact that also all of the writers (except EN native speakers) know EN as L2, making the identification of EN native speakers harder.

We then plotted a confusion matrix (reported in Figure 3) to see the trends per class. Surprisingly, EN, correctly identified 38% of the time, was almost equally confused with the other three L1s (slightly more with FR and ES, 22% and 20% respectively, than with DE, 19%). FR, correctly identified 36% of the time, was confused more with ES (25%) and EN (24%) than with DE (15%). DE, correctly identified 30% of the time, was rarely confused with ES (13%), but frequently mistaken for FR (29%) and EN (28%). ES, when incorrectly identified (40% of the time), was confused slightly more with EN and FR (15% both) than with DE (10%). This might suggest that there is not a clear distinction between the two language families. In addition, it is interesting to notice that the directionality of the confusion is not always bidirectional for the four L1s (e.g. DE is confused

with FR and ES but not vice versa).

4 Conclusion

In this paper we described two human NLI tests for Italian. Although it was a simplified NLI task, tailored bearing in mind human skills, it proved to be a difficult task even for experts.

The error analysis showed that ES was the easiest L1 to identify—correctly identified 60% of the time—while DE the hardest. L2 transfer was misleading, even when L1 clues were present. TTs knowledge of the involved L1s proved not to be a discriminant factor.

It would have been interesting to ask the TTs to point out the clues that supported their answers to be less hypothetical in the discussion, especially when dealing with texts featured by L1 and L2 transfer. For this reason, we asked our TTs to take part in another test based on the same texts in which they have to highlight the clues. At the moment, we are collecting the answers.

In the future, we will test a machine learning system on the same texts to compare its results with those of our TTs.

S	EN	FR	ES	DE	WS	EN	FR	ES	DE	WCC	EN	FR	ES	DE
T2	26%	33%	27%	27%	T11	21%	10%	47%	37%	T3	40%	44%	13%	16%
T5	31%	37%	22%	14%	T17	22%	26%	23%	22%	T14	16%	17%	63%	11%
T6	28%	44%	12%	19%	T25	40%	19%	14%	25%	T19	25%	24%	20%	21%
T16	39%	12%	16%	42%	T26	25%	23%	29%	20%	T21	14%	22%	54%	14%
T20	35%	40%	18%	22%	T28	38%	18%	8%	35%	T23	35%	34%	9%	23%
T43	40%	28%	10%	27%	T32	34%	16%	8%	35%	T29	25%	8%	45%	25%
					T36	41%	25%	11%	20%	T48	20%	25%	56%	14%
					T37	54%	18%	17%	14%					
					T38	45%	25%	22%	18%					
					T41	33%	17%	26%	30%					
					T44	31%	25%	36%	20%					

Figure 2: Test 2 results aggregated for L1 and text.

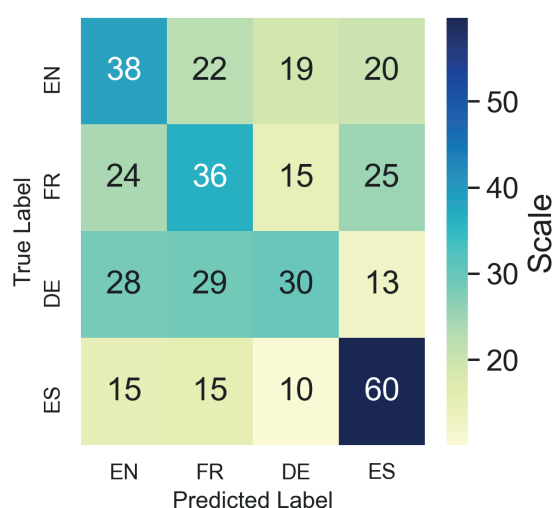


Figure 3: Confusion matrix with Test 1 aggregated data.

Acknowledgements

We would like to thank the anonymous reviewers as well as our TTs, who have dedicated their time to us by carrying out these long tests which require a lot of concentration: E. L. Baratono, C. Borge, C. Borgia, C. Bosco, V. Calabria, S. Cicillini, A. T. Cignarella, C. Conti, V. De Iacovo, G. Esposito, K. Florio, G. Giaccone, A. Giacosa, L. Giannelli, L. Inserra, I. Iubini, A. Marra, M. Pellegrini, S. Peyronel, F. Poletto, S. Racca, A. Rotolo, M. Sanguinetti, E. Truc, and M. C. Zaccone.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Ellen Broselow and F Newmeyer. 1988. Second language acquisition. *Linguistics: The Cambridge Survey: Volume 3, Language: Psychological and Biological Aspects*, pages 194–209.

Jasone Cenoz, Britta Hufeisen, and Ulrike Jessner. 2001. *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Multilingual Matters.

Andrea Cimino and Felice Dell’Orletta. 2017. Stacked Sentence-Document Classifier Approach for Improving Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437.

Elisa Corino and Carla Marengo. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*. Guerra.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian Learner Treebank in Universal Dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–8. CEUR-WS.

Rod Ellis. 2015. *Understanding second language acquisition 2nd Edition-Oxford applied linguistics*. Oxford university press.

Björn Hammarberg. 2001. Roles of L1 and L2 in L3 production and acquisition. In Jasone Cenoz, Britta Hufeisen, and Ulrike Jessner, editors, *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*, volume 31, pages 21–41. Multilingual Matters.

Georgette Ioup. 1984. Is there a structural foreign accent? A comparison of syntactic and phonological errors in second language acquisition. *Language Learning*, 34(2):1–15.

Scott Jarvis, Rosa Alonso Alonso, and Scott Crossley. 2019. Native language identification by human judges. In *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*, pages 215–231. Springer.

- Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217. Springer.
- Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389.
- Shervin Malmasi and Mark Dras. 2017. Multilingual Native Language Identification. *Natural Language Engineering*, 23(2):163–215.
- Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. Nli shared task 2013: Mq submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133.
- Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Buman Baselines for Native Language Identification. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 172–178.
- Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.
- Shervin Malmasi. 2016. *Native Language Identification: explorations and applications*. Ph.D. thesis, Macquarie University, Sydney, Australia.
- Ilia Markov, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. CIC-FBK approach to native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 374–381.
- Terence Odlin. 1996. On the recognition of transfer errors. *Language Awareness*, 5(3-4):166–179.
- Adriana Picoral. 2020. *L3 Portuguese by Spanish-English Bilinguals: Copula Construction Use and Acquisition in Corpus Data*. Ph.D. thesis, The University of Arizona.
- Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity Native Language. In *Interspeech 2016*, pages 2001–2005.
- Roumyana Slabakova. 2016. *Second language acquisition*. Oxford University Press.
- Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57.
- Laura M. Tomokiyo and Rosie Jones. 2001. You're Not From 'Round Here, Are You? Naive Bayes Detection of Non-Native Utterances. In *NAACL*.

You Don't Say...

Linguistic Features in Sarcasm Detection

Martina Ducret Lauren Kruse Carlos Martinez Anna Feldman Jing Peng

Montclair State University
Montclair, New Jersey, USA

{ducretm1,krusell,martinezcl1,feldmana,pengj}@montclair.edu

Abstract

We explore linguistic features that contribute to sarcasm detection. The linguistic features that we investigate are a combination of text and word complexity, stylistic and psychological features. We experiment with sarcastic tweets with and without context. The results of our experiments indicate that contextual information is crucial for sarcasm prediction. One important observation is that sarcastic tweets are typically incongruent with their context in terms of sentiment or emotional load.

1 Introduction

Sarcasm, or verbal irony, is a figurative language device employed to convey the opposite meaning of what is actually being said. In verbal communication, a pause, intonation, or look can provide the cues necessary to determine whether there is sarcastic intent behind a comment. In writing, these social cues are inaccessible. Thus, we must rely on our understanding of the world, the speaker, and the context beyond the statement to discern between sarcasm and sincerity. This task has proven to be so subjective that social media users moderate their own comments using symbols and hashtags such as /s and #sarcasm to denote the sentiment on Reddit and Twitter, respectively. In fact, the dataset used in this paper was collected using such hashtags (Ghosh et al., 2020).

For machines, the lack of real-world knowledge is detrimental to their understanding of sarcasm as it hinders many natural language processing applications. Beyond social-media conversations, assessing product reviews as positive or negative requires an understanding of both rhetorical and literary devices. Back in 2012, BIC rolled out a “For

Her” line of pens which led their intended female audience to poke fun at the misogynist message of the product. One reviewer commented, “Well at last pens for us ladies to use... now all we need is “for her” paper and I can finally learn to write!”. While this review seems positive and gave the product four stars, our understanding of the social climate today leads us to conclude that this review is sarcastic and should be classified as such.

In social media communication, new slang words are introduced every day and emojis are often used to negate the sentiment of the text. In addition, stylistic devices and stylometric features are also often employed to convey a meaning opposite from its literal interpretation. While deep learning models can be very effective in their detection of sarcasm, they provide a “black box” approach that gives linguists little to no insight into what features are characteristic of sarcasm. The purpose of the current work is to learn *linguistic* patterns associated with sarcastic tweets and their contexts and determine which are the strongest indicators of sarcasm. The next step is to combine these observations with transformer-based architectures to achieve a better prediction accuracy.

2 Previous Work

The field of automatic sarcasm recognition has become quite active in recent years. The most current event is the shared task (Ghosh et al., 2020) organized as a part of the 2nd FigLang workshop at ACL 2020. The task is typically framed as a binary classification task (sarcastic vs. non-sarcastic) considering either an utterance in isolation or in combination with contextual information. Early approaches to automatic sarcasm detection rely on different types of features, including sarcasm markers, word embeddings, emoticons, patterns between positive and negative sentiment (e.g., Davi-

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

dov et al. 2010; Tsur et al. 2010; González-Ibáñez et al. 2011; Riloff et al. 2013; Maynard and Greenwood 2014; Wallace et al. 2015; Ghosh et al. 2015; Joshi et al. 2015; Veale and Hao 2010; Liebrecht et al. 2013). Buschmeier et al. (2014) explore a range of features, mainly focused on sentiment, for the detection of verbal irony in product reviews. While this paper provides a good baseline for irony classification, our data differs in that it includes a multi-speaker thread of context prior to the sarcastic remark. More recent approaches apply deep learning methods (e.g., Ghosh and Veale 2016; Tay et al. 2018; Wallace et al. 2015). There is a great amount of research exploring the role of contextual information for sarcasm detection (e.g., Joshi et al. 2015; Bamman and Smith 2015; Misra and Arora 2019; Bamman and Smith 2015; Khattri et al. 2015; Amir et al. 2016; Rajadesingan et al. 2015; Ghosh and Veale 2017; Schifanella et al. 2016; Cai et al. 2019; Castro et al. 2019). Ghosh et al. (2020) report that almost all systems submitted as part of the shared task have used the transformer architecture, such as BERT (Turc et al. 2019) or RoBERTa (Liu et al. 2020), and other variants. They performed better than RNN architectures, even without any task specific fine-tuning. Unfortunately, it is difficult to interpret what these models capture about sarcastic tweets and their context. Our approach uses classical supervised algorithms to better understand which elements characterize sarcasm in a social media setting. We categorize linguistic features, experiment with different combinations, and take context into account when performing our experiments.

3 Our Approach

Our approach utilizes a combination of complex, stylometric, and psychological linguistic features to automatically detect the presence or absence of sarcasm in a given text. We intentionally experiment with classical machine learning classification algorithms to get a better understanding of the linguistic features contributing to the sarcasm detection task. Our linguistic intuition is that there will be a discordance between the linguistic features corresponding with the responses and contexts labeled as sarcastic. Sarcastic tweets are likely to be semantically or emotionally incongruent with their preceding tweets, while non-sarcastic tweets show a greater harmony with their context. To measure the emotional load of a response and its context,

we extract a number of sentiment- and emotion-related features. We also look at the distribution of these features across the two classes. Furthermore, we test the performance of our classifier and importance of our features by considering just the response tweet versus the response with its accompanying context.

4 Data Set

We use the Twitter Corpus from the CodaLab shared task on sarcasm detection (Ghosh et al., 2020). The training data consists of 2,500 tweets labeled ‘SARCASM’ and 2,500 tweets labeled ‘NON SARCASM’, the balanced test data consists of an additional 1,800 labeled tweets. Ghosh et al. (2020), this is a self-labeled data set where the tweets are annotated as sarcastic based on the hashtags used by the users. The non-sarcastic tweets are the ones that do not contain the sarcasm hashtags, but may be labeled with either positive or negative sentiment hashtags, such as ‘#happy’. Retweets, duplicates, quotes, etc., are excluded (see Ghosh et al. 2020 for more details). Each sarcastic and non-sarcastic tweet is accompanied with an hierarchical conversation thread, e.g., context/1 is the immediate context, context/0 is the context that preceded context/1, and so on. The training and test data include up to 19 preceding tweets labeled as context/0, context/1, ..., context/19 (if available).

5 Feature Extraction

Our research focuses on the role linguistic features play in sarcasm detection. We classify our features into three categories: *complexity*, *stylistic*, and *psychological*. Abonizio et al. (2020) defines complexity features as linguistic features that capture the overall objective of the context at the word and sentence level. Stylistic features use natural language techniques to gain grammatical information to better understand the syntax and style of the document. Psychological features are closest related to emotions and the cognitive aspect of NLP. We expand on these psychological features by utilizing VAD (*Valence, Arousal, Dominance*) (Warriner et al., 2013), emotional embeddings, and LIWC (Tausczik and Pennebaker, 2010). Lastly, we use word-level count vectors, word-level tf-idf, n-gram word-level tf-idf, n-gram character-level tf-idf. We stack these features and refer to them as *count vectors* for the remainder of this paper.

5.1 LIWC

LIWC (Tausczik and Pennebaker, 2010) is a text analysis program with a built-in dictionary that counts words in psychologically meaningful categories. After all the words have been reviewed, the module calculates the total percentages of words that are similar and match that of the user dictionary categories. We used LIWC to extract features to detect and categorize the meaning, emotional sentiment, and social relationship of the words in the data set.

5.2 Valence, Arousal, Dominance (VAD)

VAD (*Valence Arousal Dominance*) (Warriner et al., 2013) includes almost 14,000 lemmas rated on a 1-9 scale according to the emotions evoked by the terms. Valence refers to the pleasantness of the word, arousal determines how dull or exciting the emotion is, and dominance ranges from submission to feeling in control. The VAD dimensions allow us to further explore the affective meanings of tweets and determine their viability as a predictor of sarcasm. We compute VAD scores for each “response” and use the three scores obtained as a feature in our classifiers. Furthermore, we explore using the scores as a measure of congruity between our response and contexts. We calculate the VAD scores for each individual response and context and then subtract the response scores by their respective context scores. In other words, if a response receives a valence score of 8 and its context/0 receives a valence score of 2, the valence congruity score would be a 6. We hypothesize that sarcastic tweets might show very little affective congruity compared to their non-sarcastic counterparts.

5.3 VADER

VADER (*Valence Aware Dictionary and sEntiment Reasoner*) (Hutto and Gilbert, 2015) is a lexicon and rule-based tool built especially for sentiment analysis of social media texts. VADER maps lexical features to emotions and provides insight into the intensity of such emotions through a series of polarity indices. VADER considers capitalization, punctuation, degree modifiers, emojis, and negations to compute its negative, positive and neutral scores. Furthermore, VADER’s compound score provides a normalized, weighted composite score for a given tweet.

		Context/0							
Response	Not-Sarcasm		Anger	Fear	Joy	Love	Sadness	Surprise	Total
		Anger	221	35	176	11	69	6	518
		Fear	48	27	51	1	25	4	156
		Joy	279	72	898	34	131	20	1434
		Love	4	2	48	7	5	2	68
		Sadness	86	24	101	5	60	9	285
		Surprise	16	3	13	0	4	3	39
		Total	654	163	1287	58	294	44	2500
	Sarcasm	Anger	498	120	345	17	116	26	1122
		Fear	83	33	61	3	12	6	198
		Joy	333	72	289	13	96	22	825
		Love	9	3	6	1	3	1	23
		Sadness	98	24	88	2	41	3	256
		Surprise	30	10	28	2	5	1	76
		Total	1051	262	817	38	273	59	2500

Figure 1: Distribution of Emotions for Response vs. Context/0 in the training data.

5.4 Emotional Embeddings

The emotions conveyed in our data set are portrayed through emotional embeddings. Calculating the emotions of the text goes a level deeper than just looking at the word embeddings. Using a pre-trained model from Hugging Face (Saravia et al., 2018), we categorize the tweets into six emotions. The emotions include, *joy*, *anger*, *fear*, *surprise*, *sadness* and *love*. Figure 1 above represents an example of the distribution of emotions between response and context/0 in the balanced training data set. The results support our intuition that sarcasm is typically associated with negative emotions. When the context is labeled as “anger”, non-sarcastic tweets tend to respond with joy, while sarcastic tweets usually respond with anger. By contrast, when the context is labeled as “joy”, non-sarcastic tweets overwhelmingly respond with joy, while sarcastic tweets still largely respond with anger. There are 1,216 instances of the same emotion expressed in both response and context for the non- sarcasm class and 863 instances of this in the sarcasm class. Sarcastic tweets are generally incongruent with emotions throughout the response and context, unless associated with a negative emotion, e.g., *anger*.

5.5 Tweet-Context Similarity Scores

We use the standard document similarity estimation technique using word embeddings (GloVe, Pennington et al. 2014) and emotional embeddings (Saravia et al. 2018), which consists of measuring the similarity between the vector representations of the two documents. Let x_1, \dots, x_m and y_1, \dots, y_n be the emotion (or word embedding) vectors of two documents. The cosine similarity value between the two documents (e.g., a tweet and its context)

centroids $C_x = \frac{1}{m} \sum_{i=1}^m x_i$ and $C_y = \frac{1}{n} \sum_{i=1}^n y_i$ is calculated as follows:

$$\cos(C_x, C_y) = \frac{\langle C_x, C_y \rangle}{\|C_x\| \|C_y\|}, \quad (1)$$

where $\langle x, y \rangle$ denotes the inner product of two vectors x and y .

We compute two similarity scores: 1) semantic cosine similarity using word embeddings; 2) cosine similarity using emotional embeddings. Our linguistic intuition is that a sarcastic response is going to be semantically or emotionally incongruent with its context and this is what creates the sarcasm effect.

	Message
c/0	It's no secret that this president has routinely targeted religious and ethnic minorities. He has fanned the flames of hate against refugees, Muslims, Africans, immigrants, women and all racial and religious minorities.
c/1	He is routinely and openly hostile to any legitimate Congressional oversight. He has made clear his wanton corruption by soliciting a bribe from a foreign government for his personal political gain.
R	Yassss queen, you're so brave and bold.

Table 1: Sarcastic Tweet.

Response=R; Context0=C/0; Context1=C/1

	Message
c/0	A2 I revert back to Canvas. I am sure you can post assignments for parents in this, (haven't done this yet). Canvas = #thebomb #KidsDeserveIt
c/1	Can you talk me more about Canvas? I haven't heard of it.
R	It's Edmodo with #MorePower You can create assignments in it, post all work, the assignments can be auto graded and imported into your Skyward grade book.

Table 2: Non Sarcastic Tweet.

Response=R; Context0=C/0; Context1=C/1

Table 1 is an example of a sarcastic tweet whose context/0, context/1 and response received an emotion of anger, anger, and joy, respectively. Table 2 represents a non-sarcastic thread of tweets where each message was classified as joy. This indicates that non-sarcastic tweets tend to be more emotionally similar to the preceding context while sarcastic tweets tend to shift in emotion. As a result, when compared to its contexts, the sarcastic tweet received lower emotional similarity scores than the non-sarcastic tweet.

5.6 Feature Analysis

After running all of the features on the training data, we implemented SHAP (*SHapley Additive exPlanations*) (Lundberg and Lee, 2017) to determine which features are the most important for classification. SHAP is a theoretic output technique that explains predictions of our model, by producing a SHAPLEY score that plots the most important features in our model. The features produced by SHAP were used in our experiments and are referred to as our “select linguistic features”. The top 20 features SHAP selects contain a combination of character features such as character count, as well as a number of sentiment features, including VADER scores, emotion scores for both a response and its context as well as VAD features.

6 Experimental Evaluation

6.1 Data Preprocessing

Our preprocessing procedure consists of steps to remove noisy and unnecessary data. First, we tokenize and lemmatize the tweets using NLTK (Loper and Bird, 2002). We also remove any instance of “@USER” due to the repetition of this token in the beginning of most tweets. Prior research demonstrated that classifiers did not tend to benefit from large quantities of additional context and we noticed that a majority of the tweets only contained context/0 and context/1. While we plan to experiment further with additional context layers, in this work we only report on experiments that involve context/0 and context/1. We did not remove any stop words due to the small amount of text in each tweet. We also maintained punctuation and emojis as they proved to be useful information during the extraction of certain features, such as VADER.

7 Results

We use a Random Forest classifier and run 21 different experiments of which the most relevant ones are outlined in Table 3. The baseline scores represent an attention based LSTM model described in Ghosh et al. (2018) and used in the CodaLab Shared Task. We look at how each feature performed on just the response versus the response and context. We notice that for response, a combination of all count features and all linguistic features achieves the best F1 score of 67%. This score is further increased to 70% when the context is considered.

Experiments	P	R	A	F1
Baseline 1 Shared Task	70%	66.9%	N/A	68%
R (All Ct. Ft.)	72%	61%	63%	66%
R(All Ling Ft.)	56%	60%	59%	52%
R(Sel. Ling Ft.)	54%	60%	59%	57%
R(All Ling Ft. +All Ct.)	70%	64%	65%	67%
R(Sel. Ling Ft. +All Ct.)	65%	51%	51%	57%
R+C/0+C/1 (All Ct. Ft.)	67%	61%	62%	64%
R+C/0+C/1(All Ling Ft.)	64%	60%	60%	53%
R+C/0+C/1(Sel. Ling Ft.)	71%	60%	62%	64%
R+C/0+C/1(All Ling Ft. +All Ct.)	80%	62%	66%	70%
R+C/0+C/1(Sel. Ling Ft. +All Ct.)	70%	65%	66%	67%

Table 3: Random Forest; Various feature combinations. *Response*=R; *Context0*=C/0; *Context1*=C/1; *Count*=Ct; *Linguistics*= Ling; *Features*= ft.

8 Conclusion

In this paper we explored the role various linguistic features play in computational sarcasm detection. We investigated a combination of text and word complexity features, stylistic and psychological features. The result of our experiments indicate that contextual information is crucial for sarcasm detection. We also observed that sarcastic tweets are often incongruent with their context in terms of sentiment or emotional load. Using a Random Forest classifier and the features we extracted we obtain promising results. Our current work is concerned with combining these observations with transformer-based architectures to achieve a better prediction accuracy.

Acknowledgments

This work is supported by the US National Science Foundation under Grant No.: 1704113.

References

Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a

classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2016. *Fracking sarcasm using neural network*. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

- C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. [Harnessing context incongruity for sarcasm detection](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.
- Anupam Khattry, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an author’s historical tweets to predict sarcasm. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 25–30.
- Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. [The perfect solution for detecting sarcasm in tweets #not](#). In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia. Association for Computational Linguistics.
- Yinhan Liu, Mylène Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#).
- Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia: Association for Computational Linguistics.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Diana Maynard and Mark Greenwood. 2014. [Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *ArXiv*, abs/1908.07414.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. [Sarcasm as contrast between a positive sentiment and negative situation](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. [CARER: Contextualized affect representations for emotion recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.
- Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.
- Yla R. Tausczik and James W. Pennebaker. 2010. [The psychological meaning of words: Liwc and computerized text analysis methods](#). *Journal of Language and Social Psychology*, 29(1):24–54.
- Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. [Reasoning with sarcasm by reading in-between](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.
- Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. [Icwsn - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews](#). In *ICWSM*. The AAAI Press.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.
- Tony Veale and Yanfen Hao. 2010. [Detecting ironic intent in creative comparisons](#). In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 765–770. IOS Press.
- Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. [Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Amy Warriner, Victor Kuperman, and Marc Brysbaert.
2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45.

Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI¹

Manuel Favaro

Istituto di Linguistica Com-
putazionale “A. Zampolli” -
CNR

manuel.favaro@ilc.cnr.it

Marco Biffi

Università di Firenze
Accademia della Crusca
marco.biffi@unifi.it

Simonetta Montemagni

Istituto di Linguistica Com-
putazionale “A. Zampolli” -
CNR

simonetta.montema-
gni@ilc.cnr.it

Abstract

Italiano Questo contributo si propone di presentare il progetto TrAVaSI (*Trattamento Automatico di Varietà Storiche di Italiano*), il cui obiettivo è la creazione di risorse per il trattamento automatico di varietà storiche della lingua italiana, in particolare lessici diacronici e corpora arricchiti con annotazione linguistica da utilizzare per lo sviluppo e/o la specializzazione di strumenti di annotazione. Il contributo illustra gli obiettivi, i primi risultati conseguiti e le prospettive di sviluppo.

English *The paper is aimed at illustrating the TrAVaSI project, whose aim is the design and development of language resources for the automatic processing of historical varieties of the Italian language, in particular diachronic lexicons and corpora enriched with linguistic annotation to be used for the development and/or specialization of annotation tools. The results achieved so far are reported together with current and future directions of research.*

1 Introduzione

Sono ormai numerosi gli archivi testuali digitali che testimoniano varietà storiche dell'italiano. Tuttavia, l'accesso e l'interrogazione dei testi sono spesso elementari, per lo più limitati alle stringhe di caratteri che costituiscono il testo; ciò complica il lavoro degli studiosi e rende pressoché impossibile l'utilizzo delle risorse da parte degli utenti non addetti ai lavori. Questa situazione mostra che,

nonostante i recenti progressi nel settore delle Digital Humanities, l'accesso e l'interrogazione di testi che testimoniano varietà storiche di italiano, più o meno lontane nel tempo, rappresentano ancora oggi una sfida.

Il progetto TrAVaSI (*Trattamento Automatico delle Varietà Storiche di Italiano*), nato dalla collaborazione tra l'Istituto di Linguistica Computazionale “Antonio Zampolli” e l'Accademia della Crusca e finanziato dalla Regione Toscana, si propone di affrontare questa sfida, creando i presupposti per la navigazione e l'interrogazione sistematica di fonti che documentano le varietà storiche della lingua.

Il punto di partenza del progetto è pragmatico: potenziare due strumenti realizzati dall'Accademia della Crusca all'interno di altri progetti. Gli strumenti in questione sono la versione elettronica del *Grande Dizionario della lingua italiana (GDLI)*² e la banca dati del *Vocabolario Dinamico dell'Italiano Moderno*, che è più precisamente l'oggetto specifico di indagine del contributo che qui presentiamo. TrAVaSI ha come obiettivo principale quello di massimizzare le implementazioni pratiche, ma proiettandole nel quadro dello sviluppo di strumenti di riferimento per banche dati diacronicamente connotate e i dizionari storici in versione elettronica. In particolar modo si tratta di mettere a frutto l'occasione di lavorare in contemporanea (e quindi di intersecare ricerche, risultati e prodotti parziali) da un lato sulla strutturazione e marcatura di un dizionario storico come il

¹ Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

² Un prototipo del *GDLI* è attualmente consultabile negli Scaffali digitali del sito web dell'Accademia o direttamente all'indirizzo <<http://www.dgli.it>>.

GDLI nella sua versione informatizzata (mettendo a punto procedure di collazione semi-automatica del testo ottenuto con l'OCR e di marcatura dei campi strutturali identificati; vedi Sassolini *et alii* 2019, Biffi e Sassolini 2020), e dall'altro sulla creazione di lessici computazionali differenziati in diacronia per sopperire finalmente all'indebolimento – anche fino al grado zero – dell'efficacia di strumenti di annotazione linguistica quando ci si allontani dalle condizioni ideali dell'italiano contemporaneo scritto di quelle varietà sostanzialmente riconducibili al campione del *Lessico di frequenza dell'italiano contemporaneo* (LIF). E così determinare un allargamento dell'efficacia degli strumenti di annotazione linguistica in diacronia, in diamesia e in diafasia.

Sul versante degli studi umanistici, gli sviluppi di questi strumenti hanno ricadute notevoli, sia nell'ottica di consegnare agli studiosi metodi di indagine sempre più potenti per le loro ricerche linguistiche, sia in quella, cara a un'istituzione come l'Accademia della Crusca, di raggiungere un'utenza più vasta. La digitalizzazione dei testi e la loro disponibilità in rete – infatti – non sono sufficienti per avvicinare un'utenza non circoscritta agli addetti ai lavori al patrimonio culturale tramandato. Per rendere fruibili da una vasta e variegata utenza i contenuti di istituzioni culturali, è necessario offrire modalità di accesso e navigazione dotate di “intelligenza linguistica”.

Ad oggi, è ampiamente riconosciuto che l'applicazione di metodi e delle tecniche per il Trattamento Automatico della Lingua (TAL) a varietà storiche di una lingua presenta innumerevoli ostacoli, poiché gli strumenti sviluppati per le lingue moderne necessitano di specializzazioni a vario livello (lessicale, morfologico e sintattico) per essere utilizzati con successo nel trattamento di fonti primarie che rappresentano la base degli studi umanistici. Per la lingua italiana, Pennacchiotti e Zanzotto (2008) riportano i risultati di uno studio esplorativo sulle difficoltà derivanti dal trattamento automatico di varietà storiche della lingua italiana basato su un corpus diacronico che raccoglie testi italiani letterari (sia in prosa che in poesia) dal 1200 a

fine Ottocento: le analisi, focalizzate sulla composizione del vocabolario e sull'annotazione morfologica e morfo-sintattica, confermano che il trattamento automatico di varietà storiche della lingua italiana è una sfida aperta. Una delle sfide che il progetto TrA-VaSI intende affrontare. In questa area, l'azione del progetto ruota attorno a due direttrici principali, riguardanti la costruzione di corpora arricchiti con informazione linguistica da usarsi per la valutazione e/o addestramento di strumenti di annotazione, e di lessici computazionali diacronici in grado di supportare il processo di lemmatizzazione del testo. Al momento, l'annotazione linguistica dei testi riguarda il livello morfo-sintattico. Le risorse sviluppate su entrambi i fronti verranno integrate all'interno di infrastrutture di ricerca preesistenti (per esempio CLARIN-IT), al fine di incrementarne la visibilità, l'accessibilità e l'interoperabilità con risorse simili.

Sul versante linguistico-computazionale, la progettazione e costruzione di risorse e strumenti per varietà d'uso della lingua che si discostano in diversa misura dall'italiano contemporaneo scritto su cui gli strumenti di TAL sono tipicamente addestrati costituisce un fertile terreno di sperimentazione per lo sviluppo e il raffinamento di tecnologie innovative di TAL.

L'articolo illustra la metodologia definita per la costruzione delle risorse e i primi risultati raggiunti. La sezione 2 descrive il corpus diacronico selezionato per l'annotazione, e la sezione 3 illustra il metodo seguito per l'annotazione del corpus. La sezione 4 riporta e discute i risultati dei primi esperimenti di annotazione semi-automatica e della strategia di revisione messa a punto. La sezione 5 illustra i risultati di annotazione automatica conseguiti con sistemi esistenti di annotazione per la lingua italiana contemporanea; infine, la sezione 6 delinea gli sviluppi in corso e futuri.

2 Corpus

La costruzione delle risorse si avvarrà in primo luogo dei testi presenti nel corpus per la realizzazione del *Vocabolario Dinamico dell'Italiano Moderno* (VoDIM, Marazzini e Maconi, 2018). Il corpus è frutto di due

progetti nazionali: il PRIN 2012 (*Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni*, diretto da Claudio Marazzini) e il PRIN 2015 (*Vocabolario dinamico dell'italiano post-unitario*, sempre sotto la direzione di Claudio Marazzini); i due progetti, l'uno la prosecuzione dell'altro, hanno coinvolto diverse università (Catania, Firenze, Genova, Milano, Napoli, Piemonte Orientale e Torino) e l'Istituto di Teorie e Tecniche dell'Informazione Giuridica (ITTIG) del CNR di Firenze.

Il *VoDIM* riunisce testi, scritti e orali, attinenti all'italiano post-unitario che riguardano diversi domini dello spazio linguistico dell'italiano: arte, canzone, diritto, economia, gastronomia, poesia, politica, prosa giornalistica, letteraria, paraletteraria e scientifica. Il corpus, diacronicamente bilanciato, ha una estensione di circa 20 milioni di parole; nel prossimo futuro il progetto prevede che il corpus iniziale sia affiancato da un corpus sincronico molto più ampio (circa 2 miliardi di parole), composto da testi ricavati dalla rete (Biffi, 2016, Biffi, 2018, Biffi, 2020, Biffi e Ferrari, 2020). Inoltre, il *VoDIM* è stato di recente inserito all'interno della *Stazione Lessicografica*, consultabile tramite gli *Scaffali digitali* del sito dell'Accademia della Crusca (Biffi 2020: 360-362).³

3 Metodo

La specializzazione di strumenti di annotazione linguistica rispetto a varietà d'uso della lingua diverse da quelle testimoniate nel corpus di addestramento richiede – innanzitutto – la disponibilità di risorse (lessici e corpora annotati) rappresentative della varietà di lingua da trattare. In primo luogo per verificare il livello di accuratezza degli strumenti di annotazione disponibili, e – successivamente – per la loro specializzazione. Il corpus alla base del *VoDIM* si pone come ottima “palestra” da questo punto di vista, in quanto i testi al suo interno si distribuiscono in un vasto arco cronologico (dall'Unità a oggi), presentano una notevole differenziazione diamesica

(scritto, parlato, parlato-scritto e scritto-parlato) e appartengono a un'ampia varietà di generi e di tipologie testuali.

Per l'arricchimento dei testi con annotazioni linguistiche di varia natura ci si avvarrà, dal punto di vista metodologico, dell'esperienza maturata nel corso del progetto *Voci della Grande Guerra* (VGG, De Felice *et al.*, 2018, Lenci *et al.* 2020). In particolare, i metodi adottati per risolvere i problemi di segmentazione delle forme e di lemmatizzazione emersi durante la fase di trattamento automatico (De Felice *et al.*, 2018: 161-162) sono un fondamentale punto di partenza per ottenere risultati efficienti, come nel caso del corpus *VGG*.

Il lavoro è stato articolato nelle seguenti fasi operative:

1. selezione delle fonti a partire dal corpus *VoDIM*;
2. definizione di un sottocorpus rappresentativo delle varietà del *VoDIM*; i testi scelti appartengono a sette dei domini del corpus (arte, cucina, diritto, giornali, letteratura, paraletteratura, scienze) e sono stati bilanciati in diacronia per avere la massima copertura possibile; i campioni di ogni dominio sono tra i 2.600 e i 3.000 token, per un'estensione totale del sottocorpus – ad oggi – di circa 19.000 token;
3. annotazione morfo-sintattica e lemmatizzato del sottocorpus di cui al punto 2. Lo schema di annotazione adottato è quello sviluppato all'interno dell'iniziativa internazionale *Universal Dependencies* (Nivre, 2015), che rappresenta ad oggi uno standard *de facto* per l'annotazione morfo-sintattica e sintattica a dipendenze di testi, incluse varietà storiche di alcune delle lingue trattate. Il sottocorpus selezionato è stato annotato automaticamente con *UDPipe* (Straka e Straková, 2017), addestrato sulla *Italian Universal Dependency Treebank* (IUDT, Bosco *et al.*, 2013). A ciò, ha fatto seguito una fase di revisione manuale degli errori riscontrati nei testi annotati automaticamente.

³ <http://www.stazionelessicografica.it>

Il campione testuale, frutto delle prime fasi di lavoro, è di fatto il primo nucleo di corpus annotato per la validazione e, in prospettiva, l'addestramento e/o specializzazione degli strumenti di trattamento automatico di varietà storiche dell'italiano.

4 Primi risultati

4.1 Analisi quantitativa degli errori

L'analisi degli errori di annotazione morfosintattica e di lemmatizzazione riscontrati nei testi annotati automaticamente con *UDPipe* ha fornito dati importanti sulla mole di errori presenti in ogni testo; nella Tabella 1 sono riportate le percentuali medie di errore nei testi, distribuiti su sei intervalli cronologici che ricomprendono grossomodo la suddivisione in periodi proposta per il *DiaCORIS* (Onelli *et al.*, 2006) e successivamente per il *LIS Lessico Italiano Scritto*⁴, a cui viene aggiunto un sesto periodo comprendente i testi attinenti all'italiano contemporaneo:

	intervalli temporali	n. testi	n. token	% media errori
1	1861-1900	6	4300	8,2%
2	1901-1922	5	3400	8,6%
3	1923-1945	4	3300	5,3%
4	1946-1967	1	600	8,3%
5	1968-2001	7	3800	5,3%
6	2002-oggi	18 ⁵	4100	3,8%

Tabella 1. Percentuale di errori per intervalli cronologici.

Come ci si aspetterebbe, la percentuale di errore tende a scendere, con il passare degli anni, dall'8 al 3%; l'unica eccezione riguarda il periodo 1946-1967, testimoniato però dal campione di circa 600 token estratto da un

solo testo (un saggio di medicina di G. Brotzu, *Ricerche su di un nuovo antibiotico* del 1948); tale valore è dunque poco attendibile – se si considera, tra l'altro, che è il testo del dominio “scienze” con una percentuale di errore tra le più alte (cfr. *infra*).

Più interessanti i dati riportati nella Tabella 2, in cui sono presenti le percentuali medie di errore per ogni dominio:

domini	intervalli temporali		n. testi	n. token	% media errori
arte	2-6	1902-2009 ⁶	4	2600	6,3%
cucina	1-3	1871-1927	4	2700	9,9%
diritto	5-6	2000-2016	17	3000	3,7%
giornali	1-5	1867-1996	4	3000	5,5%
letteratura	1-5	1881-1982	4	2800	8,5%
paraletteratura	1-3	1892-1939	4	2600	5,9%
scienze	1-6	1864-2015	4	2700	5,6%

Tabella 2. Percentuale di errori per domini.

Osservando i dati si nota che le percentuali più alte riguardano i testi gastronomici e letterari, quelle più basse i testi giuridici; gli altri domini si attestano su una media del 5-6%. Una prima spiegazione di questa diffrazione risiede certamente nelle date di pubblicazione dei testi interessati: i libri di gastronomia da cui sono stati estratti i campioni per la cucina sono stati pubblicati tra il 1871 (l'anonimo ricettario *Il cuoco sapiente*) e il 1927 (il famosissimo *Talismano della felicità* di Ada Boni), mentre i testi del dominio “diritto” sono usciti

⁴ Il *LIS* corrisponde alla sezione *DIACORIS* del CILTA di Bologna (<http://corpora.dslo.unibo.it/coris_ita.html>), sviluppata all'interno del portale *Vivit Vivi Italiano. Il portale dell'italiano nel mondo* (<<http://www.viv-it.org/schede/archivi-digitali>>), e costituisce una rielaborazione informatica per omogeneizzare la banca dati *LIT Lessico Italiano Televisivo* e *LIR Lessico Italiano Radiofonico* (anch'essi presenti nel portale) in modo che possano essere interrogate contemporaneamente da un motore (come avviene appunto, affiancando la singola consultazione delle tre banche dati, nella sezione “Archivio Digitale del Vivit”: <<http://www.viv-it.org/schede/archivi-digitali>>). Il *LIS* comprende complessivamente 25

milioni di occorrenze, distribuite equamente su cinque sezioni cronologiche bilanciate (per approfondimenti cfr. Biffi, 2016, pp. 276-77 e nota 24).

⁵ Il numero è così elevato perché comprende per la maggior parte i testi del dominio “diritto”, al cui interno si trovano abstract di saggi e di opere specialistiche; i testi sono perciò molto brevi e, a differenza degli altri domini, è stato necessario attingere a molti più campioni testuali per raggiungere la quota prestabilita di token (cfr. Tabella 2).

⁶ Le date presenti nella colonna si riferiscono alla pubblicazione rispettivamente del primo e dell'ultimo testo del dominio.

tra il 2000 e il 2016. Tuttavia, va considerato che i testi paraletterari, con percentuali di errore nella media, sono stati anch'essi pubblicati tra la fine dell'Ottocento e l'inizio del Novecento; e che, d'altro canto, le opere di letteratura hanno valori elevati, nonostante la presenza di romanzi abbastanza recenti come *Se non ora, quando?* di Primo Levi (1982). Probabilmente lo scarto elevato dalla media è da ricondursi, per la cucina, alle particolari tipologie testuali (come ad esempio le ricette) “nuove” nel panorama dei campioni su cui sono stati testati finora gli strumenti di annotazione. Analoghe considerazioni valgono per la letteratura. Sebbene i corpora letterari siano stati oggetto di analisi automatiche (es. lemmatizzazione) finalizzate all'interrogazione, secondo la prospettiva adottata in questo studio essi appaiono caratterizzati da tratti morfologici quantitativamente significativi su cui gli analizzatori finora utilizzati non sono adeguatamente addestrati (es. forme verbali di prima e seconda persona).

4.2 Analisi qualitativa degli errori

Sembrerebbe essere molto forte, dunque, il legame tra errore, dominio e, soprattutto, genere testuale⁷. Per esempio, i ricettari, così come la manualistica in generale, sono costellati da verbi all'imperativo di seconda plurale (per esempio *mescolate*) che *UDPipe* tende ad assimilare con le forme del participio passato; i testi letterari, dal canto loro, sono colmi di parti dialogiche che presentano al loro interno una serie di tratti, in primo luogo interpunzioni enfatiche (punti “misti” come l'accumulo di punto esclamativo e puntini di sospensione) e interiezioni, che mettono in difficoltà l'analizzatore, sia nell'annotazione morfo-sintattica, sia nella tokenizzazione; analogamente, i testi giuridici e scientifici presentano diversi errori legati alla corposa presenza di abbreviazioni, sigle e simboli.

È stata inoltre eseguita una valutazione sulle tipologie di errore, che per circa un terzo

dei casi – oltre il 27% – coinvolgono soltanto il lemma, e sono senz'altro dovuti al fatto che *UDPipe*, nella versione di base, utilizza un dizionario costruito a partire dal corpus di addestramento integrato da euristiche di analisi morfologica utilizzate per trattare forme sconosciute (anch'esse automaticamente derivate dal corpus di addestramento). Seguono due esempi di lemmatizzazione basata su tali euristiche:

27 lavano **lare** VERB V Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin _
43 illustri **illustro** ADJ A Gender=Masc|Number=Plur _

In altri casi – circa il 20% – gli errori riguardano l'assegnazione della categoria grammaticale (sia *coarse-grained* sia *fine-grained*); tipici esempi sono il *che* pronome a cui viene attribuito il valore di congiunzione (a), e lo scambio tra aggettivo e sostantivo nella coppia nominale (b):

a.
24 Pagliarini Pagliarini PROPN SP _
25 che che **SCONJ** CS _
26 per per ADP E _
27-28 farsi _
27 far fare VERB V VerbForm=Inf _
28 si si PRON PC Clitic=Yes|Person=3|PronType=Prs _
29 ascoltare ascoltare VERB V VerbForm=Inf _
30 ha avere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _
31 parlato parlare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part _
32 in in ADP E _
33 piedi piede NOUN S Gender=Masc|Number=Plur _ SpaceAfter=No
34 . PUNCT FS _

b.
47 massimi massimo **NOUN** S Gender=Masc|Number=Plur
48 esponenti esponente **ADJ** A Number=Plur _

Circa il 6% degli errori concerne invece solo le caratteristiche morfologiche associate alla forma (*Universal Dependencies features*); appartengono a questa tipologia gli scambi sopra menzionati tra imperativo e participio:

⁷ A tal proposito, si veda il progetto EvaLatin 2020 (Sprugnoli *et al.*, 2020), che mira a implementare lo studio della portabilità degli strumenti NLP per il latino attraverso la costituzione di tre *baseline* a cui appartengono diversi generi e diversi periodi cronologici;

analogamente a quanto si osserva nel presente elaborato, i primi risultati di EvaLatin 2020 hanno mostrato l'impatto sia delle caratteristiche diacroniche sia di genere testuale sull'accuratezza dell'annotazione.

3 digrassate digrassare VERB V Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part _

Vi sono poi diversi casi – circa il 7% – di errata segmentazione dei token molto simili a quelli riscontrati nel Corpus *VGG* (De Felice *et al.*, 2018: 161), ossia il mancato riconoscimento di forme verbali enclitiche rare o desuete nell’italiano contemporaneo:

13 intendesi intendese ADJ A Number=Plur _
19 siasi siasi ADJ A Number=Sing _

Altri errori di iposegmentazione si osservano in massima parte nelle sequenze preposizionali diacronicamente marcate del tipo *pella*, *collo* ecc.:

7 co’ coco DET RD Definite=Def|Number=Sing|PronType=Art _
4 colla collare VERB V Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _
7 pei peo ADJ A Gender=Masc|Number=Plur _

Analogamente, *UDPipe* tende a non riconoscere altre combinazioni di clitici:

27 spogliatela spogliatelare NOUN S Gender=Fem|Number=Sing _
8 toglietene toglietena SCONJ CS _

I restanti errori riguardano tutte le altre possibili combinazioni presenti nella catena di annotazione:

1. lemma + categoria grammaticale + tratti morfo-sintattici (21% circa): l’annotazione automatica risulta essere totalmente sbagliata; errori di questo tipo si osservano comunemente quando, in presenza di una iniziale maiuscola, la forma viene assimilata a un nome proprio:

3 Pigliate pigliate PROPN SP _

2. categoria grammaticale + tratti morfo-sintattici (20% circa): la forma viene correttamente lemmatizzata, ma la categoria grammaticale e i tratti associati sono errati; un caso tipico, al contrario di quanto osservato sopra, è il *che* congiunzione a cui viene attribuito il valore di pronome:

1 Avuta avere ADJ A Gender=Fem|Number=Sing _
2 notizia notizia NOUN S Gender=Fem|Number=Sing _

3 che che PRON PR PronType=Rel _
4 era essere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin _
5 già già ADV B _
6 cominciato cominciare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part _

3. lemma + tratti morfo-sintattici (3% circa): soltanto la categoria grammaticale è corretta:

20 investe **investa** VERB V Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part _

La revisione ha comportato non soltanto la correzione meccanica degli errori, ma anche l’adattamento di talune annotazioni ai criteri definiti per la *IUDT treebank*, al fine di garantire la compatibilità delle annotazioni dei diversi UD corpora presenti per la lingua italiana. La selezione di tali criteri è stata una delle imprese più ardue: si è optato per una strategia conservativa e un continuo confronto con i corpora di UD per ottenere il minor rumore possibile e una massima sistematizzazione del *tagset*, soprattutto per quanto concerne avverbi e congiunzioni: per esempio *come*, nei corpora UD consultati, viene sempre classificato funzionalmente con il valore di preposizione di fronte ai sintagmi normali, di congiunzione in presenza di un’altra congiunzione (*come se*), di avverbio in tutti gli altri contesti; i casi in cui sono state introdotte delle innovazioni hanno per la maggior parte riguardato incoerenze nell’annotazione: *po’*, generalmente avverbio, è stato taggato come nome in contesti quali *un po’ di pane*, poiché la forma piena *poco* veniva già riconosciuta con lo stesso valore nei medesimi contesti; analogamente, gli infiniti sostantivati, a volte riconosciuti a volte no, sono stati etichettati sempre come nome.

5 Baseline di confronto

La scelta di *UDPipe* come strumento per la preannotazione del corpus (cfr. sezione 3) ha molteplici motivazioni, che vanno dallo schema di annotazione adottato (UD) alla possibilità di riaddestramento, che ne fanno uno strumento adeguato per la costruzione dei corpora annotati sviluppati nel progetto TrA-VaSI.

In vista del riaddestramento della catena di analisi per il trattamento di varietà storiche di italiano, i risultati ottenuti con *UDPipe* sono stati confrontati con l'output di strumenti di annotazione morfo-sintattica e lemmatizzazione sviluppati in ambito DH (per esempio l'annotatore morfologico del PiSystem, Picchi, 2003) o già testati all'interno di applicazioni di DH come LinguA (Attardi e Dell'Orletta, 2009, Attardi *et al.*, 2009, Dell'Orletta, 2009). I dati emersi mostrano che la baseline costituita dall'annotatore PiMorfo, sviluppato primariamente in funzione dell'interrogazione di vasti archivi testuali, presenta una percentuale media di errore assai più elevata di *UDPipe* – oltre il 9% –, ma è stato riscontrato che una parte sostanziale – all'incirca il 13% – riguarda alcuni errori sistematici, ad esempio il mancato riconoscimento della preposizione con vocale elisa *d'* o del valore di forme ambigue quali *ancora*, che in tutte le occorrenze in funzione di avverbio/congiunzione è stato confuso con la terza persona del presente indicativo di *ancorare*.

Per quanto riguarda LinguA, si è rilevato invece un indice di errore indubbiamente più basso rispetto a *UDPipe* – con una media intorno al 4% – e oltre il 40% degli errori riguarda soltanto sei tipologie, tra cui spicca l'assegnazione scorretta del tag SP (nome proprio, cfr. sezione 4.2), che viene associato a quasi tutte le occorrenze delle forme con l'iniziale maiuscola.

6 Conclusioni e sviluppi in corso

Abbiamo illustrato brevemente i passaggi che il progetto TrAVaSI sta seguendo per costruire risorse per il trattamento automatico di varietà storiche di italiano, differenziate anche in diamesia, diafasia e a livello di genere testuale. In questa fase preliminare abbiamo cercato di dare risposte a quesiti irrisolti sull'annotazione e la lemmatizzazione dell'italiano in diacronia, per esempio sul modo di trattare alcune forme (voci diverbate, infiniti sostantivati, varianti regionali, per citare alcuni esempi) e sui criteri di lemmatizzazione delle varianti fono-morfologiche dello stesso lemma. Le risorse sviluppate saranno sfruttate per testare strumenti esistenti

di annotazione morfo-sintattica e di lemmatizzazione e per lo sviluppo e/o specializzazione di componenti software per il trattamento di varietà storiche della lingua italiana.

I primi risultati raccolti a partire dall'analisi del sottocorpus del *VoDIM* selezionato mostrano chiaramente che le dimensioni di variazione da tenere in considerazione sono molteplici e fortemente interrelate, derivanti da diversi tipi di processi, come la variazione nel tempo (variazione diacronica), la variazione correlata a variabili sociolinguistiche oppure legata al genere testuale o allo stile di chi scrive. I luoghi di variazione spaziano dall'ortografia, alla morfologia e alla sintassi (specialmente in diacronia e in testi di domini specialistici). Questo tipo di analisi è fondamentale per arrivare a definire una metodologia per la creazione di ulteriori risorse, con il fine di allargare progressivamente l'arco cronologico ma anche la tipologia di varietà d'uso della lingua per poter costruire strumenti TAL che siano applicabili a testi italiani dei secoli precedenti, riconducibili a diverse varietà d'uso della lingua. Da questa prospettiva, il corpus selezionato come punto di partenza del progetto TrAVaSI diventa quindi particolarmente significativo in quanto crea i presupposti per la definizione di un metodo per la specializzazione di strumenti di annotazione in relazione a molteplici varietà linguistiche, diacroniche ma anche diafasiche, diastratiche o corrispondenti a tipologie testuali.

Gli sviluppi in corso includono:

1. l'addestramento e la specializzazione dei modelli di *UDPipe* mediante test set costituiti dai domini del test corpus, su cui di volta in volta verrà valutato l'incremento di accuratezza dell'annotazione;
2. la costruzione di lessici diacronici a partire da corpora annotati – iniziando dal *VoDIM* e successivamente attingendo a risorse cronologicamente antecedenti – come base per il processo di lemmatizzazione;
3. sulla base della distanza tra il corpus utilizzato per l'addestramento e i testi da analizzare, l'identificazione del modello linguistico più adeguato da utilizzarsi per l'annotazione linguistica.

Ringraziamenti

Le attività di ricerca illustrate in questo articolo sono condotte nell'ambito del progetto *TRaVaSI* del programma di intervento denominato "CNR4C" di cui al progetto congiunto di alta formazione, cofinanziato dalla Regione Toscana con le risorse del POR FSE 2014-2020 – Asse A Occupazione, all'interno di "GiovaniSi", il progetto regionale Toscano per l'autonomia dei giovani.

Bibliografia

- Attardi G., Dell'Orletta F. (2009), *Reverse Revision and Linear Tree Combination for Dependency Parsing*, in *NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (Boulder, Colorado, June 2009). Proceedings, Association for Computational Linguistics, pp. 261-264.
- Attardi G., Dell'Orletta F., Simi M., Turian J. (2009), *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*, Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italia, dicembre 2009).
- Biffi M. (2016), *Progettare il corpus per il vocabolario postunitario*, in *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*, Atti della "Piazza delle Lingue" dell'Accademia della Crusca, edizione 2014 (Firenze, 6-8 novembre 2014), a cura di Claudio Marazzini e Ludovica Maconi, Firenze, Accademia della Crusca, pp. 259-80.
- Biffi M. (2018), *Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano post-unitario*, JADT'18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data, a cura di Domenica Fioredistella Iezzi, Livia Celardo e Michelangelo Misuraca, Roma, Universitalia, 2018, vol. 1, pp. 99-107.
- Biffi M. (2020), *La galassia lessicografica della Crusca in rete, in Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale*. Atti del convegno internazionale in occasione delle 40.000 voci del TLIO, Firenze, 13-14 settembre 2018, a cura di Lino Leonardi e Paolo Squillaciotti, Alessandria, Edizioni dell'Orso, pp. 219-232.
- Biffi M., Ferrari A. (2020), *Progettare e ideare un corpus dell'italiano nella rete: il caso del CoLIWeb*, «Studi di Lessicografia Italiana», vol. XXXVII, 2020, pp. 357-374.
- Biffi M., Sassolini E., *Strategie e metodi per il recupero di dizionari storici*, in *La svolta inevitabile: sfide e prospettive per l'informatica umanistica*, Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), Milano, Università Cattolica del Sacro Cuore, 15-17 gennaio 2020, a cura di Cristina Marras, Marco Passarotti, Greta Franzini ed Eleonora Litta, dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, 2020, pp. 235-239 (pubblicazione elettronica in «Quaderni di Umanistica Digitale»: <<http://doi.org/10.6092/unibo/amsacta/6316>>).
- Bosco C., Montemagni S., Simi M. (2013). *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank*, Proceedings of the "7th Linguistic Annotation Workshop & Interoperability with Discourse", Sofia, Bulgaria, August 8-9, 2013, ACL, pp. 61-69.
- De Felice I., Dell'Orletta F., Venturi F., Lenci A. Montemagni S. (2018), *Italian in the Trenches: Linguistic Annotation and Analysis of Text of the Great War*, Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it), 10-12 dicembre 2018, Torino.
- Dell'Orletta F. (2009), *Ensemble system for Part-of-Speech tagging*, Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009).
- GDLI = *Grande dizionario della lingua italiana*, di Salvatore Battaglia (poi diretto da Giorgio Bàrberi Squarotti), Torino, UTET, 1961-2002, 21 voll.; con *Supplemento 2004* e *Supplemento 2009*, diretti da Edoardo Sanguineti, Torino, UTET, 2004 e 2008, e *Indice degli autori citati nei volumi I-XXI e nel Supplemento 2004*, a cura di Giovanni Ronco, Torino, UTET, 2004.
- Lenci A., Montemagni S., Boschetti F., De Felice I., dei Rossi F., Dell'Orletta F., Di Giorgio M., Miliani M., Passaro L. C., Puddu A., Venturi G., Labanca N. (2020), *Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War*, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marsiglia, 11-16 maggio 2020, pp. 911-918.
- LIF = Bortolini U., Tagliavini C., Zampolli A., *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia 1971 (poi: Milano, Garzanti, 1972).
- Marazzini C., Maconi L. (2018), *Il Vocabolario dinamico dell'italiano moderno rispetto ai linguaggi settoriali. Proposta di voce lessicografica per il redigendo VoDIM*, «Italiano digitale», VII/4, pp. 101-20.
- Nivre J. (2015), *Towards a Universal Grammar for Natural Language Processing*, Computational Linguistics and Intelligent Text Processing -Proceedings of the 16th International Conference, CICLing 2015, Part I., Cairo, Egitto, pp. 3-16.
- Pennacchiotti M., Zanzotto F.M. (2008). *Natural Language Processing Across Time: An Empirical Investigation on Italian*, Proceedings of GoTAL - 6th International Conference on Natural Language Processing, LNAI, volume 5221, pp. 371-382.

- Picchi E. (2003), *PiSystem: sistemi integrati per l'analisi testuale*, in *Computational Linguistics in Pisa. Linguistica Computazionale*, a cura di Zampolli A., Calzolari N., Cignoni L., Special Issue, XVIII-XIX, Pisa-Roma, IEPI, 2003, pp. 597-627.
- Sassolini E., Fahad Khan A., Biffi M., Monachini M., Montemagni S., *Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study*, in Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pe-reira, I., Kallas, J., Jakubiček, M., Krek, S. & Tiberius, C. (eds.), *Electronic lexicography in the 21st century: Smart lexicography*. Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra, Portugal), Brno, Lexical Computing CZ, s.r.o., 2019, pp. 603-621 (pubblicazione elettronica: <https://elex.link/elex2019/proceedings-download/>).
- Sprugnoli R., Passarotti M., Cecchini F. M., Pellegrini M. (2020), *Overview of the EvaLatin 2020 Evaluation Campaign*, Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, LREC 2020, Marsiglia, Francia, pp. 105-110.
- Straka M., Straková J. (2017), *Tokenizing, POS Tagging, Lemmatizing and Parsing UD2.0 with UDPipe*, Proceedings of the CoNLL 2017: Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, pp.88-99.

AriEmozione: Identifying Emotions in Opera Verses

Francesco Fernicola¹, Shibingfeng Zhang¹, Federico Garcea¹
Paolo Bonora², and Alberto Barrón-Cedeño¹

¹Department of Interpreting and Translation
Università di Bologna, Forlì, Italy

²Department of Classical Philology and Italian Studies
Università di Bologna, Bologna, Italy

{francesco.fernicola, zhang.shibingfeng}@studio.unibo.it
{federico.garcea2, paolo.bonora, a.barron}@unibo.it

Abstract

We present a new task: the identification of the emotions transmitted in Italian opera arias at the verse level. This is a relevant problem for the organization of the vast *repertoire* of Italian Opera arias available and to enable further analyses by both musicologists and the lay public.

We shape the task as a multi-class supervised problem, considering six emotions: love, joy, admiration, anger, sadness, and fear. In order to address it, we manually-annotated an opera corpus with 2.5k verses—which we release to the research community—and experimented with different classification models and representations. Our best-performing models reach macro-averaged F₁ measures of ~0.45, always considering character 3-grams representations. Such performance reflects the difficulty of the task at hand, partially caused by the size and nature of the corpus, which consists of relatively short verses written in 18th-century Italian.

1 Introduction

Opera lyrics have the function of expressing the emotional state of the singing character. In 17th- and 18th-century operas, characters brought on stage passions induced in their souls by the succession of events in the drama. Musicological studies use these affects as one of the interpretative keys of the work as a

whole (Zoppelli, 2001; McClary, 2012). Being able to automatically identify the emotions expressed by the different arias of each work would provide scholars with a useful tool for a systematic study of the *repertoire*. The technology to identify the emotion(s) expressed by an aria represents an effective tool to study the vast repertoire of arias and characters of this period for musicologists and the lay public alike. As an aria may express more than one emotion, we go one granularity level lower—at the verse level. The task is defined as follows:

Identify the emotion expressed in a verse, in the context of an aria.

In order to do that we created the AriEmozione 1.0 corpus: a collection of 678 operas with 2.5k verses, each of which has been manually annotated with respect to emotion. We experimented with different supervised models (e.g., SVMs, neural networks) and text (e.g., character *n*-grams and distributed representations).

Our experiments show that, regardless of the model, character 3-grams outperform all other representations, reaching weighted macro-averaged F₁ measures of ~0.45. Under-represented classes (e.g., **fear**) are the hardest to identify. Others, such as **anger** and **sadness**, being both negative, are often confused between each other.

The rest of the contribution is distributed as follows. Section 2 describes the AriEmozione 1.0 corpus. Section 3 describes the explored models and representations. Section 4 discusses the experiments and obtained results. Section 5 overviews some related work. Section 6 closes with conclusions and proposals for future work.

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

First of all, thank you for helping with this work. We are a group of researchers from the D. of Classical Philology and Italian Studies and the D. of Interpreting and Translation, both at UniBO. Your work will help us to produce artificial intelligence models to analyse the lyrics in music.

At this stage we are focused on opera. You will annotate arie in Italian from diverse periods, looking for the emotions that they express. Your work consists of identifying the emotion expressed in each of the verses composing an aria. You can choose among six emotions (or none of them), which are defined next: [...]

Each row is divided in six columns:

id A unique id, tied to the verse. Do not modify it.

verse A verse, inside of an aria. This is the text that you are going to analyse.

emotion Here you can select the expressed emotion (or none of them)

emotion sec. This is available to choose a secondary emotion, in case it is really difficult to choose just one

confidence Not being 100% sure is ok. If that is the case, please let us know by choosing the right confidence level (default: "I am sure").

comments Feel free to tell us something about this instance, if you feel like.

Figure 1: Instructions given to the annotators of the emotions in the AriEmozione 1.0 corpus.

2 The AriEmozione 1.0 Corpus

The corpus AriEmozione 1.0 is a subset of the materials collected by project CORAGO.¹ AriEmozione 1.0 contains a selection of 678 operas composed between 1655 and 1765. We consider the lyrical text in the arias only. A. Zeno and P. Metastasio are among the most represented librettists in the corpus (~ 30% of the operas); they are two of the most representative and prolific librettists of the 18th century. All texts are written in the 18th century Italian and articulated in verses and *stanze*.

We labeled the emotions transmitted by every single verse, as we observed that this is the right granularity to obtain full text snippets expressing one single emotion. René Descartes wrote in 1649 "Les passions de l'âme", a sort of compendium of all possible emotions and their possible causes (Garavaglia, 2018). For the sake of concreteness, we leveraged Parrott's (2001) tree of emotions classification.

¹CORAGO is the Repertoire and archive of Italian opera librettos. It constitutes the first implementation of the RADAMES prototype (Repertorizzazione e Archiviazione di Documenti Attinenti al Melodramma E allo Spettacolo) (Pompilio et al., 2005); <http://corago.unibo.it>.

	amore	gioia	ammirazione	rabbia	tristezza	paura	nessuna	total
train	289	274	289	414	503	166	38	1,973
dev	36	31	23	84	61	12	3	250
test	37	39	30	64	54	15	11	250
overall	362	344	342	562	618	193	52	2,473

Table 1: AriEmozione 1.0 corpus statistics.

The first level of such tree includes six primary emotions: love, joy, surprise, anger, sadness, and fear. Based on the nature of the material under review, we substitute surprise with admiration, ending with the following six classes:

Amore (love) incl. affection, lust, longing.

Gioia (joy) incl. cheerfulness, zest, contentment, pride, optimism, enthrallment, relief.

Ammirazione (admiration) admiration or adoration of someone's talent, skill, or other physical or mental qualities.

Rabbia (anger) incl. irritability, exasperation, rage, disgust, envy, torment.

Tristezza (sadness) incl. suffering, disappointment, shame, neglect, sympathy.

Paura (fear) incl. horror and nervousness.

An extra class **nessuna (none)** applies mostly to verses with non-actionable words only, neglected in the current experiments.

Two native speakers of Italian annotated all 2,473 instances independently considering the instructions displayed in Figure 1. They were asked to include (i) the emotion transmitted by the verse, (ii) an optional secondary label (in case they perceived a second emotion), and (iii) their level of confidence: total confidence, partial confidence, or very doubtful.

We measured the Cohen's kappa inter-annotator agreement (Fleiss et al., 1969) at this stage on the primary emotion. The result was 32.30, which is considered as a fair agreement. This value results from the perfect matching between the two annotators in 44% of the instances. When considering the secondary emotion as well, the two annotators coincided in 68% of the instances. These numbers reflect the complexity of the task. The same annotators gathered together to discuss and consolidate all dubious instances. Table 1 shows the number of instances per class for each corpus partition: training, development, and test set. The verse average length

id	verse	class
ZAP1593570_03	Non ho più lagrime; non ho più voce; non posso piangere; non so parlar I have no more tears; I have no more voice; I cannot cry; I don't know how to speak	Tristezza
ZAP1596431_00	Barbaro! Oh dio mi vedi divisa dal mio ben; barbaro, e non concedi ch'io ne dimandi almen Barbarian! Oh Lord, you see me separated from my own good; barbarian, you don't even allow me but one demand	Rabbia
ZAP1593766_01	Guardami e tutto obbligo e a vendicarti io volo; di quello sguardo solo io mi ricorderò Look at me, all else is forgotten and I haste to avenge you; only I shall remember that gaze	Amore
ZAP1594229_00	Su la pendice alpina dura la quercia antica e la stagion nemica per lei fatal non è; Up on the slope of the mountain the ancient oak tree still lives on, and the adverse season poses no fatal threat	Ammirazione
ZAP1596807_00	In questa selva oscura entrai poc'anzi ardito; or nel cammin smarrito timido errando io vo I entered this dark forest not too long ago, boldly; having now lost the path I wander around, shyly	Paura
ZAP1599979_01	Vede alfin l'amate sponde, vede il porto, e conforto prende allor di riposar Finally, the beloved shores, the harbor, are all in sight and with them come solace and sleep	Gioia

Table 2: Instances from the AriEmozione 1.0 corpus, including unique identifier, verse in Italian and **English translation**, and class. We include free (unofficial) translations for clarity.

is 72.5 ± 31.6 characters and the corpus contains 34,608 (4,458) tokens (types).²

Table 2 shows examples of verses in the corpus, including one of each of the six emotions.

3 Models and Representations

The nature of the corpus—a small amount of short verses written in 18th-century Italian—led us to select a humble set of models and representation alternatives. The baseline is a k -Nearest Neighbors algorithm (kNN), considered thanks to its success in classification tasks (Zhang and Zhou, 2007). We also experiment with multi-class SVMs, logistic regression, and neural networks. Regarding the latter, we experiment with a number of architectures with two and three hidden layers. Finally, we experiment with a FastText classifier (Joulin et al., 2017). Table 3 summarizes the explored configurations.³

²The corpus is available at <https://zenodo.org/record/4022318>.

³The code is available at <https://github.com/TinfFoil/AriEmozione>. We used Sklearn for the kNN, SVM, and logistic regression models; Keras for the neural networks, and the Facebook-provided library for FastText (cf. <https://scikit-learn.org>, <https://keras.io/>, and <https://github.com/facebookresearch/fastText>).

Model	Settings
k -NN	L2-Norm exploring with $k \in [1, \dots, 9]$.
SVM	RBF; both explored with $c \in [1, 10, 100, 1000]$ and $\gamma \in [1e-3, 1e-4]$.
Log Reg	Multinomial Logistic Regression with Newton-CG solver.
NN	2 (3) hidden layers with size $\in [32, 64, 96, 128, 256]$ ($\in [8, 16, 32, 64, 96]$); 20% dropout; ReLu for input/hidden layers; softmax for output layer; categorical cross-entropy loss function; Adam; epochs $\in [1, \dots, 15]$
FastText	300d embeddings with or without pre-training; learning rate $\in [0.3, 0.6, 1]$; epochs $\in [1, 3, 5, 10, \dots, 100]$

Table 3: Experimental settings overview.

As for the text representations, we consider TF-IDF vectors of both character 3-grams and word 1-grams (no higher n values are considered due to the corpus dimensions). For pre-processing, we employ the spacy Italian tokenizer⁴ and casefold the texts. We also explore with dense representations, derived from the TF-IDF vectors, by means of both LDA (Hoffman et al., 2010) and LSA (Halko et al., 2011).

⁴<https://spacy.io/models/it>

model representation	10-fold CV		test	
	F ₁	Acc	F ₁	Acc
kNN				
char 3-grams	0.38	38.51	0.35	35.15
words	0.36	36.08	0.35	34.73
LDA char	0.30	29.97	0.31	30.54
SVM-RBF				
char 3-grams	0.44	43.70	0.43	43.00
words	0.42	42.00	0.44	44.00
LDA char	0.28	28.00	0.30	30.00
Log reg				
char 3-grams	0.44	45.57	0.42	43.10
words	0.41	43.20	0.41	43.10
LDA char	0.28	30.63	0.29	30.96
2-layers NN				
char 3-grams	0.42	43.61	0.47	46.86
words	0.42	42.91	0.43	43.10
LDA char	0.27	29.56	0.27	31.80
3-layers NN				
char 3-grams	0.49	41.86	0.40	41.84
words	0.47	42.60	0.40	41.84
LDA char	0.26	31.41	0.30	31.80
FastText				
char 3-grams	0.43	45.00	0.41	42.37
pre-trained chars	0.43	47.00	0.41	41.00
words	0.42	42.56	0.39	44.07
pre-trained words	0.38	41.00	0.40	42.00

Table 4: F₁ and accuracy on cross-validation held-out test for some of the model/representation combinations.

In both cases, we target reductions to 16, 32, and 64 dimensions. As for embeddings, we adopted the pre-trained 300-dimensional Italian vectors of FastText (Joulin et al., 2017), and tried with character 3-grams and words.

4 Experiments

We conducted several experiments to find the best combination of parameters and representations. Given the amount of instances available, we merged the training and development partitions and performed 10-fold cross validation. As standard, the test partition was left aside and only one prediction was carried out on it, after identifying the best configurations.

We evaluate our models on the basis of accuracy and weighted macro-averaged F₁ measure to account for the class imbalance. Table 4 shows the results obtained with some interesting configurations and representations both for the cross-validation and on the test set.⁵ Character and word n -grams TF-IDF, LSA, and LDA were tested with all models except

⁵The full batch of results is available at <https://docs.google.com/spreadsheets/d/1Ztjry2mJs6ufCZM105CQRyZ8pA5YDnT0N0hONGX1nW0/edit?usp=sharing>

	ammirazione	amore	gioia	paura	rabbia	tristezza
ammirazione	0.37	0.03	0.18	0.07	0.11	0.06
amore	0.03	0.43	0.13	0.00	0.09	0.17
gioia	0.27	0.16	0.31	0.20	0.09	0.07
paura	0.10	0.03	0.00	0.40	0.02	0.07
rabbia	0.20	0.14	0.03	0.13	0.64	0.17
tristezza	0.17	0.14	0.13	0.07	0.19	0.48

Table 5: Confusion matrix for the 2-layers neural network with TF-IDF character 3-grams.

for FastText, on which we test with and without pre-trained embeddings. Notice that we are not interested in combining features, but in observing their performance in isolation.

The most promising representation on cross-validation appears to be the simple character 3-grams, with which we obtained the best results across all models; although it also features the highest variability across folds. Among all 3-gram derived representations, LDA consistently obtained the worst results across all models. Still, it is more stable across folds than the sparse 3-gram representation. As for fastText, with the same epoch number and learning rate, the character 3-gram vectors always achieved much higher accuracy than the word vectors.

Similar patterns are observed when projecting to the unseen test set. The character 3-grams in general hold the best performance, while the 3-gram LDA tends to remain the worst in spite of the model used. This behavior does not hold in all cases. For instance, the logistic regression model achieves F₁ = 0.44 on cross-validation, but drops to 0.42 on test. This might be the result of over-fitting.

It is worth noting that all models tend to confuse **rabbia** and **tristezza**. Table 5 shows the confusion matrix for the best model on test. These two emotions get confused between each other on an average of 18% of the cases. The classifiers tend to confuse **ammirazione** for **gioia** as well, which is understandable given their semantic closeness.

5 Related Work

Building on the numerous pre-existing studies focusing on sentiment analysis (Ain et al., 2017; Shi et al., 2019), some researchers have

been seeking to dig deeper, towards multi-class emotion analysis. Most of the work thus far has focused on social media (e.g. Twitter). Bouazizi and Ohtsuki (2016) built a classifier for seven emotions: **happiness**, **sadness**, **anger**, **love**, **hate**, **sarcasm** and **neutral**; i.e. an overlap of five classes with respect to the ones in ariEmozione. In contrast to our experiments, they focused on exploiting the polarity of the words from each instance to be fed to a random forest classifier.

Balabantaray et al. (2012) tried to distinguish among **happy**, **sad**, **anger**, **disgust**, **fear** and **surprise** using WordNet Affect (Valitutti et al., 2004). Given that no Word-net-Affect is currently available for Italian, such an approach is unfeasible.

Promising work has been carried out on news articles (Ye et al., 2012), news headlines (Strapparava and Mihalcea, 2007) and children’s narrative (Alm et al., 2005). While a lexical-based approach is the most frequent to determine the binary positive vs negative classification, Strapparava and Mihalcea (2007) combined a high-dimensional word space produced from word TF-IDF vectors with a set of seed words to predict the valence of a text exploiting the syntagmatic relations between words. A bottom-up semantic approach has also been proposed (Seal et al., 2020).

To the best of our knowledge, no work in the field of either emotion or sentiment analysis has been performed on operas.

6 Conclusions and Future Work

We addressed the novel problem of emotion classification of opera arias at the verse level. The task is interesting because of the lack of automated tools for the analysis of operas and challenging due to both the language used in 17th- and 18th-century lyrics and the complexity to produce the necessary amount of quality supervised data.

We explored with various classification models and representations. A neural network with two hidden layers fed with a simple TF-IDF character 3-gram representation is among the most promising approaches to the problem. Among the six possible emotions, the most difficult to identify are **rabbia** and

tristezza, which tend to be confused with each other, followed by **ammirazione**, which is often confused by **gioia**. In order to foster the research on this topic, we release the AriEmozione 1.0 corpus to the community (cf. footnote 2).

As for the future work, we intend to increase the size of the AriEmozione 1.0 corpus by means of active learning (Yang et al., 2009). Once a larger data volume is produced, we plan to explore with models to identify the emotion at the aria rather than at the verse level. Following the theory of emotion proposed by Plutchik (1980), we could identify the emotion of a whole aria by combining the emotions at the verse level, and then conduct experiments to verify which granularity is more adequate as a single emotion unit. In order to address the issue of emotional polysemy and ambiguity of aria verses, we aim at producing explainable models by highlighting the specific fragments expressing the emotion.

Another interesting alternative is the one highlighted by Zhao and Ma (2019), who adopted an efficient meta-learning approach to augment the learning ability of emotion distribution; i.e. the intensity values of a set of emotions within a single sentence, when the training dataset is small, as in the AriEmozione 1.0 corpus.

Acknowledgments

This research is carried out in the framework of CRICC: *Centro di Ricerca per l’interazione con le Industrie Culturali e Creative dell’Università di Bologna*; a POR-FESR 2014-2020 Regione Emilia-Romagna project (<https://site.unibo.it/cricc>).

We thank Ilaria Gozzi and Marco Schillaci, students at Università di Bologna, for their support in the manual annotation of the AriEmozione 1.0 corpus.

References

- Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noreen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, 8(6):424.
- Cecilia O. Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning

- for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 579–586.
- Rakesh C. Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53.
- Mondher Bouazizi and Tomoaki Ohtsuki. 2016. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter. In *IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.
- Joseph L. Fleiss, Jacob Cohen, and B.S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.
- Andrea Garavaglia. 2018. Funzioni espressive dell’aria a metà seicento secondo il “Giasone” di Cicognini e Cavalli. *Il Saggiatore Musicale*, Anno XXV(1):5–31.
- Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.
- Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. ACL, April.
- Susan McClary. 2012. *Desire and Pleasure in Seventeenth-Century Music*. University of California Press, Berkeley, CA, 1 edition.
- W. Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.
- Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.
- Angelo Pompilio, Lorenzo Bianconi, Fabio Regazzi, and Paolo Bonora. 2005. RADAMES: A new management approach to opera: Repertory, archives and related documents. In *Proceedings - First International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*. IEEE.
- Dibyendu Seal, Uttam K. Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In Milan Tuba, Shyam Akashe, and Amit Joshi, editors, *Information and Communication Technology for Sustainable Development*, pages 423–430, Singapore. Springer Singapore.
- Yong Shi, Luyao Zhu, Wei Li, Kun Guo, and Yuanchun Zheng. 2019. Survey on classic and latest textual sentiment analysis articles and techniques. *International Journal of Information Technology & Decision Making*, 18(04):1243–1287.
- Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.
- Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1).
- Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’09*, page 917–925, New York, NY. Association for Computing Machinery.
- Lu Ye, Rui-Feng Xu, and Jun Xu. 2012. Emotion prediction of news articles from reader’s perspective based on multi-label classification. In *2012 international conference on machine learning and cybernetics*, volume 5, pages 2019–2024. IEEE.
- Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.
- Zhenjie Zhao and Xiaojuan Ma. 2019. Text emotion distribution learning from small sample: A meta-learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3957–3967, Hong Kong, China, November. Association for Computational Linguistics.
- Luca Zoppelli. 2001. Il teatro dell’umane passioni: note sull’antropologia dell’aria secentesca. In *I luoghi dell’immaginario barocco*. Liguori, Napoli, Italia.

The AEREST Reading Database

Marcello Ferro

Istituto di
Linguistica Computazionale
ILC-CNR Pisa, Italy
marcello.ferro@ilc.cnr.it

Sara Giulivi

Scuola Professionale
della Svizzera Italiana
SUPSI Locarno, Switzerland
sara.giulivi@supsi.ch

Claudia Cappa

Istituto di
Fisiologia Clinica
IFC-CNR Pisa, Italy
claudia.cappa@cnr.it

Abstract

Aerest is a reading assessment protocol for the concurrent evaluation of a child's decoding and comprehension skills. Reading data complying with the *Aerest* protocol were automatically collected and structured with the *ReadLet* web-based platform in a pilot study, to form the *Aerest* Reading Database. The content, structure and potential of the database are described here, together with the main directions of current and future developments.

Aerest è un protocollo di valutazione della lettura che misura in parallelo la capacità di decodifica e quella di comprensione del testo. Il protocollo è stato applicato in uno studio pilota i cui dati sono stati raccolti attraverso la piattaforma web ReadLet. L'articolo descrive il contenuto, la struttura e le potenzialità del data set risultante, insieme a future direzioni di sviluppo.

1 Introduction

In the PISA 2000 report (OECD, 2003), a distinction is introduced between the concept of “reading literacy” as opposed to “reading”, the latter being restricted to the ability of decoding or reading aloud, the former including a much wider and more complex range of cognitive and meta-cognitive competencies: decoding, vocabulary, grammar, mastery of larger linguistic and textual structures and features, knowledge about the world, but also use of appropriate strategies necessary to process a text (p. 23). In the PISA 2019 report (OECD, 2019) “reading literacy” is

defined as “an individual's capacity to understand, use, evaluate, reflect on and engage with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society”, and as the “range of cognitive and linguistic competencies, from basic decoding to knowledge of words, grammar and the larger linguistic and textual structures needed for comprehension, as well as integration of meaning with one's knowledge about the world” (p.28). Achieving reading literacy is crucial for an individuals' participation in society and ultimately for their realization in academic context, in workplace or, more generally, in life.

To achieve reading literacy, pupils need first and foremost to be able to read accurately, understand what they read, and do this in a reasonably small amount of time. This multifaceted ability is defined here as “reading efficiency”. Efficient reading implies on its turn, in the subject, the development of deep comprehension skills. As a matter of fact, comprehension is a complex construct that requires coordination and processing of several cognitive abilities at word, sentence, and text level (Perfetti et al., 2005; Padovani, 2006), including, but not limited to, building coherent semantic representations of what is being read (Nation and Snowling, 2000), making lexical and semantic inferences, using reading strategies, activating metacognitive control (Carretti et al., 2002).

When it comes to assessment, the above described complexity is not given due consideration and is, among other aspects, at the basis of the inadequacy of most protocols currently available. The latter often measure comprehension performance (in a way the “product” of reading comprehension) without considering the underlying processes, or treat those processes as if they were independent, not in interaction with one another. In addition, reading comprehension tests often tend to be used interchangeably, while they actually

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

measure different skills or processes and are not really comparable to one another (Colenbrander et al., 2017; Keenan et al., 2008; Cutting and Scarborough, 2006; Calet et al., 2020; Joshi, 2019). Finally, most currently available reading assessment tools fail to focus on reading efficiency, as they normally measure decoding and reading comprehension separately. This leads to failure in the identification of kids having difficulties in integrating the above mentioned abilities.

The AEREST protocol for reading assessment was designed and developed to fill this gap, by testing student skills in three tasks: reading aloud, silent reading, and listening comprehension. In the last two conditions, the student's comprehension of the text being read is assessed through a questionnaire. Only in the reading aloud condition, the text can also contain non-words.

In 2019, AEREST was tested in schools located in Southern Tuscany (Italy) and in the Canton of Ticino (Switzerland), involving a total of 433 children, from the 3rd grade of the Italian primary school through to the first grade of the Italian middle school (6th grade). The protocol was automatically administered using a prototype version of *ReadLet* (Ferro et al., 2018a; Ferro et al., 2018b), a web-based platform that records large streams of time-aligned, multimodal reading data.

2 ReadLet

The ReadLet platform monitors and records a user's behaviour during the execution of various reading tasks. It includes a central repository and a set of web applications, background services for pre- and post-processing analysis and query tools. The ReadLet endpoint is an ordinary tablet running a web application which is responsible for the administration of the reading protocol. The ReadLet app overrides most of the actions taken by a tablet to respond to typical touch events on the screen (tapping, scrolling etc.), which is needed to allow a reader to slide across the text displayed on the touchscreen as one would normally do on a printed text on paper.

The child is asked to read a short story displayed on the tablet screen either silently or aloud, and to finger-point to the text while reading. The story is displayed on the tablet one page at a time and the child is free to flip the pages back and forth. During each reading session, the audio stream is recorded along with the time-stamped

touch events caused by the interaction of the user with the touchscreen. At the end of a session, all data are sent to the central repository, ready for post-processing and for further analysis. In the listening task, ReadLet provides an audio-player playing a pre-recorded story. As the user finishes reading or listening, a multiple-choice questionnaire is presented one question at a time. In answering each question, the reader/listener can get back to the full text or play back the audio-player, and search for relevant information.

Captured data are recorded, anonymized, and encrypted locally by the application, and sent to a remote server: i) the user information along with the session settings; ii) the text disposition and layout on the screen; iii) the audio stream (i.e. the user's voice while reading aloud), iv) the time-stamped finger interaction during the reading task and in filling the questionnaire; v) the timing of the answers to each question, along with possible self-corrections. ReadLet is equipped with tools for the automated linguistic analysis of texts. The tools, together with a finger-tracking-to-text alignment module, make it possible to capture the user finger-tracking behaviour (e.g. forward tracking, regressions, tracking pauses) and the time spent on the text for different text unit levels (page, paragraph, sentence, token, syllable, morpheme, n -gram, letter) and different linguistic levels (e.g. morphological, lexical, syntactic). Furthermore, the ReadLet speech-to-text alignment module (currently under development) will allow the automatic assessment of decoding accuracy during reading-aloud sessions, by analysing hesitations, reading errors, and self-corrections.

3 The AEREST protocol

As already mentioned, the AEREST protocol was created to provide teachers and education professionals with an accurate, non-invasive, child-friendly assessment tool that could identify the full range of students with low reading efficiency. Unlike current protocols, that usually fail to identify students who do well in the single abilities underlying reading when assessed one at a time, but struggle in the integration of those abilities, the AEREST protocol allows identification of all children manifesting difficulties, in so doing favoring access to specifically tailored enhancement training programs for all those who may need them. The AEREST assessment protocol includes three

tasks: 1. Reading comprehension; 2. Listening comprehension; 3. Decoding.

3.1 Reading comprehension

In order to carry out this task, subjects are provided with a tablet, displaying a story that contains narrative as well as descriptive parts. The texts used for comprehension assessment are based on existing stories written by well-known authors and modified by adding or cutting out text, in order to achieve two main objectives.

The first objective is to obtain a balanced mixture of narrative and descriptive text. In our opinion, this reflects more closely the kind of texts we normally encounter in life, which are hardly ever barely descriptive or barely narrative. Keeping this separation (as most reading assessment tools actually do) would lead, in our opinion, to a less ecological way of assessing reading comprehension.

The second objective is to obtain a text that would allow assessment of all (or most of) the cognitive processes involved in reading comprehension (this is usually not found in other assessment tools currently available). This is made possible through 15 comprehension questions that engage subjects in:

1. retrieving the general content of the text;
2. identifying specific information in the text; (who/what/where/when/...). Usually 4 questions out of 15 concerns this kind of information;
3. identifying temporal relations;
4. identifying cause-effect and sequential relations;
5. making inferences of different kinds;
6. retrieving information from syntactic structure (for example understanding if some event in the story has actually happened or not, based on the verb tenses used by the author);
7. forming mental representations (in general, subjects are prompted with 4 different images of a character or situation in the story and are asked to determine which image corresponds to what they have read);
8. spotting incongruities and errors;
9. retrieving word meaning from context;
10. identifying text register and style;
11. identifying text type.

For each question, the subject can choose among four different answers, out of which only one is correct.

Before starting the task, kids are told that they have no time limit. Subjects are instructed to read the story silently from beginning to end, always pointing their finger to the text being read. Once they reach the end of the story, they are prompted with 15 comprehension questions. These are displayed, one at a time, on the bottom part of the screen, while the text is available in the top part. They can re-read the text, or chunks of it, as many times as they want, by scrolling up and down the text on the screen.

Analysing the responses to the comprehension questions, built as described above, allows to understand which of the processes underlying comprehension are leveraged by the subject and which ones are not efficient and need support through specific, personalised training.

In order to consider comprehension abilities independent of decoding skills (that may be weaker in some subjects, for example in kids with dyslexia) the listening comprehension test described underneath was included in the protocol.

3.2 Listening comprehension

As with the reading comprehension task, subjects are given a tablet and headphones for story listening. After hearing the whole story for the first time, kids start answering comprehension questions one by one, upon hearing them through their headphones and reading them on the tablet's screen. In order to reduce the child's working memory load, some of the questions are asked only after the text passage containing the relevant information is heard for the second time.

3.3 Reading aloud

In this task, children are asked to read aloud stories with a similar narrative structure. At the end of each story, one of the story characters (typically with some kind of supernatural powers: an alien, a witch, ecc.) starts speaking an unknown language, which consists of non-words following the phonology and morpho-syntax of Italian, and some Italian function words. We include here an example of text used for this task.

E come se stesse leggendo su quel vetro,
rivelò a Lucilla la ricetta della segretis-

sima pozione: "Prendi una sirta mellusa e gafala in un tulo. Spisola una rifa e lubica una buva. Non zudugnare e non tapire le vughe. Quita le puggie, zuba i mumini e ralla un tifurno."

The administrator takes notes on the subject's errors, hesitations and self-corrections throughout the task. Meanwhile, the subject's performance is also recorded by the tablet. In addition, as for the reading comprehension task, children are instructed to always finger-point to the text being read. The child's reading score is then calculated taking off 1 point for each spelling error, 0.5 point for each word stress error, 0.5 point for each self-correction. No points or fractions of point are subtracted for hesitations, as they already have an impact on reading time.

4 Data structure

Data are stored at different levels. Texts are pre-processed with NLP tools (Dell'Orletta et al., 2011) for text tokenization, POS tagging, dependency parsing, readability analysis, syllabification, n -gram splitting, and, finally, frequency information by means of a reference corpus.

Session settings are stored to include metadata such as the administrator identifier, user information (a unique identifier, child's affiliation and grade level, possible annotations), the text being read and its layout (e.g. margins, font size and family, letter and line spacing), task type (i.e. silent reading, reading aloud, or listening comprehension).

At the end of each session, all recorded data are sent to a remote server. Basic data include information about the tablet (e.g. the user agent string, the screen resolution), time-stamps of the beginning and end of the reading task and of questionnaire answering. More detailed data include the disposition of the text on the tablet screen (i.e. coordinates of the bounding box of each letter), touchscreen events (i.e. event type, time-stamp, and finger coordinates), the audio stream (sampled at 48KHz stereo and compressed in MP3 format at 128kbps), answers to the questionnaire and their timing.

Post-processing tools enrich stored data offline. A finger-tracking-to-text alignment algorithm binds touchscreen events over time to the text layout at the character level. This is done by

creating two black and white images and performing a convolution operation over them: the first image represents the text disposition on the screen, where each line is rendered as a filled black rectangle on a white background; the second represents the user finger-tracking over time, where each segment between a touch-begin and a touch-end event is rendered as a black rectangle on a white background. During the execution of the convolution operation, the vertical and horizontal offsets which maximize the overlapping of the black areas within the two images indicate the optimal alignment to be taken into account. Such binding allows for subsequent modelling and evaluation of the reading dynamic, as well as for measurement of the reading time at different levels of granularity: from single letters and syllables through to sentences, and whole pages or documents.

5 Collected Data

In 2019, the AEREST protocol was administered to a total of 433 students. A total of 12 narrative texts was used, one for each of the four grade levels and the three assessment tasks. Details of participants and texts are reported respectively in Tables 1 and 2.

Grade	Italy		Switzerland	
	<i>N</i>	<i>Age</i>	<i>N</i>	<i>Age</i>
3	78 (13)	8.6 (0.4)	22 (4)	8.8 (0.4)
4	71 (14)	9.6 (0.3)	21 (2)	9.7 (0.5)
5	94 (25)	10.6 (0.4)	23 (2)	10.7 (0.4)
6	54 (6)	11.5 (0.4)	70 (2)	11.9 (0.4)
TOT	297 (58)	10.0 (1.1)	136 (10)	10.9 (1.3)

Table 1: Sample size (number of children with disorders between brackets) and mean age (standard deviation between brackets) of the participants involved in the study, across grades (from the 3rd to the 6th grade level) and countries (Italy and Switzerland).

Grade	silent	aloud		listening
	<i>words</i>	<i>words</i>	<i>nonwords</i>	<i>words</i>
3	588	177	53	572
4	750	180	74	527
5	951	216	80	941
6	711	352	83	734

Table 2: Number of tokens in the texts administered during the study, across grades (from the 3rd to the 6th grade level) and decoding conditions (silent reading, reading aloud, and listening).

6 Results and discussion

Tablets proved to be easy to use and well accepted devices, extremely instrumental and accurate for data collection with toddlers and older children (Frank et al., 2016; Semmelmann et al., 2016). Tablet data confirmed high standards of ecological validity, and a high correspondence with data collected with other, more traditional tools (e.g. eye-tracking, see Lio *et al.* (2019)), and protocols. Within the present work, the collected data allowed for the evaluation of the decoding and comprehension skills of the children involved in the study. For each grade level, Aerest decoding performance, expressed in syllables per second, was shown to be in line with more classical reading assessment reports (Cornoldi et al., 2010), for both words and non-words. Furthermore, the use of the finger tracking allowed for the validation of the correlation of the time spent on each word with basic features such as frequency and length: statistical analysis with linear mixed-effect models shows a highly significant correlation ($p < 0.0001$), thus confirming the reliability of the adopted technique.

Decoding and comprehension performance scores are shown in Fig. 1. Data are normalized for each grade level group, so that all data groups can be overlapped on the same plot. Indeed, data belonging to each group was divided by the median value of control children only. In this way data can be graphically compared, being a value of 0.5 equal to half the mean performance of control children, a value of 1 equal to average behaviour, and a value of 2 indicates a double outperforming with respect of the average performance.

7 Conclusions and future work

The AEREST protocol was shown to be effective in characterizing the decoding and comprehension performance of children of late primary school and early middle school in text reading tasks. Results are clear and encouraging, opening the way to further, more detailed, dynamic, and multimodal analysis. Completion of the current AEREST protocol with a second battery of tests is foreseen in the near future. This will provide schools with two different test batteries, to be used for assessment at the beginning and end of school year, for adequate monitoring of pupils' reading and reading comprehension skills. A version of the protocol conceived for clinical context is also foreseen, as

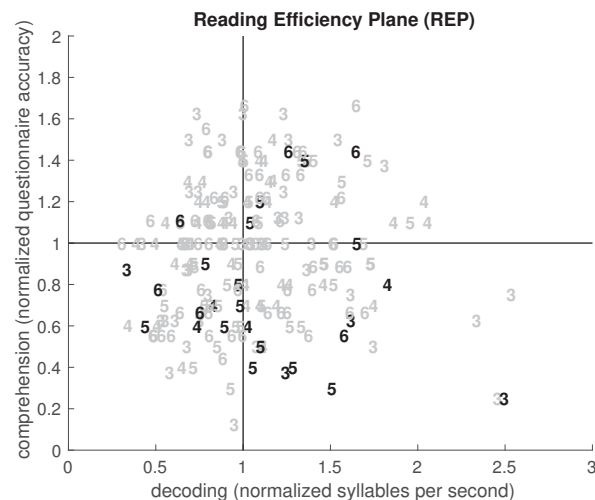


Figure 1: Reading Efficiency Plane for the reading comprehension task (silent reading and comprehension questions). The decoding performance on silent reading (expressed as the normalized syllables per second) is shown in the horizontal axis, while the comprehension performance (expressed as the normalized questionnaire accuracy) is shown in the vertical axis. For each grade level group (from the 3rd to the 6th grade level), the two measures are normalized on the basis of the performance of children with typical reading development. Each child is represented by a digit marker indicating the grade level. Typically and atypically developing readers are shown respectively in gray and black.

well as translation and adaptation of the protocol to languages other than Italian.

The collected data will be assembled in a multimodal linguistic resource and made freely available to the scientific community.

Acknowledgments

This work was supported by the Swiss grant "AEREST: An Ecological Reading Efficiency Screening Tool" (2017-2020) funded by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI), and by the Italian project "(Bio-)computational models of language usage" (2018-) funded by the Italian National Research Council (DUS.AD016.075.004, ILC-CNR).

A special thanks goes to all schools that took part in the study, in particular: Ist. Comprensivo di Manciano-Capalbio (Grosseto, Italy), elementary school of Novaggio, (Ticino Switzerland), lower

secondary school of Bedigliora (Ticino, Switzerland).

References

- Nuria Calet, Rocío López-Reyes, and Gracia Jiménez-Fernández. 2020. Do reading comprehension assessment tests result in the same reading profile? a study of spanish primary school children. *Journal of Research in Reading*, 43:98–115.
- Barbara Carretti, Cesare Cornoldi, and Rossana De Beni. 2002. Il disturbo specifico di comprensione del testo scritto. In S. Vicari and M.C. Caselli, editors, *I disturbi dello sviluppo: neuropsicologia clinica e ipotesi riabilitative*, pages 169–189. Il Mulino, Bologna.
- Danielle Colenbrander, Lyndsey Nickels, and Saskia Kohnen. 2017. Similar but different: differences in comprehension diagnosis on the neale analysis of reading ability and the york assessment of reading for comprehension. *Journal of Research in Reading*, 40(4):403–419.
- Cesare Cornoldi, Patrizio E. Tressoldi, and Nicoletta Perini. 2010. Valutare la rapidità e la correttezza della lettura di brani. nuove norme e alcune chiarificazioni per l'uso delle prove mt. *Dislessia*, 7:89–101.
- Laurie E. Cutting and Hollis S. Scarborough. 2006. Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading*, 10(3):277–299.
- Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Franco Alberto Cardillo, and Vito Pirrelli. 2018a. ReadLet: an ICT platform for the assessment of reading efficiency in early graders. page 61, Edmonton, Alberta (Canada), 25-29 September, 2018. 11th International Conference on the Mental Lexicon.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouaphae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018b. Readlet: Reading for understanding. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 1–6.
- Michael C. Frank, Elise Sugarman, Alexandra C. Horowitz, Molly L. Lewis, and Daniel Yurovsky. 2016. Using tablets to collect data from young children. *Journal of Cognition and Development*, 17(1):1–17.
- R. Malatesha Joshi. 2019. Componential model of reading (cmr): Implications for assessment and instruction of literacy problems. In D. A. Kilpatrick, R. M. Joshi, and R. K. Wagner, editors, *Reading development and difficulties*, pages 3–18. Springer, Dordrecht (The Netherlands).
- Janice M. Keenan, Rebecca S. Betjemann, and Richard K. Olson. 2008. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3):281–300.
- Guillaume Lio, Roberta Fadda, Giuseppe Doneddu, Jean-René Duhamel, and Angela Sirigu. 2019. Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(5392):1–13.
- Kate Nation and Maggie J. Snowling. 2000. Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied psycholinguistics*, 21(2):229–241.
- OECD. 2003. Learners for life. student approaches to learning. results from PISA 2000. <https://doi.org/10.1787/9789264103917-en>, OECD Publishing, Paris.
- OECD. 2019. Assessment and analytical framework. <https://doi.org/10.1787/b25efab8-en>, OECD Publishing, Paris.
- Roberto Padovani. 2006. La comprensione del testo scritto in età scolare. una rassegna sullo sviluppo normale e atipico. *Psicologia clinica dello sviluppo*, x(3):369–398.
- Charles A. Perfetti, Nicole Landi, and Jane Oakhill. 2005. The acquisition of reading comprehension skill. In M. J. Snowling and C. Hulme, editors, *The science of reading: a handbook*, chapter 13, pages 227–247. Blackwell, Oxford.
- Kilian Semmelmann, Marisa Nordt, Katharina Sommer, Rebecka Röhnke, Luzie Mount, Helen Prüfer, Sophia Terwiel, Tobias W Meissner, Kami Koldewyn, and Sarah Weigelt. 2016. U Can Touch This: How Tablets Can Be Used to Study Cognitive Development. *Frontiers in psychology*, 7:1021, jul.

Græcissāre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin

Greta Franzini, Marco Passarotti,
Francesco Mambrini, Giovanni Moretti
Università Cattolica del Sacro Cuore
CIRCSE Research Centre
Largo Gemelli 1,
20123 Milan, Italy
name.surname@unicatt.it

Federica Zampedri
Università degli Studi di Pavia
Pavia, Italia
federica.zampedri01@universitadipavia.it

Abstract

English. This paper describes the addition of an index of 1,763 Ancient Greek loanwords to the collection of Latin lemmas of the *LiLa: Linking Latin* Knowledge Base of interoperable linguistic resources. This lexical resource increases LiLa's lemma count and tunes its underlying data model to etymological borrowing.

1 Introduction

*"Graecia capta ferum victorem cepit"*¹
HORACE, *Epistles*, II, 1, 156

Boasting over two thousand years' worth of written attestation, Latin's evolutionary history is among the longest in existence. The diachronic and geographical reach of the Roman Empire exposed Latin, an Indo-European Italic language, to many regional dialects and languages, including Ancient Greek. The mutually profitable linguistic contact between Latin and Ancient Greek², facilitated by their similar morphosyntactic structures and characteristic syntheticity (Ledgeway, 2012, pp. 10-28), is most evident in their vocabulary, chiefly calques and loanwords. Both lexemes presuppose a certain knowledge of the donor language, but while the former takes from the donor *with* translation, the latter does not (Hock and Joseph, 2009, p. 252).

Examples of Latin words calqued from Ancient Greek are *unicornuus* "unicorn" (*unus* "one" + *cornu* "horn") from *μονόκερως* (*μόνος* "one" + *κέρας* "horn"), and *infans* "infant" (*in-* "not" + *fans* "speaking") from *νήπιος* (negative prefix *νη-*

+ *ἔπος* "speech"). Calques can also involve affixes, as is the case of Latin's suffix *-us* being substituted for the Greek *-os* (Hock and Joseph, 2009, p. 253). The adjective "dramatic", for instance, is attested as both *dramaticos* and *dramaticus*.

Example loanwords in Latin are *crocodilus* "crocodile", imported from the Ancient Greek *κροκόδειλος*, and *liquiritia* "liquorice" from *γλυκύριζα*. Adams identifies three categories of Greek loans in Latin (2003, p. 443):

- (1) words for which there existed a Latin equivalent; the writer was so familiar with the local Greek term that he adopted it in response to local conditions;
- (2) local Greek technical terms for which it might have been difficult to find a Latin equivalent; and
- (3) transfers determined by a writer's lack of fluency in Latin, as a result of which he either adopted Greek words because he was unaware of their Latin equivalents, or did so unconsciously because of his poor command of Latin.

For each category, Adams provides a handful of examples, including (1) *(h)amaxa* from *ἄμαξα* "wagon", (2) *buneurum* from *βούνευρον* "whip of oxhide" and (3) *arura* from *ἄρουρα* "land".

Over the course of its long history, Latin lexicography has produced a plethora of lexical resources, notably dictionaries, thesauri and lexica. Many are available in machine-readable form but their differing annotation schemes and formats are seldom interoperable. In an effort to offset the issue, the *LiLa: Linking Latin* project is leveraging Linked Data technology to dovetail a wide range of Latin resources into an interoperable whole, producing an ever-growing lexically-based data model capable of accommodating etymological, morphological, syntactic and semantic informa-

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹"Captive Greece captured her savage conqueror" (our translation).

²In Egypt, for instance.

tion, and more besides (Passarotti et al., 2020)³. In LiLa, glossaries, lexica, treebanks, textual resources and tools intersect and interact through their common denominator, the lemma (itself, incidentally, a loanword from the Ancient Greek λήμμα). Indeed, the LiLa Knowledge Base hinges on a lemma bank of approximately 130,000 lemmas largely derived from the lexical basis of LEM-LAT (Passarotti et al., 2017). As textual and lexical resources are added to the Knowledge Base, LiLa’s lemma bank and coverage of the Latin lexicon grow in size.

Though chiefly targeting readily available lemmatised resources on the web, LiLa also creates linguistic resources in-house as a means of further developing its underlying data model. Examples of these are the *Index Thomisticus* Treebank (Passarotti, 2019) and *Latin VALLEX* (Passarotti et al., 2016). Here, we describe the addition of a new homegrown lexical resource, the *Index Graecorum Vocabulorum in Linguam Latinam* (Saalfeld, 1874), to the LiLa Knowledge Base of Linguistic Resources for Latin.

2 Data and Methodology

Etymological data is not new to LiLa. Mambriani and Passarotti (2020) describe the inclusion of 1,391 entries from the *Etymological Dictionary of Latin and the other Italic Languages* (De Vaan, 2008) modelled against the *lemonETY* etymological extension (Khan, 2018) of the *OntoLex Lexicon Model for Ontologies (lemon)* (McCrae et al., 2017), which have provided LiLa with 1,465 Proto-Italic and 1,393 Proto-Indo-European reconstructed forms. Whereas those entries came to Latin via inheritance, the work described here targets (nativised) loans from Ancient Greek⁴.

The *Index Graecorum vocabulorum in linguam Latinam translatorum quaestiunculis auctus* (hereafter *IGVLL*) is a list of 1,763 Ancient Greek loanwords in the Latin language published in 1874 by classical scholar Günther Alexander E. A. Saalfeld. An extended edition of the *Index*, published in 1884 as *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der Griechischen Lehn- und Fremdwörter im Latenischen*, is the most comprehensive lexicographic

collection of its kind, counting roughly six to eight thousand entries (Saalfeld, 1884)⁵.

Of the two, the size and Optical Character Recognition (OCR) quality of the 1874 edition best suited a first development of a derivative linguistic resource, conducted as part of a Master’s internship at the CIRCSE Research Centre in Milan⁶.

IGVLL is structured into three columns of information: the Latin loanword (occasionally accompanied by variants), the Ancient Greek source lemma(s) (multiple lemmas include graphical, morphological and dialectal variants), and a record of attestations (see Figure 1). Explanatory notes at the bottom of the page provide additional context. In thirteen cases, question marks indicate some level of uncertainty⁷, and, as is convention, asterisks are used to identify thirty-nine unattested –and thus reconstructed– Ancient Greek forms.

acanthus ἀκανθός³) Verg. ge. 4.
acapnos ἀκαπνος Mart. 13, 15 in lemm.
acatalectus ἀκατάληκτος Gramm.

Figure 1: Three lexical entries in IGVLL, translating to “Bear’s Foot (plant)”, “without smoke”, and “acatalectic (line of verse)”, respectively.

2.1 Data Preparation

Judging by the illegible Greek, the engine used to produce the OCR’d text available from Internet Archive (ABBYY FineReader 8.0) was set to recognise the Latin alphabet only (see Figure 2).

acanthus axav&og 3) Verg. ge. 4.
acapnos awrcvog Mart. 13, 15 in lemm.
acatalectus axaTah]%%og Gramm.

Figure 2: Latin OCR of Fig. 1, Internet Archive.

The OCR quality of the text written in the Latin alphabet, however, was sufficient to automatically isolate and tabulate the Latin lemmas, which were then manually cleaned. Next, this list was automatically mapped against the LiLa lemma bank to measure the degree of lexical overlap, which came up at 1,488 unique matches (84.40%), 207

³<https://lila-erc.eu/>

⁴The Latin verb *gracissō* used in the title of this paper is a nativised version of the Greek γράμιζω “to imitate the Greeks; speak Greek”. For a detailed overview of linguistic “nativisation”, see Hock and Joseph (2009, pp. 247-57).

⁵Crude estimate of an average ten to fifteen entries per page, for a total 592 pages.

⁶https://centridiricerca.unicatt.it/circse_index.html

⁷The specific uncertainty remains unclear as no editorial documentation is provided.

ambiguous matches (11.74%) and 68 unmatched lemmas (3.85%). Unique matches inherited their respective LiLa identifier, ambiguous matches were manually disambiguated, and unmatched lemmas were added –once again, manually– to the LiLa lemma bank. Ambiguities were caused by homography between lemmas belonging to different categories, be those morphosyntactic (the lemma *philosophus*, for instance, matched against LiLa’s adjective *philosophus* “philosophical” and common noun *philosophus* “philosopher”) or inflectional (the common noun *er* might refer to the masculine *er*, *eris* “hedgehog” or the invariable *er* (graphical variant of *R*) “seventeenth letter of the Latin alphabet”). Of the 68 unmatched lemmas, 33 were graphical variants of lemmas already present in LiLa and 35 were new additions.

Next, we OCR’d IGVLL with Tesseract v. 4.1.1 set to Ancient Greek recognition (Smith, 2007)⁸. As Figure 3 shows, contrary to the Latin OCR the noise affecting Greek lemmas required heavy manual intervention for clean tabulation, e.g. the rectification of instances of χ (cappa) misread as χ (chi) or of π (pi) misread as $\tau\tau$ (double tau) and viceversa, missing breathings and incorrect accents, to mention but a few.

’ βοδηίμα95 ἄχανθος 3) γοῖρ. 56. 4
0 ΒΟΒΟΠΟΒ’ ἄχατχυνος Ματῦ. 18, 15 πῖ Ἰθιηπι.
’ δοδία]θοῖι8. ἄχατάληκχτος γιαιτ.

Figure 3: Ancient Greek OCR of Fig. 1, Tesseract.

In LiLa, a lemma can have one or more graphical variants, known as “written representations” (e.g. the verb *sacrifico* “to sacrifice” is also attested as *sacrufico*), as well as inflectional variants, with which it holds a symmetric “lemma variant” property or relation in the Knowledge Base (the active *sacrifico*, *sacrufico* vs. the dependent *sacrificor*, *sacruficor*).

Therefore, for the purposes of LiLa, where the editor provides multiple Ancient Greek lemmas for a single Latin loanword, e.g. *burrus* “red” πυρρός (πυρρός); *cyperum* “rush (botany)” κύπειρον (κύπειρος), these were distinguished into written representations of the same lemma (i.e. πυρρός vs. πυρρός) and lemma variants (i.e. the neuter κύπειρον vs. the masculine κύπειρος).

Compounds such as *authepsa* “an urn, boiler”,

⁸For the most recent overview of Ancient Greek optical character recognition, see Robertson and Boschetti (2017).

(αὐτός & ἔψω) were tabulated as two separate words, and entries followed by a question mark (13 in total) were marked as “uncertain”.

2.2 Data Model

The transformation of IGVLL into an RDF lexicographic resource bound for LiLa relied on a combination of vocabularies. In line with previous etymological work, we integrated the aforementioned lemon and lemonETY modules of OntoLex to represent lexical entries in IGVLL. The example lemma *abacus* “sideboard” shown in Listing 1 is treated as an `ontolex:LexicalEntry` linked to LiLa’s own *abacus* (lemma ID 86829) through the property `ontolex:canonicalForm`.

```
a                                ontolex:LexicalEntry;  
rdfs:label                        "abacus";  
ontolex:canonicalForm <..lemma/86829>;
```

Listing 1: Latin

We employed the *Simple Knowledge Organization System (SKOS)* (Miles and Bechhofer, 2009) to point Ancient Greek lemmas to their corresponding canonical forms in a machine-readable version of the Greek-English *Liddell-Scott Jones (LSJ)* lexicon (Blackwell, 2018). As Listing 2 shows, we modelled the Ancient Greek source lemma of *abacus*, ἄβαξ, as an `etymon`, which, in the absence of a Linked Data Knowledge Base for Ancient Greek, currently points to a blank node.

```
a                                lemonEty:etymon;  
rdfs:label                        "ἄβαξ";  
lime:language                     "grc";  
ontolex:canonicalForm  
                                [ontolex:writtenRep "ἄβαξ"];  
skos:exactMatch <urn:cite2...:n51>.
```

Listing 2: Ancient Greek

The `skos` property stores the LSJ identifier of ἄβαξ as an `exactMatch` to denote an exact correspondence between the Ancient Greek lemma of IGVLL and that of LSJ. Failing an exact match, the property `skos:broadMatch` is used to indicate that the IGVLL lemma is incorporated in a different entry of LSJ (e.g. the IGVLL noun φυσική “science of nature, physics” does not have its own entry in LSJ but is listed as a nominalised adjective under the adjectival entry φυσικός “natural”); further, failing both exact and broad matches, the property `skos:relatedMatch` is used to indicate a loose relation between IGVLL and LSJ (e.g. IGVLL’s πορφύριζον “purple

dye pigment”, neuter present participle of πορφυρίζειν, and LSJ’s verb πορφυρίζω “to be purplish”). As LSJ is not currently equipped with a URN resolver, no actionable link can be made between LiLa and LSJ.

If multiple written representations of a Greek word are listed in the IGVLL, those are all assigned to the canonical form of the related etymon, for instance `ontolex:canonicalForm` [`ontolex:writtenRep` "πυρρός", "πυρρός"].

In the case of multiple Ancient Greek variant lemmas, these are all treated as individual etyma, with the difference that the primary etymon points to the URI(s) of the other etyma –classified as both `lemonEty:etymon` and `lemonEty:cognate`– via the additional property `lemonEty:cognate` (Listing 3).

```
a lemonEty:etymon;
rdfs:label "χύπειρον";
lime:language "grc";
ontolex:canonicalForm
  [ ontolex:writtenRep "χύπειρον" ];
skos:exactMatch <urn:cite2...n60988>;
lemonEty:cognate
  <http://lila.../IGVLL/id/etymon/499>.
```

Listing 3: Lemma variants: χύπειρον/ος

Latin composite words in IGVLL never point to an Ancient Greek compound but to the two constituent lemmas. In contrast, in the LSJ lexicon seven of the total thirteen multi-word lexical entries in IGVLL are traced back to a Greek compound lemma, e.g. *authepsa* (IGVLL: αὐτός & ἔψω; LSJ: αὐθέψης⁹). In keeping with the IGVLL, we employed the `decomp:subterm` property of *lemon*¹⁰ to point the Latin lexical entry to its two constituent Ancient Greek etyma and reconciled these with LSJ using the `skos:relatedMatch` property.

The etymology of *abacus* is expressed with the *CIDOC Conceptual Reference Model (CRM)* class *E89 Propositional Object*¹¹ as a borrowing by way of the `lemonEty:etymon` property. This set-up is also valid for calques, should these become available in future.

The *CRM_{inf}* extension of CRM and the *Open Vocabulary* (Davis, 2004) were used to rep-

⁹http://folio2.furman.edu/ljsj/?urn=urn:cite2:hmt:lsj.chicago_md:n17373

¹⁰<https://lemon-model.net/lemon-cookbook/node21.html>

¹¹<http://www.cidoc-crm.org/Entity/e89-propositional-object/version-6.0>

resent uncertainty as a “belief” or confidence value (Stead et al., 2019; Doerr, 2003; Mambri and Passarotti, 2020). Specifically, we coded uncertainty as a *CRM_{inf}* *Belief* class (`crminf:I2`) carrying an arbitrary *Belief Value* (`crminf:I6`) of 0.5 (Listing 4).

```
a crminf:I2;
crminf:I6 [a crminf:I6;
           ov:confidence 0.5].
```

Listing 4: Uncertainty

Additionally, we employed the *Dublin CoreTM Metadata Terms* vocabulary to supply the resource with descriptive metadata, such as publisher and licence (DCMI, 2020).

All editorial notes in IGVLL were excluded from the data model.

As previously mentioned, with this development LiLa’s etymological purview now covers both direct inheritance *and* borrowing. Figure 4, for example, shows all etymological information in the Knowledge Base associated with LiLa’s common noun *muscus* “moss, musk” (top row, centre node). LiLa’s “muscus” is connected to the “muscus” lexical entries of both IGVLL and the Brill Etymological Dictionary via the bidirectional *OntoLex* property `canonicalForm`. These lexical entries point to their respective etyma via the directed *lemonETY* *etymology* and *etymon* properties.

3 Conclusion

This paper describes the preparation and integration of Saalfeld’s *Index Graecorum Vocabulorum in Linguam Latinam* (1874) in the LiLa Knowledge Base of Linguistic Resources for Latin. This first list of 1,763 Latin loans from Ancient Greek adds 68 new Latin lemmas to LiLa, stretches its data model to include borrowing and has been mapped to the digitised Greek-English Liddell-Scott-Jones lexicon. Beyond LiLa, this linguistic resource might be integrated in other resources, such as dictionaries (Bowers and Romary, 2016) or digital scholarly editions. Future improvements might acquire a list of calques (Detreville, 2015; Fruyt, 2011) and the extended edition of Saalfeld’s *Index* (1884).

The data and code for the project are available at: <https://github.com/CIRCSE/index-graecorum-vocabulorum>.

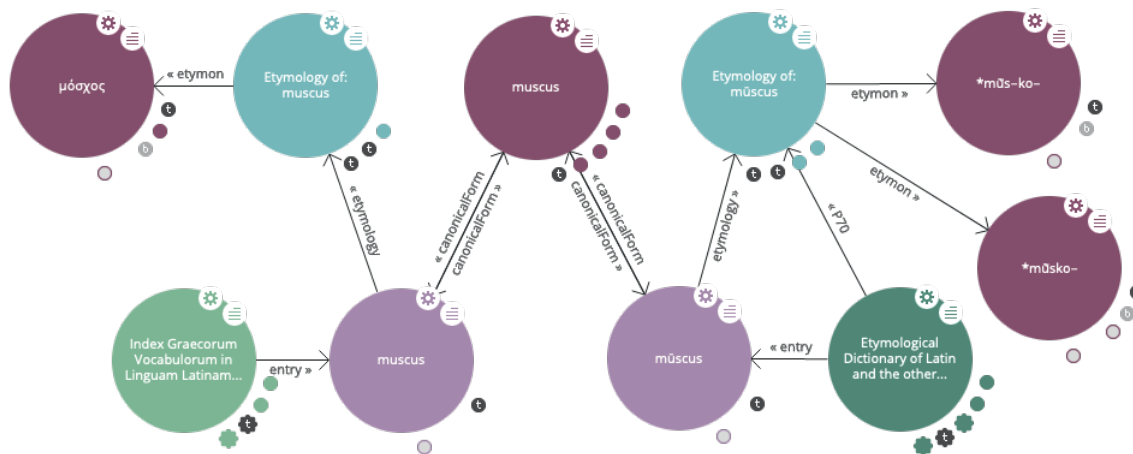


Figure 4: Etymology of *muscus* “moss, musk” in LiLa.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

- James Noel Adams. 2003. *Bilingualism and the Latin Language*. Cambridge University Press.
- Christopher W. Blackwell. 2018. CITE LSJ Browser, v. 1.4.0. <http://folio2.furman.edu/ljsj/index.html>.
- Jack Bowers and Laurent Romary. 2016. Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative*. <https://doi.org/10.4000/jtei.1643>.
- Ian Davis. 2004. vocab.org, v. 20040205. <https://vocab.org/>.
- Usage Board DCMI. 2020. DCMI Metadata Terms. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.
- Michiel De Vaan. 2008. *Etymological Dictionary of Latin: And the other Italic Languages*, volume 7 of *Leiden Indo-European Etymological Dictionary Series*. Brill, Amsterdam. <https://brill.com/view/title/12612?language=en>.
- Eleanor Detreville. 2015. An Overview of Latin Morphological Calques on Greek Technical Terms: Formation and Success. Master’s thesis, University of North Carolina, Asheville, NC.
- Martin Doerr. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75. <https://dl.acm.org/doi/10.5555/958671.958678>.
- Michèle Fruyt. 2011. Latin Vocabulary. In James Clackson, editor, *A Companion to the Latin Language*, pages 144–156. Wiley-Blackwell.
- Hans Henrich Hock and Brian D. Joseph. 2009. *Language history, language change, and language relationship: an introduction to historical and comparative linguistics*. Trends in Linguistics. Studies and monographs, 218. Mouton de Gruyter, 2nd revised edition. <http://gen.lib.rus.ec/book/index.php?md5=819682013cda444e5dad7bf866a45d64>.
- Anas Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9:304–320. <https://doi.org/10.3390/info9120304>.
- Adam Ledgeway. 2012. *From Latin to Romance. Morphosyntactic Typology and Change*. Oxford University Press, Oxford.
- Francesco Mambrini and Marco Passarotti. 2020. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi, and Besim Kabashi, editors, *Proceedings of the Globalex Workshop on Linked Lexicography (LREC 2020)*, pages 20–28, Paris, France. European Language Resources Association (ELRA). <https://doi.org/10.5281/zenodo.3862156>.
- John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of the Electronic lexicography in the 21st century conference (eLex 2017)*, pages 237–251. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- Alistair Miles and Sean Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference. *W3C recommendation*, 18. <https://www.w3.org/TR/skos-reference/#L4858>.

- Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2599–2606. <https://www.aclweb.org/anthology/L16-1414/>.
- Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31. <http://www.ep.liu.se/ecp/article.asp?issue=133&article=006&volume=>.
- Marco Passarotti, Francesco Mambrini, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas: The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212. <https://doi.org/10.4454/ssl.v58i1.277>.
- Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, volume 10 of *Age of Access? Grundfragen der Informationsgesellschaft*, pages 299–319, Berlin-Boston. De Gruyter GmbH. <https://doi.org/10.1515/9783110599572>.
- Bruce Roberston and Federico Boschetti. 2017. Large-Scale Optical Character Recognition of Ancient Greek. *Mouseion*, 14:341–359. <https://doi.org/10.3138/mous.14.3-3>.
- Günther Alexander E. A. Saalfeld. 1874. *Index graecorum vocabulorum in lingvam latinam translatorum quaestivncvllis avctvs*. apvd F. Berggold, Berlin. <https://archive.org/details/indexgraecorumvo00saal/>.
- Günther Alexander E. A. Saalfeld. 1884. *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen*. Carl Gerold’s Sohn, Wien. <https://archive.org/details/tensaurusitalogr00saal>.
- Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. <https://doi.org/10.1109/ICDAR.2007.4376991>.
- Stephen Stead, Martin Doerr, Christian-Emil Ore, and Athina et al. Kritsotaki. 2019. CRMinf: the Argumentation Model, Version 0.10.1 (draft). <http://new.cidoc-crm.org/crminf/sites/default/files/CRMinf%20ver%2010.1.pdf>.

L'impatto emotivo della comunicazione istituzionale durante la pandemia di Covid-19: uno studio di Twitter Sentiment Analysis.

Gloria Gagliardi
Università di Napoli "L'Orientale"
Napoli, Italy
ggagliardi@unior.it

Lorenzo Gregori
Università di Firenze
Firenze, Italy
lorenzo.gregori@unifi.it

Alice Suozzi
Università Ca' Foscari Venezia
Venezia, Italy
956549@stud.unive.it

Abstract

Italiano. L'articolo si propone di investigare l'impatto delle comunicazioni istituzionali durante la crisi sanitaria causata dal Covid-19 in Italia, attraverso l'analisi computazionale dell'attività di micro-blogging su Twitter. A partire dal corpus TWITA, abbiamo realizzato una Sentiment Analysis per osservare eventuali cambiamenti nella polarità delle scritture degli utenti, in reazione a undici discorsi istituzionali tenuti durante l'emergenza sanitaria. Variazioni sono state osservate in reazione a tre discorsi tenuti dal Presidente del Consiglio rispettivamente il 4 marzo, il 9 marzo e il 26 aprile 2020.

English. *This paper aims at investigating the impact of institutional communications during the health crisis due to Covid-19 pandemic in Italy, through the analysis of micro-blogging activities on Twitter by means of NLP techniques. We performed a Sentiment Analysis on the TWITA corpus, to pinpoint potential correlations between opinion polarity (positive or negative) of the users and public speeches during the outbreak. Our findings show changes in sentiment polarity related to three institutional speeches delivered by the Italian Prime Minister Giuseppe Conte on March, 4th, March, 9th, and April, 26th 2020.*

1 Introduzione

L'epidemia Covid-19 si è rapidamente diffusa da Wuhan, in Cina, a numerose nazioni in tutto il mondo e il 28 marzo 2020 l'Organizzazione

Mondiale della sanità (OMS) ha ufficialmente dichiarato lo stato di pandemia del Covid-19. In Italia, il primo caso è stato riportato il 21 febbraio 2020; i contagi si sono poi diffusi a diverse regioni e l'8 marzo 2020 l'intera nazione è stata dichiarata zona rossa. È stato inoltre stabilito un lockdown nazionale, che è durato circa due mesi, fino al 4 maggio 2020. La situazione di emergenza è stata drammatica e completamente nuova: le comunicazioni istituzionali, finalizzate ad illustrare le misure di contenimento e ad indirizzare il comportamento della popolazione, hanno dunque giocato un ruolo determinante. La loro diffusione è avvenuta sia attraverso la televisione pubblica, sia attraverso i social network, come Facebook, Instagram e Twitter; di conseguenza, questi sono diventati i principali canali di diffusione di informazioni e di condivisione di opinioni.

La Sentiment Analysis è un campo di studi multidisciplinare finalizzato ad investigare e valutare le opinioni espresse nei testi (Beigi et al., 2016; Mejova et al., 2013; Zimbra et al., 2018) cioè l'orientamento (positivo o negativo) che il parlante esprime verso un oggetto (Jurafsky and Martin, 2019) attraverso dispositivi lessicali. Twitter è una fonte privilegiata di dati per l'analisi di emozioni e opinioni, tanto che oggi si parla di Twitter Sentiment Analysis (TSA) per identificare la specifica branca di ricerca basata su dati estratti da questo social network. Oggi la TSA è usata in vari ambiti e per diversi scopi, come, ad esempio, monitorare le opinioni degli utenti su prodotti commerciali (Ghiassi et al., 2013; Jansen et al., 2009), studiare gli orientamenti politici (Mejova and Srinivasan, 2012; Mejova et al., 2013; Garcia and Thelwall, 2013; Wang et al., 2012; Wang et al., 2014), analizzare i livelli e le cause di stress tra gli adolescenti (Basili et al., 2017) o l'opinione pubblica riguardo i vaccini (Tavoschi et al., 2020).

Questo studio utilizza la TSA per investigare l'impatto che le comunicazioni istituzionali hanno

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Data	Parlante	Argomento	Durata	Parole
04/03/20	G. Conte	Prime azioni del governo per limitare la diffusione del virus (es. chiusura di scuole e università, sospensione degli eventi sportivi)	4'58"	831
05/03/20	S. Mattarella	Primo discorso sull'emergenza coronavirus	3'40"	455
08/03/20	G. Conte	La Lombardia e altre 14 province del Nord sono dichiarate "zona rossa"	15'44"	2357
09/03/20	G. Conte	Decreto "Io resto a casa"	6'37"	995
11/03/20	G. Conte	L'OMS dichiara la pandemia; tutte le attività vengono chiuse (ad eccezione di supermercati e farmacie)	8'59"	1392
21/03/20	G. Conte	Decreto "Chiudi Italia": vengono imposte ulteriori limitazioni alle attività produttive e agli spostamenti individuali	7'06"	942
27/03/20	S. Mattarella	Secondo discorso sull'emergenza coronavirus	7'28"	1105
01/04/20	G. Conte	Il governo estende il lockdown fino al 13 aprile	4'50"	777
10/04/20	G. Conte	Il governo estende il lockdown fino al 3 maggio	19'47"	3124
26/04/20	G. Conte	Annuncio dell'allentamento delle restrizioni a partire dal 4 maggio ("Fase 2")	28'34"	4538
16/05/20	G. Conte	Annuncio di un ulteriore allentamento delle restrizioni	12'04"	1767

Table 1: Corpus ItalC-Covid19

avuto durante la crisi sanitaria di Covid-19 in Italia. Nello specifico, è stato analizzato il sentiment di tweet italiani sul coronavirus per identificare correlazioni tra la polarità delle opinioni degli utenti (positiva o negativa) e i discorsi tenuti dal Presidente del Consiglio e dal Presidente della Repubblica durante la pandemia.

2 Materiali

2.1 Il corpus ItalC-Covid19

Preliminarmente allo studio, è stato costruito un piccolo corpus di dichiarazioni e conferenze stampa pronunciate dal Presidente del Consiglio Giuseppe Conte e dal Presidente della Repubblica Sergio Mattarella, chiamato ItalC-Covid19 (Gagliardi and Suozzi, 2020).¹ Tutti i dati provengono dal canale YouTube ufficiale di Palazzo Chigi² e del Quirinale³.

Attualmente, il corpus include undici testi orali, pronunciati tra il 3 marzo e il 16 maggio 2020. Tab. 1 riassume il loro contenuto. Per ogni video è stata realizzata una trascrizione ortografica, conforme al formato L-AcT (Cresti and Moneglia, 2018), una versione del formato standardizzato CHAT (MacWhinney, 2000).

2.2 Il corpus Twitter Covid-19

I tweet utilizzati per questo studio sono derivati dal corpus 40wita,⁴ un dataset di tweet italiani

raccolto da Basile e Caselli (2020) durante il periodo di emergenza della pandemia, come parte dell'iniziativa *Computational Linguistics and the Covid-19 Outbreak*⁵ promossa dall'Associazione Italiana di Linguistica Computazionale (AILC). Il dataset è stato costruito giornalmente dal 1 febbraio 2020, filtrando TWITA (Basile and Nissim, 2013; Basile et al., 2018), la più grande raccolta di messaggi pubblicati su social media in lingua italiana, con le parole-chiave relative al Covid-19,⁶.

2.3 Il dataset

Al fine di ottenere delle statistiche generali, è stato esaminato un lasso di tempo di 88 giorni, dal 26 febbraio al 23 maggio 2020. Il periodo considerato inizia una settimana prima della prima comunicazione istituzionale, e finisce una settimana dopo l'ultima. Il numero di tweet giornalieri sul coronavirus durante questo periodo va da un minimo di 9.245 a un massimo di 51.490, per un totale di 2.298.566 post. In fig. 1 è illustrata la distribuzione di questi numeri in funzione del tempo. Si può chiaramente osservare una crescita significativa nei primi giorni di marzo, quando, in alcune regioni, il lockdown è iniziato; il picco viene raggiunto il 10 marzo 2020, in corrispondenza

⁵<http://www.ai-lc.it/covid-19/>

⁶covid, covid19, covid-19 corona virus, coronavirus, quarantena, autoisolamento, auto-isolamento, iorestoacasa, stateacasa, COVID19Italia, redditodicittadinaza, eurobond, coronabond, restiamoacasa, preghiamoinsieme, NoMes, milanononsiferma, bergamononsiferma, l'italianonsiferma, abbracciauncinese, iononsonounvirus, iononmiferma, aperisera, covidunstria, italiazonarossa, bergamoisrunning, quarantena, chiudetetutto, aprietetutto, CuraItalia, circordiamotutto, oggisciopero, chiudiamolefabbriche, iononrinuncioallegtrazioni, andratuttobene, INPSdown, percheQuando, cercareDi, ringraziarevoglio, 600euro, CineINPS, COVID19Pandemic.

¹<https://euiccovid19.altervista.org/>

²<https://www.youtube.com/user/governoit>

³<https://www.youtube.com/user/presidenzarepubblica>

⁴<http://twita.di.unito.it/dataset/40wita>

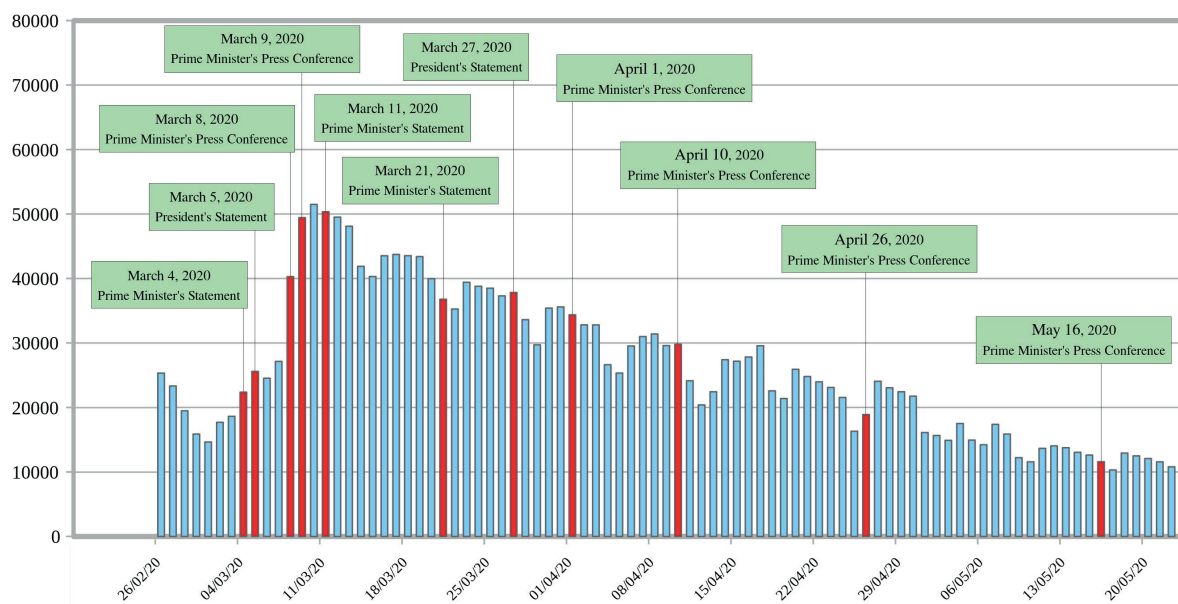


Figure 1: Numero di tweet sul Covid-19 in Italia durante il lockdown.

dell'estensione del lockdown a tutto il territorio nazionale. Segue una diminuzione graduale fino alla fine di maggio, quando la maggior parte delle attività viene riaperta.

La Sentiment Analysis è stata realizzata con Sentita (Nicola, 2018).⁷ Lo strumento, specifico per i tweet italiani, fornisce due valori indipendenti: un valore tra 0 e 1 per la polarità positiva (0 = neutro, 1 = completamente positivo) e uno tra 0 e 1 per la polarità negativa (0 = neutro, 1 = completamente negativo). Per ridurre la complessità del compito, sono stati inclusi nell'analisi solo i tweet con una polarità chiaramente orientata, escludendo quelli con bassa polarità (positivi e negativi sotto la soglia di 0.5) e quelli polarizzati sia in positivo che in negativo (con entrambi i valori superiori a 0.5). Due esempi di tweet analizzati con Sentita sono riportati in tabella 2.

3 Analisi della *sentiment polarity*

Il dataset appena descritto è stato utilizzato per misurare se e in che misura le reazioni dei cittadini alle comunicazioni istituzionali siano riflesse nei loro tweet.

A tal fine, è necessario considerare che la polarità dei tweet non è uniforme durante il giorno (Larsen et al., 2015): i tweet pubblicati al mattino sono solitamente più neutri di quelli della sera. Per avere una panoramica di questa vari-

⁷<https://nicgian.github.io/Sentita/>

Tweet	pos.	neg.
Che palle questo coronavirus mi sta rovinando tutti i piani e mi sta togliendo troppe gioie mi uccido	0.046	0.963
So che non si può abbracciare per il Covid-19 ma io voglio mandare un enorme abbraccio virtuale a medici, infermieri, operatori ospedalieri che in queste settimane lavorano senza sosta per garantire la salute di tutti. Grazie, grazie, grazie.	0.985	0.093

Table 2: Due esempi di tweet con polarità negativa e positiva.

azione all'interno del corpus, è stato misurato l'andamento di tweet positivi e negativi nell'arco delle 24 ore e la media è stata calcolata sull'intero periodo di osservazione (88 giorni). Questo ha permesso di comparare l'andamento di tweet positivi e negativi intorno agli eventi comunicativi con la media giornaliera, per determinare eventuali deviazioni. Per l'analisi è stata considerata una finestra temporale che va da quattro ore prima a quattro ore dopo ogni comunicazione: all'interno di essa è stata misurata la percentuale di tweet positivi e negativi per ogni ora. Per evitare ogni potenziale fluttuazione oraria, è stata misurata la deviazione

dist	hour	pos	neg	pos_dev	neg_dev
-4	15	1.71%	10.91%	-1.44%	1.56%
-3	16	1.80%	11.73%	-1.20%	2.74%
-2	17	2.26%	9.93%	-0.55%	1.07%
-1	18	2.22%	10.17%	-0.53%	1.30%
0	19	1.99%	11.26%	-1.16%	1.08%
1	20	1.79%	12.72%	-1.66%	0.77%
2	21	2.44%	13.01%	-1.70%	-0.01%
3	22	1.14%	10.13%	-2.98%	-3.35%
4	23	2.15%	10.39%	-1.76%	-3.45%

Table 3: Analisi dei tweet nella finestra temporale [-4,+4] rispetto alla comunicazione istituzionale del Presidente del Consiglio del 4 marzo 2020.

di polarità rispetto alla media calcolata su tutti gli 88 giorni. In tab. 3 sono riportati, a titolo di esempio, i dati relativi alla prima comunicazione, la dichiarazione del Presidente del Consiglio del 4 marzo 2020, trasmessa tra le 19:00 e le 20:00.

I dati rappresentati sono i seguenti: distanza in ore dall'evento comunicativo (*dist*), ora del giorno (*hour*), percentuale di tweet positivi (*pos*), differenza tra *pos* e la media percentuale di tweet positivi per la stessa ora del giorno (*pos_dev*); analogamente sono calcolati *neg* e *neg_dev* sui tweet negativi.

I valori delle ultime due colonne (*pos_dev* e *neg_dev*) sono stati divisi in due gruppi: le percentuali relative ai tweet precedenti all'evento (A) e quelle relative ai tweet successivi all'evento (B). Numericamente i tweet considerati per l'analisi (nella finestra temporale [-4,+4] ore da ogni evento) sono 161.197, di cui 27.057 con polarità positiva o negativa. Abbiamo applicato un modello di regressione lineare con una variabile *dummy* per ogni evento, assegnando un valore di 1.0 agli elementi di (A) e di 0.0 a quelli di (B). I *p*-value risultanti sono stati utilizzati per identificare gli eventi dopo i quali la polarità dei tweet è cambiata significativamente.

4 Risultati e discussione

I risultati della regressione lineare sono riportati nella tab. 4.

I dati mostrano che per tre comunicazioni istituzionali si registrano variazioni significative: sono quelle rilasciate il 4 marzo, il 9 marzo e il 26 aprile 2020. Tali risultati sono coerenti con il contenuto delle comunicazioni: le prime due (4 e 9 marzo) hanno annunciato, rispettivamente, l'istituzione delle prime zone rosse e l'estensione del lockdown a tutto il territorio nazionale. L'ultima, al contrario, ha annunciato

Date	POS <i>p</i>	NEG <i>p</i>
03/04/20	0.03152	0.03406
03/05/20	0.9412	0.3768
03/08/20	0.1195	0.721
03/09/20	0.006432	0.001612
03/11/20	0.07579	0.371
03/21/20	0.6074	0.2086
03/27/20	0.9202	0.08326
04/01/20	0.07515	0.8863
04/10/20	0.09524	0.7186
04/26/20	0.449	0.0006091
05/16/20	0.7202	0.4424

Table 4: *p*-value della Regressione Lineare con variabile dummy applicata ai dati riferiti alla finestra temporale [-4,+4].

le misure contenitive relative alla cosiddetta Fase 2 (cioè la parziale riapertura delle attività, e la fine del lockdown). È importante sottolineare che i tre eventi hanno suscitato tre reazioni diverse (fig. 2):

- 4 marzo 2020: significativa diminuzione sia di tweet positivi sia di negativi;
- 9 marzo 2020: significativa crescita sia di tweet positivi sia di tweet negativi;
- 26 aprile 2020: significativa crescita esclusivamente di tweet negativi, con percentuale di tweet positivi invariata.

La prima dichiarazione è stata seguita da un forte calo sia di tweet positivi sia di negativi: questo è coerente con il clima di grande incertezza che ha caratterizzato i primi giorni di marzo 2020. Il Covid-19 si stava, infatti, ancora diffondendo prevalentemente nel Nord Italia, e la dichiarazione analizzata ha stabilito le prime misure contro un'ulteriore diffusione del virus. La percezione di pericolo crescente, da un lato, e l'attesa generale di altre misure contenitive, dall'altro, si traduce in opinioni meno chiaramente orientate.

La seconda dichiarazione è servita ad annunciare il decreto "Io resto a casa": da quel momento, tutta la nazione è diventata "zona protetta". Le reazioni emotive a questa comunicazione sono ambivalenti, con una crescita importante di tweet sia positivi, che negativi. Questa "doppiezza" emotiva mostra da un lato la paura causata dall'epidemia in corso, e il lutto per le vittime, in quel momento particolarmente numerose; dall'altro, l'approvazione che il lockdown nazionale ha suscitato nella popolazione, e la tendenza a rimanere ottimisti e soprattutto uniti, esemplificata dall'hashtag *andràtuttobene*.

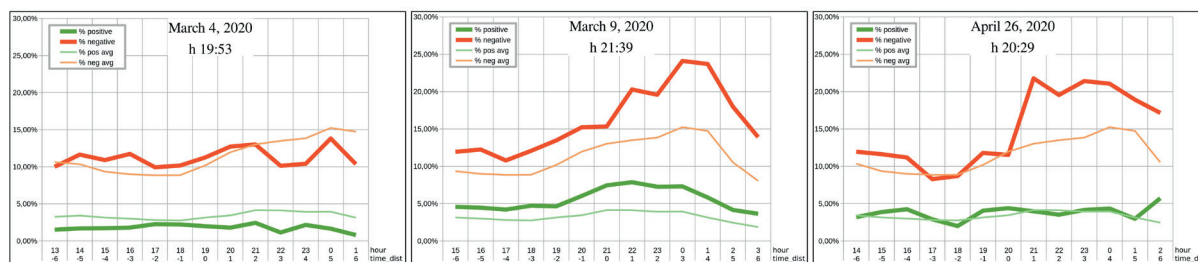


Figure 2: Percentuale di tweet con sentiment positivo e negativo intorno alle comunicazioni del 4 marzo, 9 marzo e 26 aprile 2020 (linee spesse); sono riportate anche le percentuali delle medie orarie (linee fini).

La terza comunicazione ha, invece, provocato solo una significativa crescita di tweet negativi: il 26 aprile sono state annunciate le linee guida per la progressiva riapertura, e l'ingresso nella cosiddetta Fase 2. Una reazione così negativa è imputabile prevalentemente alla delusione: le misure annunciate sono state percepite come ancora troppo restrittive, tanto che molti tweet ironizzavano sulla sostanziale identità di Fase 1 e Fase 2. La delusione è derivata da una precedente aspettativa: un rapido ritorno alla normalità. Questo desiderio, a sua volta, è stato provocato da effetti più generali che la pandemia ha avuto sul benessere mentale ed emotivo della popolazione italiana. La pandemia di Covid-19, analogamente a quanto osservato per altre epidemie, ha suscitato reazioni psicologiche negative, come un'incidenza maggiore di depressione, stress (Shultz et al., 2016), preoccupazione (Thompson et al., 2017), e ansia di essere contagiati (Horney et al., 2010); il tutto è stato esacerbato dall'impossibilità di accedere a servizi di supporto e di dedicarsi ad attività come hobbies e sport.

5 Conclusioni

Per concludere, questo studio ha lo scopo di investigare le reazioni dei cittadini italiani alle comunicazioni istituzionali durante l'epidemia di Covid-19. Grazie al crescente utilizzo dei social network, è infatti possibile esplorare le reazioni psicologiche ad eventi traumatici, sia individuali sia collettivi: secondo i nostri dati, le comunicazioni istituzionali che hanno provocato reazioni psicologico-emotive più forti sono quelle pronunciate il 4 e 9 marzo e il 26 aprile 2020, che hanno annunciato le misure più drastiche di contenimento dell'epidemia. Ulteriori studi potrebbero espandere questo lavoro, considerando contesti nazionali differenti (sia dove sono state attuate

misure di contenimento del virus simili a quelle italiane, sia dove sono state attuate misure molto diverse). Inoltre, potrebbero essere oggetti di analisi eventi socio-politici diversi, come suggerito dallo studio EmoItaly⁸ (es. elezioni politiche, Giornata Internazionale dei Lavoratori, omicidio di George Floyd).

References

- Valerio Basile and Tommaso Caselli. 2020. 40wita 1.0: An collection of Italian Tweets during the COVID-19 Pandemic. <http://twita.di.unito.it/dataset/40wita>.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In Alexandra Balahur, Erik van der Goot, and Andres Montoyo, editors, *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, page 100–107. Association for Computational Linguistics.
- Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term Social Media Data Collection at the University of Turin. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, pages 40–45. CEUR-WS.org.
- Roberto Basili, Valentina Bellomaria, Niels J. Bugge, Danilo Croce, Francesco De Michele, Federico Fiori Nastro, Paolo Fiori Nastro, Chantal Michel, Stefanie J. Schmidt, and Frauke Schultze-Lutter. 2017. Monitoring Adolescents' Distress using Social Web data as a Source: the InsideOut Project. In Roberto Basili, Malvina Nissim, and Giorgio Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it*, Rome.
- Ghazaleh Beigi, Xia (Ben) Hu, Maciejewski Ross, and Huan Liu, 2016. *An Overview of Sentiment Analy-*

⁸<http://corpora.ficlit.unibo.it/EmoItaly/>

- sis in Social Media and Its Applications in Disaster Relief*. Springer, New York.
- Emanuela Cresti and Massimo Moneglia. 2018. Chapter 13. The illocutionary basis of information structure: The Language into Act Theory (L-Act). In Evangelia Adamou, Katharina Haude, and Martine Vanhove, editors, *Information Structure in Lesser-described Languages. Studies in prosody and syntax*, pages 360–402. John Benjamins, Amsterdam - Philadelphia.
- Gloria Gagliardi and Alice Suozzi. 2020. ItaIC-Covid19: Institutional Communication in Times of COVID-19. <http://gloriagagliardi.altervista.org/psychological-reactions-to-the-pandemic/>.
- David Garcia and Mike Thelwall. 2013. Political alignment and emotional expression in Spanish Tweets. In *Proceedings of the TASS Workshop at SEPLN*.
- Manoochehr Ghiassi, J. Skinner, and David Zimbra. 2013. Twitter brand Sentiment Analysis: A hybrid system using n-gram analysis and Dynamic Artificial Neural Network. *Expert Systems with Applications*, 40(16).
- Jennifer A. Horney, Zack Moore, Meredith Davis, and Pia D. M. MacDonald. 2010. Intent to receive pandemic influenza A (H1N1) vaccine, compliance with social distancing and sources of information in NC, 2009. *PloS One*, 5(5).
- Jim Jansen, Mimi Zang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11).
- Dan Jurafsky and H. Martin, James. 2019. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Third Edition Draft*. Pearson, New York.
- Mark E Larsen, Tjeerd W Boonstra, Philip J Batterham, Bridianne O’Dea, Cecile Paris, and Helen Christensen. 2015. We feel: mapping emotion on Twitter. *IEEE journal of biomedical and health informatics*, 19(4):1246–1252.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah.
- Yelena Mejova and Padmini Srinivasan. 2012. Political Speech in Social Media Streams: YouTube Comments and Twitter Posts. In Noshir S Contractor, Brian Uzzi, Michael Macy, and W. Nejdl, editors, *Proceedings of the 3rd Annual ACM Web Science Conference*, New York. Association for Computing Machinery.
- Yelena Mejova, Padmini Srinivasan, and Bob Boynton. 2013. GOP primary season on Twitter: popular political sentiment in social media. In *Proceedings of the ACM Web Search and Data Mining Conference (WSDM’13)*, Rome. ACM.
- Giancarlo Nicola. 2018. Bidirectional Attentional LSTM for Aspect Based Sentiment Analysis on Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- M. James Shultz, Janice L. Cooper, Florence Bain-gana, Zelde Oquendo, Maria A. and Espinel, Benjamin M. Althouse, Louis Herns Marcelin, Sherry Towers, Maria Espinola, Laurie McCoy, Clyde B. and Mazurik, Milton L. Wainberg, Yuval Neria, and Andreas Reckemmer. 2016. The role of fear-related behaviors in the 2013–2016 West Africa Ebola virus disease outbreak. *Current Psychiatry Reports*, 18(104).
- Lara Tavo-schi, Filippo Quattrone, Eleonora D’Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni, and Pier Luigi Lopalco. 2020. Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics*, 16(5).
- Rebecca R. Thompson, Dana Rose Garfin, E. Alison Holman, and Roxane Cohen Silver. 2017. Distress, worry, and functioning following a global health crisis: A national study of Americans’ responses to Ebola. *Clinical Psychological Science*, 5(3).
- Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju. Association for Computational Linguistics.
- Yu Wang, Tom Clark, Eugene Agichtein, and Jeffrey Staton. 2014. Towards Tracking Political Sentiment through Microblog Data. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 88–93, Baltimore.
- David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Transactions on Management Information Systems*, 9(5).

On Knowledge Distillation for Direct Speech Translation

Marco Gaido^{1,2}, Mattia Antonino Di Gangi^{3*}, Matteo Negri¹, Marco Turchi¹

¹Fondazione Bruno Kessler, Trento, Italy

²University of Trento, Trento, Italy

³AppTek, Aachen, Germany

{mgaido, digangi, negri, turchi}@fbk.eu

Abstract

English. Direct speech translation (ST) has shown to be a complex task requiring knowledge transfer from its sub-tasks: automatic speech recognition (ASR) and machine translation (MT). For MT, one of the most promising techniques to transfer knowledge is knowledge distillation. In this paper, we compare the different solutions to distill knowledge in a sequence-to-sequence task like ST. Moreover, we analyze eventual drawbacks of this approach and how to alleviate them maintaining the benefits in terms of translation quality.

Italiano. È stato dimostrato che la speech translation (ST) diretta è un'operazione complessa che richiede l'adozione di tecniche di knowledge transfer sia da automatic speech recognition (ASR) che da machine translation (MT). Per quanto riguarda MT, una delle tecniche più promettenti è la knowledge distillation (KD). In questo lavoro, confrontiamo diverse possibili soluzioni di KD per addestrare modelli sequence-to-sequence come quelli di ST. Inoltre, analizziamo eventuali problemi causati da questa tecnica e come attenuarli mantenendo i benefici in termini di qualità della traduzione.

1 Introduction

Speech translation (ST) refers to the process of translating utterances in one language into text in a different language. Direct ST is an emerging paradigm that consists in translating without

intermediate representations (Bérard et al., 2016; Weiss et al., 2017). It is a newer and alternative approach to cascade solutions (Stentiford and Steer, 1988; Waibel et al., 1991), in which the input audio is first transcribed with an automatic speech recognition (ASR) model and then the transcript is translated into the target language with a machine translation (MT) model.

The rise of the direct ST paradigm is motivated by its theoretical and practical advantages, namely: *i*) during the translation phase it has access to information present in the audio that is lost in its transcripts (eg. prosody, characteristic of the speaker¹), *ii*) there is no *error propagation* (in cascade systems the errors introduced by the ASR are propagated to the MT, which has no cues to recover them), *iii*) the latency is lower (as data flows through a single system instead of two), and *iv*) the management is easier (as there is a single model to maintain and no integration between separate modules is needed).

On the downside, direct ST suffers from the lack of large ST training corpora. This problem has been addressed by researchers through transfer learning from the high-resource sub-tasks (Bérard et al., 2018; Bansal et al., 2019; Liu et al., 2019), multi-task trainings (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bahar et al., 2019a), and the proposal of data augmentation techniques (Jia et al., 2019; Bahar et al., 2019b; Nguyen et al., 2020). In this work, we focus on the transfer learning from MT. The classic approach consists in pre-training the decoder with that of an MT model. Its benefit, however, is controversial: indeed, (Bahar et al., 2019a) showed that it is effective only with

^{*}Work done during the PhD at Fondazione Bruno Kessler and University of Trento.

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹For instance, the pitch of the voice is a cue for the sex of the speaker. Although the gender is a social aspect and does not depend on physical attributes, in many cases sex and gender coincide, so systems relying on this are likely to have a better accuracy than those that do not have access to any information regarding the speaker (Bentivogli et al., 2020; Gaido et al., 2020b).

the addition of an *adapter* layer, but this has not been confirmed in (Gaido et al., 2020a), while in (Inaguma et al., 2020) it always brought improvements. Another, more promising possibility consists in distilling knowledge from an MT model.

Knowledge distillation (KD) is a knowledge transfer technique introduced for model compression (Hinton et al., 2015). A small *student* model is trained computing the KL-divergence (Kullback and Leibler, 1951) with the output probability distribution of a big *teacher* model. Although KD was introduced in the context of image processing, its effectiveness suggested its adoption in other fields. Specifically, Liu et al. (2019) showed that using an MT system as teacher brings significant improvements to direct ST models. However, they did not compare the different methods to distill knowledge in sequence-to-sequence models (Kim and Rush, 2016) and they did not analyze possible negative effects of adopting this technique.

In this paper, we analyze different sequence-to-sequence KD techniques (*word level*, *sequence-level*, *sequence interpolation*) and their combination in the context of direct ST. Then, we study the effect of the best technique on a strong system trained on a large amount of data to reach state-of-the-art results. We show that word-level KD is the best approach and that fine-tuning the resulting model without KD brings further improvements. Finally, we analyze the limitations and the problems present in models trained with KD, which are partly solved by the final finetuning.

2 Sequence-level Knowledge Distillation

We focus on distilling knowledge from an MT model to an ST model. This is helpful due to the better results achieved by MT, which is an easier task than ST, as it does not involve the recognition of the audio content, and it also benefits from the availability of large training corpora. Our student (ST) model is trained to produce the same output distribution of the teacher (MT) model when the latter is fed with the transcript of the utterances passed as input to the ST model. As KD was introduced in the context of *classification* tasks, while ST and MT are sequence-to-sequence *generation* tasks, an adaptation is required for its application. Kim and Rush (2016) introduced three methods to distill knowledge in sequence-to-sequence models: *i)* word-level KD, *ii)* sequence-level KD, and *iii)* sequence interpolation.

Word-level KD (Word-KD) refers to computing the KL-divergence between the distribution of the teacher and student models on each token to be predicted. As recomputing the teacher output at each iteration is computationally expensive (it needs a forward pass of the MT model), we explored the possibility to pre-compute and store the teacher outputs. To this aim, we experimented with truncating the output distribution to have a lower memory footprint, as proposed in MT (Tan et al., 2019).

Sequence-level KD (Seq-KD) consists in considering as target the output generated by the teacher model using the beam search.

Sequence interpolation (Seq-Inter) is similar to Seq-KD, but the target is the sentence with the highest BLEU score (Papineni et al., 2002) with respect to the ground truth among the *n*-best generated by the beam search with the teacher model.

As done in (Kim and Rush, 2016), we also combine these methods to analyze whether they are complementary or not. Finally, we experiment with fine-tuning the model trained with KD on the reference translations.

3 Experimental Settings

We performed preliminary experiments on a limited amount of data to compare the three KD methods. Then, we created a model exploiting all the available corpora with the best technique to analyze the KD behavior in a real scenario.

3.1 Data

We first experiment using only Librispeech (Kocabiyyikoglu et al., 2018), an ST corpus with English audio, transcripts and French translations. We use the (*audio*, *transcript*) pairs for the ASR pre-training, the (*transcript*, *translation*) pairs to train the MT teacher, and the (*audio*, *translation*) pairs for the ST training.

Then, we built an English-Italian model. In addition to Librispeech, the ASR pre-training involves TED-LIUM 3 (Hernandez et al., 2018), Mozilla Common Voice,² How2 (Sanabria et al., 2018) and the en-it section of MuST-C (Di Gangi et al., 2019a). The MT teacher is trained on the OPUS datasets (Tiedemann, 2016), cleaned using the ModernMT framework (Bertoldi et al., 2017).³

²<https://voice.mozilla.org/>

³With the `CleaningPipelineMain` class.

For ST, we use the en-it section of MuST-C and Europarl-ST (Iranzo-Sánchez et al., 2020).

We pre-process the input audio extracting a 40-dimensional feature vector from a span of 25 ms every 10 ms using Mel filter bank. During this pre-processing performed with XNMT (Neubig et al., 2018), we also apply speaker normalization. The text is tokenized and the punctuation is normalized with Moses (Koehn et al., 2007). We create 8,000 shared BPE merge rules on the MT data of each experiment and apply them to divide the text into sub-word units. Samples lasting more than 20 seconds are discarded in order to avoid out of memory issues during training.

3.2 Models

For ST and ASR we use the S-Transformer architecture (Di Gangi et al., 2019b; Di Gangi et al., 2019c) with logarithmic distance penalty in the encoder. In particular, in the experiments on Librispeech we train a small model using the basic configuration by Di Gangi et al. (2019b), while in the experiment with all the data we follow the BIG configuration. In the second case, we also slightly modify the architecture to improve performance by removing the 2D attention layers and changing the number of Transformer Encoder layers and Transformer Decoder layers to be respectively 11 and 4 in ST and 8 and 6 in the ASR pre-training (Gaido et al., 2020a). The different number of layers between ASR and ST is motivated by the idea of having adaptation layers (Jia et al., 2019; Bahar et al., 2019a).

For MT we use a Transformer with 6 layers for both the encoder and the decoder. In the preliminary experiments, we use a small model with 512 hidden features in the attention layers, 2,048 hidden units in the feed-forward layers and 8 attention heads; in the experiment with more data we double all these parameters.

3.3 Training

We optimize our models with Adam (Kingma and Ba, 2015) using betas (0.9, 0.98). The learning rate increases linearly for 4,000 steps starting from $1e-7$ to $5e-3$. Then it decays according to the inverse square root policy. In fine-tunings, the learning rate is fixed at $1e-4$. A 0.1 dropout is applied and the total batch size is 64. When we do not use KD, the loss is label smoothed cross entropy (Szegedy et al., 2016) with 0.1 smoothing factor.

In the final training with all the data, we apply SpecAugment (Park et al., 2019) with probability 0.5, 13 *frequency masking pars*, 20 *time masking pars*, 2 *frequency masking num*, and 2 *time masking num*. We also increase the overall batch size to 512. Moreover, the ASR pre-training is performed as a multi-task training in which we add a CTC loss (predicting the output transcripts) on the encoder output (Kim et al., 2017).

Our code is based on the Fairseq library (Ott et al., 2019), which relies on PyTorch (Paszke et al., 2019), and it is available open source at <https://github.com/mgaido91/FBK-fairseq-ST>. The models are trained on 8 GPU K80 with 11 GB of RAM.

4 Results

First, we experiment truncating the output distribution generated by the teacher model. Table 1 shows that truncating the output to few top tokens does not affect significantly the performance. On the contrary, the best result is obtained using the top 8 tokens. Hence, all our experiments with Word-KD use the top 8 tokens of the teacher.

Top K	BLEU
4	16.43
8	16.50
64	16.37
1024	16.34

Table 1: Results with different K values, where K is the number of tokens considered for Word-KD.

Then, we try different values for the temperature T parameter. The temperature is a parameter used to sharpen (if $T < 1$) or soften (if $T > 1$) the output distribution. In particular, by adding the temperature, the *softmax* function that converts the *logits* z_i into probabilities p_i becomes:

$$p_i = \frac{e^{z_i/T}}{\sum(e^{z_i/T})} \quad (1)$$

A higher temperature has been claimed to help learning the so-called *dark knowledge* (Hinton et al., 2015), one of the possible reasons alluded to justify the success of KD. Indeed, with a high temperature, the cost function is similar to minimizing the squared distance between the logits produced by the student and teacher networks. So logits with very negative values – which are basically ignored with low temperature – become important to be learnt by the student network. For a demon-

stration, please refer to (Hinton et al., 2015). Table 2 reports the BLEU score for different values of T and indicates that the default $T = 1$ is the best value. This result suggests that, in ST, the networks do not have the capacity of MT models trained on the same data. So focusing on the mode of the probability distribution works best.

T	BLEU
1.0	16.50
4.0	16.11
8.0	14.27

Table 2: Results with different temperatures (T).

	BLEU
Baseline	9.4
Word-KD	16.5
Seq-KD	13.4
Seq-Inter	13.3
Seq-KD + Word-KD	15.7
Word-KD + FT Seq-KD	16.7
Seq-KD + FT Word-KD	16.8
Word-KD + FT w/o KD	16.8

Table 3: Results of the small model on Librispeech with different KD methods and combining them in a single training or in consecutive trainings through a fine-tuning (FT).

Then, we compare the different sequence-level KD techniques. We also combine them either in the same training or in consecutive trainings through a fine-tuning (FT). The results are presented in Table 3. We can notice that all the methods improve significantly over the baseline: KD makes the training easier and more effective. Among them, Word-KD achieves the best results by a large margin. Combining it with another method in the same training is harmful (Seq-KD + Word-KD), while a fine-tuning on a different KD method or without KD (i.e. using the ground-truth target and label smoothed cross entropy) improves results by up to 0.3 BLEU (Seq-KD + FT Word-KD and Word-KD + FT w/o KD). These results confirm the choice by (Liu et al., 2019), but differ from those of (Kim and Rush, 2016). So, we can conclude that the best sequence-to-sequence KD technique is task-dependent and that the best option to distill knowledge from MT to ST is the word-level KD.

To validate the effectiveness of KD in a real case, we create a model translating English utterances into Italian text leveraging all the available corpora for each task. Our ASR pre-trained model scores 10.21 WER on the MuST-C test set, while

the teacher MT model scores 30.3 BLEU on the Italian reference for same test set. We train our ST model first on the ASR corpora for which we generated the target with the MT model (resulting in a Seq-KD + Word-KD training). Note that we could not use this data without Seq-KD or Seq-Inter, hence we opted for the best training including one of them (Seq-KD + Word-KD). Second, we fine-tune the model on the ST corpora with Word-KD. Third, we fine-tune without KD as in the case leading to the best result (Table 3). So, our training is: Seq-KD + Word-KD (on ASR data) + FT Word-KD + FT w/o KD. After the first two steps, our ST model scores 22.8 BLEU on the MuST-C test set, while after the final fine-tuning the result is it scores 27.7 BLEU. This highlights the importance of fine-tuning without KD.

5 Analysis

We analyze the outputs of the en-it model to assess whether, despite the benefits in terms of translation quality, KD introduces limitations or issues. Namely, we checked whether the lack of access of the MT teacher to information present in the audio and not in the text (such as the gender⁴ of the speaker) hinders the ability of the final model to exploit such knowledge. Moreover, we compared the output generated by the model before fine-tuning without KD and after it to determine the reasons of the significant BLEU improvement.

Direct ST systems have been shown to be able to exploit the audio to determine the gender of the speaker and reflect it better in the translations into languages rich of gender marked words (Bentivogli et al., 2020). This is not possible for an MT system that has no clue regarding the speaker’s gender. We tested the performance of our models on the category 1 of the MuST-SHE test set (Bentivogli et al., 2020) (which contains gender marked word related to the speaker) to check whether distilling knowledge from MT harms this advantage of ST systems or not. Table 4 shows that, indeed, systems trained with KD inherit the bias from the MT system and, although the final fine-tuning mitigates the issue, the final model has a higher gender bias than a base ST system without KD (regarding the words related to the speaker).

The better translation of speaker’s gender marked words does not explain the big BLEU im-

⁴This is true if the gender identity coincides with the biological sex. This assumption holds true in nearly all our data.

	BLEU	Female			Male			Bias
		Corr.	Wrong	Diff.	Corr.	Wrong	Diff.	Diff. M - Diff. F
Base ST (Bentivogli et al., 2020)	21.5	26.7	27.2	-0.5	46.3	6.8	39.5	40.0
MT	30.3	10.8	55.5	-44.7	54.4	7.1	47.3	92.0
Seq-KD + Word-KD + FT Word-KD	22.8	12.3	46.5	-34.2	45.4	8.1	37.3	71.5
+ FT w/o KD	27.7	19.8	39.0	-19.2	43.2	10.5	32.7	51.9

Table 4: Accuracy on Category 1 of the MuST-SHE test set of a base direct ST model and models created using KD. A high *Diff.* means that the model is able to recognize the speaker’s gender and the gap between the *Diff.* on the two genders indicates the bias towards one of them. The reported BLEU score refers to the MuST-C test set and shows the translation quality of the model.

provement obtained with fine-tuning. Hence, we performed a manual analysis of sentences with the highest TER (Snover et al., 2006) reduction. The analysis revealed three main types of enhancements, with the first being the most significant.

Samples with multiple sentences. Some utterances contain more than one sentence. In this case, the model trained with KD tends to generate the translation of only the first sentence, ignoring the others. This is likely caused by the fact that MT training data is mostly sentence-level. For this reason, the MT model tends to assign a high probability of the EOS symbol after the dot. The student ST model learns to mimic this harmful behavior and, as in ST training and test samples often include more than one sentence, to wrongly truncate the generation once the first sentence is completed. The fine-tuned model, instead, generates all the sentences.

Verbal tenses. The fine-tuned model tends to produce the correct verbal tense, while before the fine-tuning the verbal tense is often not precise, likely because the MT model favors more generic forms. For instance, “*That meant I was going to be on television*” should be translated as “*Significava che sarei andata in televisione*”. The model before fine-tuning produces “*Questo significava che stavo andando in tv*” while the fine-tuned model uses the correct verbal tense “*Questo significava che sarei andata in televisione*”. Despite relevant for the final score, it is debatable whether this is a real improvement of the fine-tuned model, as in some cases both verbal tenses are acceptable or their correctness depends on the context (e.g. in informal conversations, the usage of conjunctive forms is often replaced with indicative tenses).

Lexical choices. In some cases, the fine-tuned model chooses more appropriate words, probably thanks to the fine-tuning on in-domain data. For instance, the reference translation for “*She has taken a course in a business school, and she has*

become a veterinary doctor” is “*Ha seguito un corso in una scuola di business, ed è diventata una veterinaria*”. The corresponding utterance was translated by the model before the fine-tuning into “*Ha frequentato una lezione di economia ed è diventata una dottoressa veterinaria*”, while after the fine-tuning the translation is “*Ha frequentato un corso in una business school, ed è diventata una dottoressa veterinaria*”.

We can conclude that KD provides a benefit in terms of overall translation quality, but the resulting ST system also learns negative behaviors (such as the masculine default for the speaker-related words that exacerbates the gender bias). These are partly solved by performing a fine-tuning without KD, which keeps (and even enhances) on the other side the translation capabilities.

6 Conclusions

We presented and analyzed the benefits and issues brought by distilling knowledge from an MT system for direct ST models. We compared the different KD techniques and our experiments indicated that the best training procedure consists in a pre-training with word-level KD and a fine-tuning without KD. Then, we showed that KD from MT models causes an increased gender bias, omission of sentences in multi-sentential utterances and more generic word/verbal-tense choices. Finally, we demonstrated that a fine-tuning helps resolving these issues, although the exacerbation of gender bias is not solved, but only alleviated.

Acknowledgments

This work is part of the “End-to-end Spoken Language Translation in Rich Data Conditions” project,⁵ which is financially supported by an Amazon AWS ML Grant.

⁵<https://ict.fbk.eu/units-hlt-mt-e2eslt/>

References

- Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 82–91, New Orleans, Louisiana.
- Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019a. A Comparative Study on End-to-end Speech to Text Translation. In *Proc. of International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore.
- Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019b. On Using SpecAugment for End-to-End Speech Translation. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, China.
- Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on High-resource Speech Recognition Improves Low-resource Speech-to-text Translation. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68, Minneapolis, Minnesota.
- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6923–6933, Virtual.
- Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks. In *Proc. of ICASSP 2018*, pages 6224–6228, Calgary, Alberta, Canada.
- Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, et al. 2017. MMT: New Open Source MT for the Translation Industry. In *Proc. of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 86–91, Prague, Czech Republic.
- Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.
- Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 2012–2017, Minneapolis, Minnesota.
- Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing Transformer for End-to-end Speech-to-Text Translation. In *Proc. of Machine Translation Summit XVII*, pages 21–31, Dublin, Ireland.
- Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting Transformer to End-to-end Spoken Language Translation. In *Proc. of INTERSPEECH*, Graz, Austria, September.
- Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020a. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proc. of the 17th International Conference on Spoken Language Translation*, pages 80–88, Virtual.
- Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020b. Breeding Gender-aware Direct Speech Translation Systems. In *Proc. of The 28th International Conference on Computational Linguistics (COLING 2020)*, Virtual.
- François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In *Proc. of the Speech and Computer - 20th International Conference (SPECOM)*, pages 198–208, Leipzig, Germany.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. In *Proc. of NIPS Deep Learning and Representation Learning Workshop*, Montréal, Canada.
- Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-One Speech Translation Toolkit. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Virtual.
- Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Giménez. Adrià, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *Proc. of ICASSP 2020*, pages 8229–8233, Barcelona, Spain.
- Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *Proc. of ICASSP 2019*, pages 7180–7184, Brighton, UK.
- Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas.

- Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning. In *Proc. of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839, New Orleans, Louisiana.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of 3rd International Conference on Learning Representations (ICLR)*, San Diego, California.
- Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.
- Solomon Kullback and Richard Arthur Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.
- Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. of Interspeech 2019*, pages 1128–1132, Graz, Austria.
- Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proc. of the 13th Conference of the Association for Machine Translation in the Americas*, pages 185–192, Boston, MA.
- Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proc. of the 2020 International Conference on Acoustics, Speech, and Signal Processing – IEEE-ICASSP-2020*, Barcelona, Spain.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.
- Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. of Interspeech 2019*, pages 2613–2617, Graz, Austria.
- Adam Paszke, Sam Gross, Francisco Massa, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. of Advances in Neural Information Processing Systems 32 (NIPS)*, pages 8024–8035. Curran Associates, Inc.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proc. of Visually Grounded Interaction and Language (ViGIL)*, Montréal, Canada.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge.
- Fred W. M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States.
- Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual Neural Machine Translation with Knowledge Distillation. In *Proc. of International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, United States.
- Jörg Tiedemann. 2016. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384. Special Issue: Proc. of the 19th Annual Conference of the European Association of Machine Translation (EAMT).
- Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing, ICASSP 1991*, pages 793–796, Toronto, Canada.
- Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proc. of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.

Predicting Social Exclusion: A Study of Linguistic Ostracism in Social Networks

Greta Gandolfi

University of Trento

greta.gandolfi@alumni.unitn.it

Carlo Strapparava

Fondazione Bruno Kessler (FBK)

strappa@fbk.eu

Abstract

Ostracism is a community-level phenomenon, shared by most social animals, including humans. Its detection plays a crucial role for the individual, with possible evolutionary consequences for the species. Considering (1) its bound with communication and (2) its social nature, we hypothesise the combination of (a) linguistic and (b) community-level features to have a positive impact on the automatic recognition of ostracism in human online communities. We model an English linguistic community through Reddit data and we analyse the performance of simple classification algorithms. We show how models based on the combination of (a) and (b) generally outperform the same architectures when fed by (a) or (b) in isolation.¹

1 Introduction

Ostracism is a social phenomenon meant to ignore or exclude an individual from a group, performed by an individual or a group. Due to its relevance in our everyday life - as a threat to basic needs (Wesselmann et al., 2012) - and its impact on community-level essential patterns - such as mother-infant attachment, xenophobia, and leadership (Raleigh and McGuire, 1986) - each person must develop a system to predict and avoid it. Humans and other social animals (such as rhesus monkeys, for example) use ostracism as a form of social control on problematic group members, as a way to strengthen their group and to remove members that do not conform to social norms. Moreover, it reinforces the hierarchical role of the per-

petrators while causing the social or even the actual death of their direct victims. For these reasons, the scope of ostracism allows researchers to assume that its identification has adaptive advantages (Wesselmann et al., 2012).

Given its intrinsic relation with communication and its community-level impact, we assume that its detection can be automatised relying on linguistic and extra-linguistic, community-level, social features. We expect both the types of information to be predictive but to work best when combined.

Reddit communities² can be used as proxies of linguistic communities since they provide huge amounts of linguistic data³ paired with social information. The performance of minimal binary classifiers, such as Naïve Bayes and SVM, can be investigated to analyse the relevance of such cues to distinguish between prospective ostracised or not-ostracised members of a group, modelling our adaptive ability to detect ostracism in advance.

2 Background

As far as we know, this can be defined as the first attempt to analyse the phenomenon of ostracism from the point of view of computational linguistics.

Linguistic behaviours have been analysed as predictors of social exclusion. Researchers focused both on the treatment of silence - i.e. the voluntary suspension of any linguistic utterance - (Williams, 2002) and on the proactive use of language - i.e. the voluntary application of particular linguistic acts. An example of such linguistic acts is the use of gender-exclusive language (e.g., using *he* to indicate both a male member or a female one), experienced as ostracism by female members of the group (Stout and Dasgupta, 2011).

Also non-linguistic cues have been considered,

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²Described in Section 3.

³Mainly written in the English language.

such as members' competitive behavior (Wu et al., 2015) or agreeableness (Hales et al., 2016).

Predictors, in both of the cases, have been searched in the victims' behaviour or personality type. Critically, our approach is meant to focus primarily on cues coming from the perpetrators.

The following proposal is purely observational; we will define a set of possible predictors of social exclusion, not relying on a proper theoretical model. We think that this exploration can help other researchers to define a paradigm of social exclusion, that focuses on general empirical linguistic and extra-linguistic data.

3 Methods and Tools

Reddit is an American news aggregation and discussion website, it ranks as the fifth most visited website in the U.S., with an average of 430M monthly active users and more than 130K active communities⁴. It is organised in *subreddits* i.e. hubs for discussion, controlled by moderators and administrators and characterized by a transparent hierarchical structure. Moderators and administrators are listed in each community page and the importance of each user on the platform is represented by its *karma*⁵.

Reddit provides a good balance of linguistic and extra-linguistic data. Even if some sort of jargon is present, the linguistic analysis is not constrained by particular boundaries of length and form (being more reliable, in this case, than Twitter data). The extra-linguistic features that are particularly relevant for this work are the ones reflecting the structure and the hierarchical organisation of the Reddit community. A more detailed description of these features and their selection will be provided below.

3.1 Dataset

To collect data we used *PRAW* (*Python Reddit API Wrapper*), a Python package that allows for simple access to Reddit's API (<http://praw.readthedocs.io>).

The dataset creation has been strongly controlled. Having in mind the work of Raleigh and McGuire (1986), that focused on the behaviour of sub-adults and adults non-human primates leaving a group after they failed to maintain their role as dominant figures, we selected all *reactions* (i.e.,

⁴Data from <https://www.redditinc.com>.

⁵i.e. a number that is computed relying on the *popularity* (ratio between *upwards* and *downwards*) of the total amount of its *comments* and *submissions* (discussion posts).

comments to submissions and comments to posts) addressed to ten moderators during nine years⁶.

3.1.1 Moderator selection

We distinguished between moderators that left the linguistic community and moderators that are still relevant (in terms of karma), trying to match their period of activity on Reddit, for future longitudinal comparisons⁷.

Ostracised moderators are defined on the basis of two identification processes. First, we automatically searched for all the post in the subreddit */r/redditrequest*. It can be defined as a space in which users are allowed to ask to remove a moderator from a group, due to his/her/their inactivity or abusive, harmful or irrespective behaviour towards the other users (in that particular group or in the whole Reddit community)⁸.

We identified 5 users. These are proxies of directly ostracised individuals that violated the social norms of their groups. Secondly, we automatically searched for all the moderators' posts that stated their willingness to leave the Reddit community followed by their actual inactivity. We simply performed a word-based search. We selected other 5 moderators, representing a subset of individuals that left the community deliberately.

3.1.2 Sampling

To create a balanced dataset, we searched for popular moderators, who shared the same period of activity with the target ones. We selected the ones with the highest karma. For each year of production, then, we randomly extracted a sample of comments received, to obtain the same number of reactions per year, for each moderator.

We created a dataset⁹ of 4,200 linguistic reactions, 50% of which are addressed to the moderators that

⁶From 2010 to 2019.

⁷For example, if one of the ostracised moderators have been active in the community from the summer of 2013 to the winter of 2015, we searched for another admin that has been productive in the same period of time, without being excluded from the community.

⁸We could select only the posts in which the user name of the target moderator was explicit (e.g. "*Please remove moderator X from the subreddit Y*"), several times, however, it was more likely to find posts with this form: "*Please remove the moderator of the subreddit Y*", which is more ambiguous. Then we reduced the set of moderators, keeping only the ones that actually stopped their activity i.e. that are no more active with respect to the definition of *inactivity* provided by the Reddit administrators: 3 months of silence in whole Reddit environment.

⁹Relevant materials can be found here: <https://github.com/gretagandolfi/ostracism>.

left the community. The remaining 50% is composed by reactions addressed to active and popular moderators.

3.2 Models

We trained and tested a Naïve Bayes and a SVM algorithm (10-fold cross-validation) and we analysed the fluctuations of their accuracy scores. We took 0.50 as the baseline since the corpus is new and perfectly balanced.

4 Feature selection

To select the right features to detect ostracism, we tried to focus on the formal properties of written English, intentionally ignoring semantically relevant information. This choice is justified by our willingness to proceed in a domain-general fashion and by the awareness of the fact that, generally, ostracism differs from hate speech or swear, being more subtle.

4.1 Linguistic Features

Punctuation and Stop-words Punctuation marks and function words can reveal the syntactic structure of a text, being useful in authorship attribution and gender classification tasks (Koppel et al., 2006; Sarawgi et al., 2011). Their analysis does not involve semantics, thus promoting generalisation. Moreover, punctuation has been considered helpful in performing sentiment detection (Barbosa and Feng, 2010).

Length The length of the comments can give hints on the conversation modality. Short posts, for example, can sometimes show a closer relationship between users if compared to longer ones. Intuitively, fewer words are uttered when interlocutors feel aligned one with each other, while re-phrasing and the need for long explanations are signs of misalignment and misunderstanding, plausible manifestations of conflict (Clark and Henetz, 2014). We computed the median length of the sentences (identified by the sentence tokeniser provided by NLTK python package) that compose each comment, coding long and short comments differently.

Emoticons Emoticons are meant to express feelings. They have been shown to play a crucial role in sentiment analysis (Shin and Maldonado, 2013). The use of emoticon can reveal an author's

positive or negative attitude towards a target individual. We compute the informativeness of the emoticons performing the VADER analysis that provides polarity scores for each reaction passed to the model (Hutto and Gilbert, 2015).

4.2 Extra-linguistic features

In this context, we define extra-linguistic features the set of relevant data which is not related to the users' language in use. Extra-linguistic features mainly relate to the hierarchical organization of subreddits or the users' popularity.

Moderators Raleigh and McGuire (1986) showed how the behaviour of ostracised ruling primates can be seen as a function of the relations between the prospective ruling individuals and other members of the group. Considering this fact, we decided to study the reactions addressed to moderators from the Reddit community, as a way of formalising and implement the idea of the balancing of power in human and animal communities. Reactions can come from normal users, administrators or moderators themselves. Here, we took as a feature the role of the author of each reaction, computing its relevance for the classification task¹⁰.

Score Each Reddit post is associated with a publicly visible score. Being defined as the sum of the *upvotes* (likes, positive integers) and *downvotes* (dislikes, negative integers) that the target post or comment has obtained since it was written, the score provides an idea of how much the product is useful, funny or appreciated, from the point of view of the community members.

Reddit Karma The karma is a measure of the appreciation and the respect that a user gains in years of activity. Its computation is based on the ratio of the scores of each post and comment he/she/they produced. We considered the karma of the users addressing our targets.

5 Experiment

We can operationalise the impact of linguistic and extra-linguistic features on the binary classification task looking at the fluctuations of the models' accuracy. We focused on minimal questions, such as: do the linguistic features have an impact on the classification accuracy? Which is the best (i.e.

¹⁰We coded *basic users* with 0, *moderators* with 0.5 and *admins* with 1.

most accurate) combination? What is the impact of each extra-linguistic feature on the classification accuracy? Does the performance get better if we combine linguistic and extra-linguistic features?

6 Results

6.1 Linguistic and Extra-linguistic Features

The relevance of the linguistic features and extra-linguistic features taken singularly is given by the scores reported in Table 1¹¹. The best linguistic combination is C3, which contains all the linguistic features considered. It is possible to notice that, at this level, the accuracy depends on the number of linguistic features considered, increasing as the latter increases. Regarding the set of extra-linguistic features, the social status of the reaction’s author (*moderator*) seems to be the most relevant.

Table 1: Linguistic Features and Extra-linguistic features

Features	NB	SVM
Punctuation	.550	.579
Stopwords	.569	.604
Length	.580	.580
Emoticons	.499	.499
C1	.588	.615
C2	.590	.620
C3	.609	.623
Moderator	.595	.595
Reddit Karma	.508	.508
Score	.532	.532

6.2 Linguistic + Extralinguistic Features

Table 2 shows the result of combinations of linguistic and extralinguistic features¹².

The mean accuracy of each combination (provided by the 10-fold cross-validation measure) is, in a statistically relevant way ($p\text{-values} < 0.05$), different from the mean accuracy of both the

¹¹C1 stands for the combination of punctuation and stop words, C2 for punctuation, stop words and sentence length and C3 for punctuation, stop words, sentence length and emoticons.

¹²C1, C2 and C3 represent the sets of linguistic features listed above, and each row of the table contains the accuracy scores given by the summation of the social feature(s) (on the left). EL1 stands for the combination of moderator and score; EL2 for score and Reddit karma; EL3 for moderator, score and Reddit karma.

Table 2: Linguistic + Extralinguistic Features

Features	C1		C2		C3	
	NB	SVM	NB	SVM	NB	SVM
Moderator	.625	.636	.625	.638	.614	.639
Karma	.597	.616	.607	.620	.608	.623
Score	.603	.619	.605	.620	.612	.624
EL1	.626	.641	.620	.644	.618	.643
EL2	.605	.620	.609	.621	.612	.625
EL3	.622	.642	.621	.642	.617	.646

models when trained only on linguistic or extra-linguistic features. Moreover, for all the combinations, the SVM models outperform the Naïve Bayes models.

7 Conclusion

We explored the phenomenon of social exclusion through Reddit data within a period of 9 years. We collected reactions addressed to moderators, here considered as leading figures of the groups. We selected 10 moderators that left the community influenced by the linguistic and non-linguistic behaviour of the group they lead. We performed a binary classification task on a total of 14200 linguistic reactions addressed to each of the target moderators, analysing the influence of linguistic and extra-linguistic or social patterns on two simple models’ performance.

We showed how the performance of both models increases if linguistic and extra-linguistic features are combined. The best combination of features, concerning the SVM model, is given by the combination of all the linguistic features and all the social features considered. We can consider this work as an attempt to follow the statements of the sociolinguistics that considers language as intrinsically bound up with society (Hovy, 2018).

Our experiment and the relative techniques are simple and easy to replicate. We think that they can be also applied in non-English domains, just using a translating system for the stop-words. All the other features can be directly generalised to other languages.

References

- Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *23rd International Conference on Computational Linguistics, COLING*, volume 2, pages 36–44.
- Herbert H Clark and Tania Henetz. 2014. Working together. In *The Oxford handbook of language and social psychology*, page 85. Oxford University Press, USA.
- Andrew H. Hales, Matthew P. Kassner, Kipling D. Williams, and William G. Graziano. 2016. Disagreeableness as a cause and consequence of ostracism. *Personality and Social Psychology Bulletin*, 42(6):782–797. PMID: 27044246.
- Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, jun.
- C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.
- Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Michael J. Raleigh and Michael T. McGuire. 1986. Animal analogues of ostracism: Biological mechanisms and social consequences. *Ethology and Sociobiology*, 7(3):201–214.
- Ruchita Sarawagi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of 2011 Conference on Computational Natural Language Learning - CoNLL*.
- S.Y. Shin and J.C. Maldonado, editors. 2013. *Exploiting emoticons in sentiment analysis*. Association for Computing Machinery, Inc.
- Jane G. Stout and Nilanjana Dasgupta. 2011. When he doesn’t mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin*, 37(6):757–769. PMID: 21558556.
- Eric Wesselmann, James Nairne, and Kipling Williams. 2012. An evolutionary social psychological approach to studying the effects of ostracism. *Journal of Social, Evolutionary, and Cultural Psychology*, 6:309, 09.
- Kipling D Williams. 2002. *Ostracism: The power of silence*. Guilford Press.
- Long-Zeng Wu, D. Lance Ferris, Ho Kwong Kwan, Flora Chiang, Ed Snape, and Lindie H. Liang. 2015. Breaking (or making) the silence: How goal interdependence and social skill predict being ostracized. *Organizational Behavior and Human Decision Processes*, 131:51 – 66.

Grounded and Ungrounded Referring Expressions in Human Dialogues: Language Mirrors Different Grounding Conditions

Eleonora Gualdoni, Raffaella Bernardi
University of Trento

eleonora.gualdoni@studenti.unitn.it
raffaella.bernardi@unitn.it

Raquel Fernández, Sandro Pezzelle
University of Amsterdam

raquel.fernandez@uva.nl
s.pezzelle@uva.nl

Abstract

We study how language use differs between dialogue partners in a visually grounded reference task when a referent is mutually identifiable by both interlocutors vs. when it is only available to one of them. In the latter case, the addressee needs to *disconfirm* a proposed description – a skill largely neglected by both the theoretical and the computational linguistics communities. We consider a number of linguistic features that we expect to vary across conditions. We then analyze their effectiveness in distinguishing among the two conditions by means of statistical tests and a feature-based classifier. Overall, we show that language mirrors different grounding conditions, paving the way to future deeper investigation of referential disconfirmation.

1 Introduction

Communication is a joint activity in which interlocutors share or synchronize aspects of their private mental states and act together in the world. To understand what our minds indeed do during communication, Brennan et al. (2010) highlight the need to study language in interpersonal coordination scenarios. When a conversation focuses on objects, interlocutors have to reach the mutual belief that the addressee has identified the discussed referent by means of visual grounding. In this frame, Clark and Wilkes-Gibbs (1986) have pointed to *referring* as a collaborative process, that requires action and coordination by both speakers and interlocutors, and that needs to be studied with a collaborative model. Clark and Wilkes-Gibbs (1986), in fact, have highlighted that – in

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

grounded condition

L: *i have grapefruit with carrots and celery*

F: *yep me too might be a blood orange though really dark*



non-grounded condition

L: *what about a guy in a suit and black hat holding a blue plaid umbrella with more of them around him*

F: *i do not have that one*



Figure 1: Examples of dialogue segments where the image referent is visible to both leader and follower (grounded condition) or only visible to the leader (non-grounded condition).

order to refer to an object in the world – speakers must believe that the referent is *mutually identifiable* to them and their addressees. This is an important skill that human speakers leverage to succeed in communication.

However, humans are not only able to identify an object described by the interlocutor – that is, *grounding* a referring expression – but also to *understand that such an object is not in the scene and, therefore, it cannot be grounded*. It can happen, indeed, that a referent is not mutually identifiable by the speakers, due to the speakers being in different grounding conditions. In this case, the addressee is able to *disconfirm a description* stated by the interlocutor by communicating that he/she does not see it (as in Figure 1). This is a crucial skill of human speakers. However, it is often neglected in the computational modelling of conversational agents.

We conjecture that the participants’ visual grounding conditions have an impact on the linguistic form and structure of their utterances. If confirmed, our hypothesis would lead to the claim that mature AI dialogue systems should learn to

master their language with the flexibility shown by humans. In particular, their language use should differ when the referred object is mutually identifiable or not. It has been shown that current AI multimodal systems are not able to decide if a visual question is answerable or not (Bhattacharya et al., 2019), and they fail to identify whether the entity to which an expression refer is present in the visual scene or not (Shekhar et al., 2017b; Shekhar et al., 2017a). We believe models can acquire this skill if they learn to play the “language game” properly.

In this paper, we investigate how the language of human conversational partners changes when they are in a mutually grounded (they both see the image they are speaking about) or non-mutually grounded setting (one sees the image while the other does not).

We find that, indeed, there are statistically significant differences along various linguistic dimensions, including utterance length, parts of speech, and the degree of concreteness of the words used. Moreover, a simple SVM classifier based on these same features is shown to be able to distinguish between the two conditions with a relatively high performance.

2 Dataset

We take the PhotoBook dataset (Haber et al., 2019) as our testbed: two participants play a game where each sees a different grid with six images showing everyday scenes.¹ Some of the images are common to both players, while others are only displayed to one of them. In each grid, three of the images are highlighted. By chatting with their dialogue partner, each player needs to decide whether each of the three highlighted images is also visible to their partner or not.

A full game consists of five rounds, and the players can decide to move to the next round when they are confident about their decisions. As the game progresses, some images may reappear in subsequent rounds. The corpus is divided into dialogue *segments*: the consecutive utterances that, as a whole, discuss a given target image and include expressions referring to it. From the set of all segments in PhotoBook, we create our dataset by focusing on segments belonging to the first round of a game (since at that point all images are new to the participants) and where a sin-

gle image is being discussed.² This results in a dataset composed of 3,777 segments paired with a given image referent and an action label indicating whether the referent is visible to both participants or only to one. The annotated dataset, together with other relevant materials, is available at: <https://dmg-photobook.github.io/>

The PhotoBook task does not impose a specific role on the players, unlike for example the Map-Task corpus (Anderson et al., 1991), where there are predefined *information giver* and *information follower* roles. In PhotoBook, the dialogues typically follow this scheme: one of the participants spontaneously decides to describe one of the images highlighted in their grid and the other participant indicates whether they also have it in their own grid or not. We call the former player the *leader* and the latter the *follower*.³ We refer to situations where the follower also sees the image described by the leader as the **grounded condition** and those where the follower does not see the image as the **non-grounded condition**. Naturally, the leader always sees the referent image.

Out of the 3,777 dialogue segments in our dataset, 1,624 belong to the grounded condition and 2,153 to the non-grounded one.

3 Linguistics Features

We hypothesize that the language used by the dialogue participants will differ in the grounded vs. non-grounded condition. To test this hypothesis, we first identify several linguistic features that we expect to vary across conditions.

Length. We expect that the length of the utterances and the overall dialogue segments may depend on the players’ possibility to see the referent. For example, in the non-grounded condition more utterances may be needed to conclude that the follower does not see the referent (thus leading to longer segments). Furthermore, not seeing the referred image could limit the expressivity of the utterances by non-grounded follower (thus leading to shorter utterances).

²We discard segments that refer to more than one image as well as those labelled with the wrong image by the original heuristics (Haber et al., 2019).

³We use simple heuristics to assign these roles a posteriori: when the image is not in common, we label as the follower the participant who does not see the image, while when the image is visible to both participants we consider the follower the player who produces the last utterance of the segment. We manually corrected the classification of the few segments that did not follow this general rule.

¹The images used in the PhotoBook task are taken from the MS COCO 2014 Trainset (Lin et al., 2014).

We compute utterance length as number of tokens per utterance and segment length as both number of tokens per segment and number of utterances per segment.

Word frequency. Frequency effects are key in psycholinguistics. Word frequency is one of the strongest predictors of processing efficiency (Monsell et al., 1989) and experiments have confirmed its link to memory performances (Yonelinas, 2002). It is plausible that different grounding conditions lead to different word choices, and that word frequency turns out to be a key aspect of this linguistic variation.

To estimate word frequency, we use off-the-shelf lemma frequency scores (frequency per million tokens) from the British National Corpus (Leech et al., 2014).⁴ For each segment in our dataset, we compute the average word frequency by first lemmatizing the words in the segment and then calculating the average frequency score for all lemma types in the segment.⁵

Concreteness. Concreteness is fundamental to human language processing since it helps to clearly convey information about the world (Hill and Korhonen, 2014). We use the *concreteness scores* by Brysbaert et al. (2014), corresponding to 40K English word lemmas, and collected via crowd-sourcing, where participants were requested to evaluate *word-concreteness* by using a 5-point rating scale ranging from abstract to concrete. We compute the average word concreteness by first lemmatizing the words in the segment and then calculating the average score for all lemma types in the segment without repetitions, divided by part-of-speech (POS).⁶

Parts of Speech distributions. Different POS differ in their function and descriptive power. We thus expect that their distribution will vary between grounded and non-grounded conditions. For example, we expect *nouns* and *adjectives* to be more likely in visually grounded referential acts, while determiners may signal whether the referent is in common ground or not (*the* vs. *a*) and give clues about the polarity of the context where they are used (*any* vs. *each*).

We extract POS distributions by first POS-tagging the utterances in the dataset⁷ and then computing the proportion of words per segment that are nouns, adjectives, verbs, or determiners, respectively. Given the different functions of different determiners, we break down this class and independently compute proportions for each of the following determiners: *a/an, the, that, those, this, these, some, all, each, any, half, both*.

4 Statistical Analysis

To test our hypothesis that the language used by the participants differs in the grounded vs. non-grounded condition, we perform a statistical analysis on our data. We compare: (1) the utterances by the *leaders* in the grounded and non-grounded conditions, and (2) the utterances by the *followers* in the grounded and non-grounded conditions. We evaluate the statistical significance of these comparisons with a Mann-Whitney U Test, which does not assume the data fits any specific distribution type. Below we report the results of each of these comparisons. Unless otherwise specified, statistical significance is tested for $p < 0.001$.

Length. Followers use significantly fewer words while leaders use significantly more words in the non-grounded condition than in the grounded condition. This trend is also illustrated in the example in Figure 1. Although followers use fewer words in the non-grounded condition, they produce a significantly higher number of utterances per segment, while no reliable differences are observed for the leaders (see Figure 2a and 2e, respectively). These findings indicate that establishing that a referring expression cannot be commonly grounded requires more evidence and more information than resolving the expression.

Frequency. Followers use significantly more high-frequency words in the grounded condition than the non-grounded condition, in particular for nouns and conjunctions. This is consistent with the reported production of more utterances per segment in the non-grounded condition, and suggests that the non-grounded follower uses them to talk about fine-grained details described by low-frequency words. In contrast, high-frequency verbs are reliably more common in the non-grounded condition (see Figure 2b).

⁴Available at <http://ucrel.lancs.ac.uk/bncfreq/flists.html>

⁵Lemmas not present in the BNC lists are ignored.

⁶Lemmas not present in the corpus are ignored.

⁷We use the NLTK Python library (Bird et al., 2009) in its “universal” tagset version.

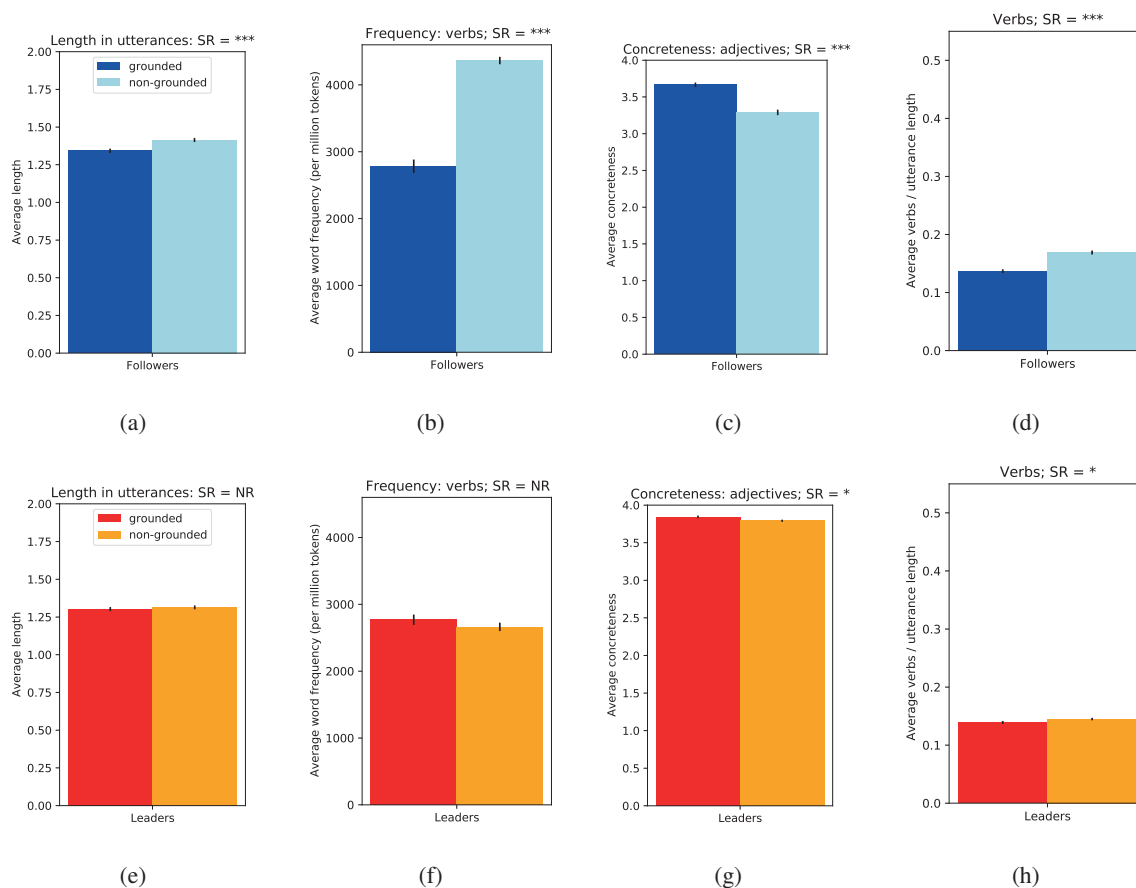


Figure 2: From left to right, difference between grounded and non-grounded condition for: (a/e) number of utterances per segment; (b/f) frequency of used verbs; (c/g) concreteness of used adjectives; (d/h) proportion of verbs. Top: followers; bottom: leaders. We use *** to refer to statistical significance at $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$; . for $p < 0.1$. Best viewed in color.

For example, note the high-frequency verbs *do* and *have* used by the non-grounded follower in Figure 1. The language of leaders, in contrast, shows marginally reliable or no difference across conditions regarding word frequency (see, e.g., the case of verbs in Figure 2f), except for high-frequency nouns and conjunctions, which are reliably more common in the grounded condition ($p < 0.01$).

Concreteness. Somehow counterintuitively, followers use overall significantly more concrete words in the non-grounded than in the grounded condition. However, an opposite pattern is found for adjectives, which usually describe the colors of the objects in the scene (see Figure 2c). This latter result is in line with our intuitions: in the non-grounded condition, followers do not have direct access to the specific perceptual properties of the entities in the image and hence use less concrete adjectives. As for the leaders, while nouns are re-

liably different, for the other POS there is either no or marginally reliable difference (see adjectives in Figure 2g, adverbs, conjunctions, and numerals) between the two conditions. This is expected since their language is always visually grounded.

Parts of speech. Followers use significantly more nouns and the determiners *a/an*, *the*, *each* in the grounded condition, while in the non-grounded condition they use significantly more verbs (see Figure 2d) and determiners *all* and *any*. That is, the grounded condition leads followers to more directly describe what they see by focusing on a specific object, as in the grounded example in Figure 1. In contrast, the non-grounded condition elicits utterances with more ‘confirmation’ verbs such as *do* and *have* and a more *vague* language signalled by the use of quantifiers, e.g., “*I don’t have any of a cake*”. As for the leaders, we observe a mixed pattern of results, though, overall, there are less reliable differences between the two

conditions compared to the followers (see the case of verbs in Figure 2h).

5 Automatic Classification

To more formally investigate the effectiveness of our selected features in distinguishing between various grounding conditions, we feed them into an SVM classifier which predicts GFC or NGFC. We run two SVM models: one for leaders, *SVM leaders*, and one for followers, *SVM followers*.⁸ Our hypothesis is that *SVM leaders* should not be very effective in the binary classification task since the language of the leaders differs only on few aspects, and less reliably between the two conditions compared to the followers'. In contrast, we expect *SVM followers* to achieve a good performance in the task, given the significant differences observed between the two conditions.

Starting from all our linguistic features (see above), we excluded those that turned out to be multicollinear in a Variance Inflation Factor test (VIF).⁹ The resulting N features (27 for the leaders, 28 for the followers), were used to build, for each datapoint, an N -dimensional vector of features that was fed into the classifier. We performed 10-fold cross-validation on the entire dataset.

Table 1 reports the accuracy, precision, recall and F1-score of the two SVM models. While *SVM leaders* is at chance level, *SVM followers* achieves a fairly high performance in the binary classification task. This indicates that our linguistic features are effective in distinguishing among the two conditions in the followers' segments. These results confirm that the language of the speakers in the *follower* role is affected by their grounding condition, and that a well-informed model is able to capture that by means of their language's linguistic features.

Table 2 reports the confusion matrices produced by our SVM models after 10-fold cross-validation. We can notice that *SVM leaders* wrongly labels NGFC datapoints as GFC in 1,381 cases, thus producing a high number of false positives. This does not happen with *SVM followers*, which is overall more accurate.

⁸We experiment with the `scikit-learn` Python library (Pedregosa et al., 2011) for C-Support Vector Classification. We use the default Radial Basis Function (`rbf`) kernel. Parameter `C` set to 100 gives the best results.

⁹The VIF test indicates whether there is a strong linear association between a predictor and the others (Pituch and Stevens, 2016). When the VIF index exceeded 10, we performed a variable deletion (Myers, 1990).

6 Related Work

Current multimodal systems are trained to process and relate modalities capturing correspondences between "sensory" information (Baltrusaitis et al., 2017). It has been shown they have trouble deciding if a question is *answerable* or not (Bhattacharya et al., 2019). Moreover, they fail to identify whether the entity to which an expression refers is present in the visual scene or not (Shekhar et al., 2017b; Shekhar et al., 2017a). Connected to this weakness is the limitation they encounter when put to work as dialogue systems, where they fail to build *common ground* from minimally-shared information (Udagawa and Aizawa, 2019). To be successful in communication, speakers are supposed to attribute mental states to their interlocutors even when they are different from their own (Rabinowitz et al., 2018; Chandrasekaran et al., 2017). This, in multimodal situations, can happen when the visual scene is only partially common between them. AI models have difficulties in such conditions (Udagawa and Aizawa, 2019).

We study how the language of conversational partners changes when (i) speakers refer to an image their interlocutor does not see and (ii) neither of the two is aware of this unshared visual ground. Though the idea that the grounding conditions of the addressees can affect their interlocutor's language is not new in psycholinguistics (Brennan et al., 2010; Brown and Dell, 1987; Lockridge and Brennan, 2002; Bard and Aylett, 2000), our approach differs from previous ones since it proposes a computational analysis of visual dialogues. Moreover, differently from other computational approaches (Bhattacharya et al., 2019; Gurari et al., 2018), we investigate scenarios where the disconfirmation of a referent's presence *is the answer* instead of suggesting a case of *unanswerability*.

7 Conclusion

Our findings confirm that, in a visually-grounded dialogue, different linguistic strategies are employed by speakers based on different grounding conditions. Our statistical analyses reliably indicate that several aspects of the language used in the conversation mirror whether the referred image is – or not – *mutually shared* by the interlocutors. Moreover, the effectiveness of a simple feature-based classifier to distinguish between the two followers' conditions further indicates that the lan-

	Accuracy	Precision			Recall			F1-score		
		GFC	NGFC	Av.	GFC	NGFC	Av.	GFC	NGFC	Av.
SVM leaders	0.57	0.15	0.89	0.40	0.50	0.58	0.55	0.23	0.70	0.50
SVM followers	0.80	0.77	0.79	0.78	0.73	0.82	0.78	0.75	0.80	0.78

Table 1: *Accuracy, Precision, Recall, and F1-score* of our SVM models, computed per class on a 10-fold cross-validation, with the corresponding weighted averages (Av.). Since our two classes (GFC and NGFC) are not balanced, chance level is 0.57.

	SVM leaders		SVM followers	
	GFC	NGFC	GFC	NGFC
GFC	243	1381	1245	379
NGFC	242	1911	461	1692

Table 2: The confusion matrices produced by our SVM models on a 10-fold cross-validation.

guage used by the speakers differs along several dimensions. We believe this capability of humans to flexibly tune their language underpins their success in communication. We suggest that efforts should be put in developing conversational AI systems that are capable to master language with a similar flexibility. This could be achieved, for example, by exposing models to one or the other condition during training to encourage them encode the relevant linguistic features. Alternatively, they should first *understand* whether the grounded information which is referred to is available to them or not. These are open challenges that we plan to tackle in future work.

Acknowledgments

EG carried out part of the work while being an ERASMUS+ visiting student at the University of Amsterdam. SP and RF are funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819455 awarded to RF).

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-

Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Ellen G Bard and MP Aylett. 2000. Accessibility, duration, and modeling the listener in spoken dialogue. In *Proceedings of the Göta 2000 Fourth Workshop on the Semantics and Pragmatics of Dialogue*.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing text with the natural language toolkit*. O’Reilly Media, Inc.

Susan E Brennan, Alexia Galati, and Anna K Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation*, volume 53, pages 301–344. Elsevier.

Paula M Brown and Gary S Dell. 1987. Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19(4):441 – 472.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of AI’s mind. *CoRR*, abs/1704.00717.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández.

2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.
- Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–731.
- Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.
- Calion Lockridge and Susan Brennan. 2002. Addressees’ needs influence speakers’ early syntactic choices. *Psychonomic bulletin & review*, 9:550–7, 10.
- Stephen Monsell, Michael C Doyle, and Patrick N Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1):43.
- Raymond H Myers. 1990. *Classical and modern regression with applications*. Duxbury, Boston, MA, 2nd edition.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Keenan A Pituch and James P Stevens. 2016. *Applied Multivariate Statistics for the Social Sciences*. Routledge, 6th edition.
- Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227.
- Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: Moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.
- Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. *CoRR*, abs/1907.03399.
- Andrew P Yonelinas. 2002. The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3):441–517.

Predicting movie-elicited emotions from dialogue in screenplay text: A study on “Forrest Gump”

Benedetta Iavarone^{*◊}, Felice Dell’Orletta[◊]

^{*} Scuola Normale Superiore, Pisa

[◊]ItaliaNLP Lab, Istituto di Linguistica Computazionale “Antonio Zampolli”, Pisa
benedetta.iavarone@sns.it, felice.dellorletta@ilc.cnr.it

Abstract

We present a new dataset of sentences¹ extracted from the movie *Forrest Gump*, annotated with the emotions perceived by a group of subjects while watching the movie. We run experiments to predict these emotions using two classifiers, one based on a Support Vector Machine with linguistic and lexical features, the other based on BERT. The experiments showed that contextual embeddings are effective in predicting human-perceived emotions.

1 Introduction

Emotional intelligence, described as the set of skills that contributes to the accurate appraisal, expression and regulation of emotions in oneself and in others (Salovey and Mayer, 1990), is recognised to be one of the facets that make us humans and the fundamental ability of human-like intelligence (Goleman, 2006). Emotional intelligence has played a crucial role in numerous applications during the last years (Krakovsky, 2018), and being able to pinpoint expressions of human emotions is essential to advance further in technological innovation. Emotions can be identified in many sources, among which there are semantics and sentiment in texts (Calefato et al., 2017). In NLP, Sentiment Analysis already boasts many state-of-the-art tools that can accurately predict or classify the polarity of a text. However, real applications often need to go beyond the dichotomy positive-negative and identify the emotional content of a text with a finer granularity. Nevertheless, the task of predicting a precise emotion from text brings many challenges, mostly because there is a need of context: emotions can’t be easily understood in isolation, as

¹Data can be downloaded at www.italianlp.it/dataset_release.zip

they are conveyed by a complex of explicit (e.g. speech) and implicit (e.g. gesture and posture) behavioural cues. Still, there has been an increasing interest in research for text-based emotion detection (Acheampong et al., 2020). In this work, we study how textual information extracted from the screenplay of a movie can be used to predict the emotions perceived by a group of people during the view of the movie itself. We create a new dataset of sentences extracted from the screenplay, annotated with six different perceived emotions and their perceived intensity and create a binary classification task to predict emotional elicitation during the view of the movie. We use two predicting models, with different kind of features that capture diverse language information. We determine which model and which kind of features are the best for predicting the emotions perceived by the subjects.

2 Data

Our dataset was retrieved from *studyforrest*², a research project centered around the use of the movie *Forrest Gump*. The project repository contains data contributions from various research groups, divided in three areas: (i) behavior and brain function, (ii) brain structure and connectivity, and (iii) movie stimulus annotations. We focused on the latter, retrieving two types of data: the speech present in the movie and the emotions that the vision of the movie elicited in a group of subjects. As for the speech, each screenplay line pronounced by the characters is transcribed in sentences and associated with two timestamps in terms of tenths of a second t_{begin} and t_{end} , that respectively indicate the moment of the movie in which the character starts talking and the moment in which they stop. Emotional data comes from the contribution to the project given by Lettieri et al. (2019). A group of 12 subjects was asked to watch the movie and

²<http://studyforrest.org/>

Subject	Happiness	Surprise	Fear	Sadness	Anger	Disgust	Neutral	Emotion
1	592	172	101	557	111	166	22	876
2	628	87	83	539	120	42	61	837
3	345	471	212	340	123	37	30	868
4	274	179	137	255	119	133	276	622
5	244	84	98	224	83	6	305	593
6	496	92	147	264	60	13	113	785
7	277	255	88	132	88	23	286	612
8	357	218	119	305	103	77	231	667
9	299	389	15	147	109	22	312	586
10	213	125	81	255	60	0	377	521
11	352	320	116	307	150	30	120	778
12	180	36	22	149	34	25	526	372
Total	4257	2428	1219	3474	1160	574	2659	8117

Table 1: Emotions distribution in the dataset.

report the emotions they were experiencing during the vision, among a list of six emotions (happiness, surprise, fear, sadness, anger, disgust). Emotion reporting was performed by pressing the keys of a keyboard, with which subjects could indicate the emotion they were experiencing and its intensity, within a range from 0 (no emotion) to 100.

2.1 Data creation

Emotional data was collected from a continuous output $\mathbf{z} = (z_1, z_2, \dots, z_n)$ from the keyboard, such that each z_i corresponds to an increment of 0.1 seconds in the playing time of the movie ($z_i = 0.1, z_{i+1} = 0.2, z_{i+2} = 0.3, \dots$). Each z_i is associated to a list $x_{i1}, x_{i2}, \dots, x_{ij}$, with $x_j \in [0, 100]$ and $j \in [\text{happiness}, \text{surprise}, \text{fear}, \text{sadness}, \text{anger}, \text{disgust}]$, where each x_j indicates the intensity that one emotion assumes at a given timestamp. For our purpose, this information was too detailed and it could not be mapped to textual data properly, thus we proceeded to resample emotional information. We generated new timestamps $\mathbf{s} = (s_1, s_2, \dots, s_m)$, such that each s_i corresponds to the sum of 20 consecutive z_i , thus to an increment of 2 seconds in the playing time of the movie. Each s_i is associated to a new list of emotional values, where each new value is the average of the values associated to the summed z_i .

After resampling, we aligned the text to emotional data. As one of our aims is to determine how much text is needed for accurate emotion prediction, we considered three progressively larger time windows for each s_k , such that $\text{window}_i = [s_k - m, s_k]$, where $m = (2, 4, 6)$. For each sentence, we retrieve its t_{end} and align the sentence verifying if $s_k - m \leq t_{\text{end}} \leq s_k$, thus checking if the moment in which the sentence ends falls within the given time window. In this way, the larger the

time window, the larger the amount of text that gets aligned with a specific timestamp. With this process, we created three different datasets, one for each time window. We then removed all the lines in which no text was aligned to s_k . For each dataset, we end up having 898 timestamps associated with a line of text and 6 emotion declarations for each of the 12 subjects.

2.2 Data statistics and data selection

We first looked at the distribution of our data, examining how many times each subject declared a specific emotion. Whenever the subject assigned a value different than zero to a certain emotion, we considered that emotion as present at a given timestamp, regardless of its intensity. If all 6 emotions were zero at the same time (all $x_j = 0$), we assigned to that case the class *neutral*. Furthermore, if any given emotion was declared (at least one $x_j \neq 0$), we assigned to that case the class *emotion*, to indicate a generic emotional response.

As shown in Table 1, the most represented emotions in the dataset are happiness and sadness, while the others are underrepresented. Table 1 also shows that emotions distribution is quite uneven among the different subjects, as there were some subjects that declared emotions frequently and others that entered fewer declarations. This is due to the fact that emotive phenomena are strongly subjective, meaning that emotion processing is specific to each person and that everyone experiences emotions at a different granularity (Barrett, 2006). To account for this factor, we measured the level of agreement between the 12 subjects using *Fleiss' Kappa*. Table 2 reports the percentage of agreement for each emotion in the data. The lowest agreement was found on surprise and disgust. As disgust is also the less declared emotion, it is fair to assume

Emotion	Agreement
happiness	0.32
surprise	0.14
fear	0.41
sadness	0.31
anger	0.42
disgust	0.17

Table 2: Annotators agreement (Fleiss’ Kappa) on all emotions

that the movie does not contain many moments that elicit this emotion in the subjects. On the other side, the strongest agreement is found on fear and anger, showing that these emotions are evoked in specific scenes of the movie and that subjects had a similar emotional response to those scenes. In Table 3 we report examples of sentences on which the subjects agreed the most, for all six emotions. For every emotion, there are many sentences on which a large number of subjects agreed, meaning that there were various moments of the movie that elicited the same emotions in the subjects. In the case of disgust, the highest level of agreement was achieved at 8 subjects, only on one sentence. There were no other sentences for which 8 subjects (or more) agreed. This is justified by the fact that disgust is the less represented emotion in the data.

Given the information on the agreement and on emotions distribution, we decided not to examine underrepresented emotions directly, even if their agreement was strong (i.e. surprise). In order to still account for underrepresented emotions, we relied on the general class *emotion*. Hence we assessed three different scenarios: (i) the presence of any kind of emotion (at least one $x_j \neq 0$), (ii) the presence of happiness ($x_{happiness} \neq 0$) and (iii) the presence of sadness ($x_{sadness} \neq 0$). Furthermore, we decided to conduct our experiments only on two subjects, subject 4 and subject 8. We focused on these specific subjects as they declared all emotions evenly, without neglecting any of them, and because the number of declarations for each emotion was quite similar between the two subjects.

3 Emotions prediction

We evaluated the three scenarios described in 2.2 in contrast to the absence of any emotion (all $x_j = 0$), producing three binary classification tasks. We relied on two sets of features: automatically extracted linguistic and lexical features, and contextual word embeddings from a language model.

Emotion	N subs	Text
happiness	12	I had never seen anything so beautiful in my life. She was like an angel.
surprise	11	Jenny! Forrest!
fear	12	(into radio) Ah, Jesus! My unit is down hard and hurting! 6 pulling back to the blue line, Leg Lima 6 out! Pull back! Pull back!
sadness	12	Bubba was my best good friend. And even I know that ain’t something you can find just around the corner. Bubba was gonna be a shrimp in’ Boat captain, But instead he died right there by that river in Vietnam.
anger	12	Are you retarded, Or just plain stupid? Look, I’m Forrest Gump.
disgust	8	You don’t say much, do you?

Table 3: Examples of sentences on which subjects agreed the most, for all emotions.

3.1 Prediction with linguistic and lexical features

For the first set of features, sentences were first POS tagged and parsed using UDPipe (Straka and Straková, 2017). We extracted a wide set of features, like the ones described in Brunato et al. (2020). These features capture various linguistic phenomena, that range from raw information to information related to the morpho-syntactic and syntactic structure of the sentence (rows 1, 2 and 3 in Table 4, hereafter *linguistic* features). Additionally, we extracted other features that are able to capture some lexical information (row 4 in Table 4, hereafter *lexical* features), as they identify set of characters or words that appear more frequently within a sentence. We trained two SVM models, one on the linguistic features (*SVMling*), one on the lexical features (*SVMlex*). We trained the models with a linear kernel and standard parameters, performing 10-cross-fold validation to evaluate the models accuracy.

3.2 Prediction with language model

For the second set of features, we relied on BERT (Devlin et al., 2019), a Neural Language Model that encodes contextual information. We retrieved the pre-trained base model and fine tuned it on our data. The pre-trained BERT model already includes a lot of information about the language, as it has already been trained on a large amount of data. By fine tuning it on our data, we are able to exploit the information already acquired by the model and use it for our task. We performed differ-

ent fine tuning stages, then used the so fine-tuned models to perform the binary classification task on our data. We evaluated model accuracy using 10 cross-fold-validation. Specifically, we tested three different fine tuning approaches: (1) original data (*BERTorig*), (2) oversampled data to balance the neutral class (*BERTover*), (3) oversampled data + transfer learning tuning (*BERTtransf*). In the case of (3), we first fine tuned the model on data different than ours but conceived for a similar task. Notably, we relied on data created for SemEval-2018 Task 1E-c (Mohammad et al., 2018), containing tweets annotated with 11 emotion classes. After this first tuning, we tuned the model again on our oversampled data and proceeded with the classification task.

4 Results and discussion

Figure 1 shows the accuracy scores for all the models, for both subjects and the three datasets. In all cases, the baseline was determined with a majority classifier. The results appear similar for both subjects.

SVM models are always outperformed by BERT ones. In any case, *SVMling* is the model that gave the lowest performance, remaining below or around the baseline value. On the contrary, *SVMlex* tends to bring a higher performance, despite remaining close to the baseline in most cases. On one side, this is due to the fact that features that look at the raw, morpho-syntactic and syntactic aspects of text, do not encode any relevant information regarding the emotional cues in the text. *SVMlex* always performs better than *SVMling* because lexical features look at patterns of words and characters that are repeated in the input text and thus record information about the lexicon of the dataset. However, as our dataset is too small, it is hard for the model to retrieve the same lexical patterns in both the training and test set.

BERT models outperform the SVM ones in both happiness and sadness prediction. In the case of emotion prediction, BERT models obtain very good results only on the 6 seconds dataset. This is due to the fact that, in this case, we have flattened all emotions into a single category, thus it may be difficult for the model to distinguish between general emotionally charged sentences and those that are not perceived as emotionally charged. When emotions are specific and clearly separated, as in happiness and sadness cases, BERT is able to infer the per-

Level of Annotation	Feature
Raw Text	Sentence length
	Word length
	Type/Token Ratio for words and lemmas
POS tagging	Distribution of POS
	Lexical density
	Inflectional morphology of lexical verbs and auxiliaries (Mood, Number, Person, Tense and VerbForm)
Dependency Parsing	Depth of the whole syntactic tree
	Average length of dependency links and of the longest link
	Average length of prepositional chains and distribution by depth
	Clause length (n. tokens/verbal heads)
	Order of subject and object
	Distribution of verbs by arity
	Distribution of verbal heads and verbal roots
	Distribution of dependency relations
	Distribution of subordinate and principal clauses
	Average length of subordination chains and distribution by depth
	Relative order of subordinate clauses
Lexical Patterns	Bigrams, trigrams and quadrigrams of characters, words and lemmas

Table 4: Linguistic and Lexical Features.

ceived emotions even from small amounts of text (2 and 4 seconds datasets). *BERTover* and *BERTtransf* tend to give better performances than what happens with *BERTorig*. In the case of *BERTover*, there is a very slight difference in the prediction of happiness and sadness, as in these cases the classes to be predicted were distributed quite evenly. In the case of emotion prediction, the model is helped by the higher representation of the neutral class. With *BERTtransf*, the performances stay in line with the ones obtained with the bare oversampling. Fine tuning the model on similar data did not add any more useful information. This is due to the fact that SemEval data were too distant from the ones of our dataset. Therefore, even though the task is similar to ours, the input text is too different from our sentences to actually make a huge difference for the prediction. We also tried another form of transfer learning, tuning the model on one subject and testing it on the other one. However, the results were too low and we did not report them. This is because emotion perception is a very personal phenomenon and it cannot be easily generalised to different individuals.

To further evaluate the results, we computed the percentage of agreement between the two models that overall had the best performances, *BERTover*

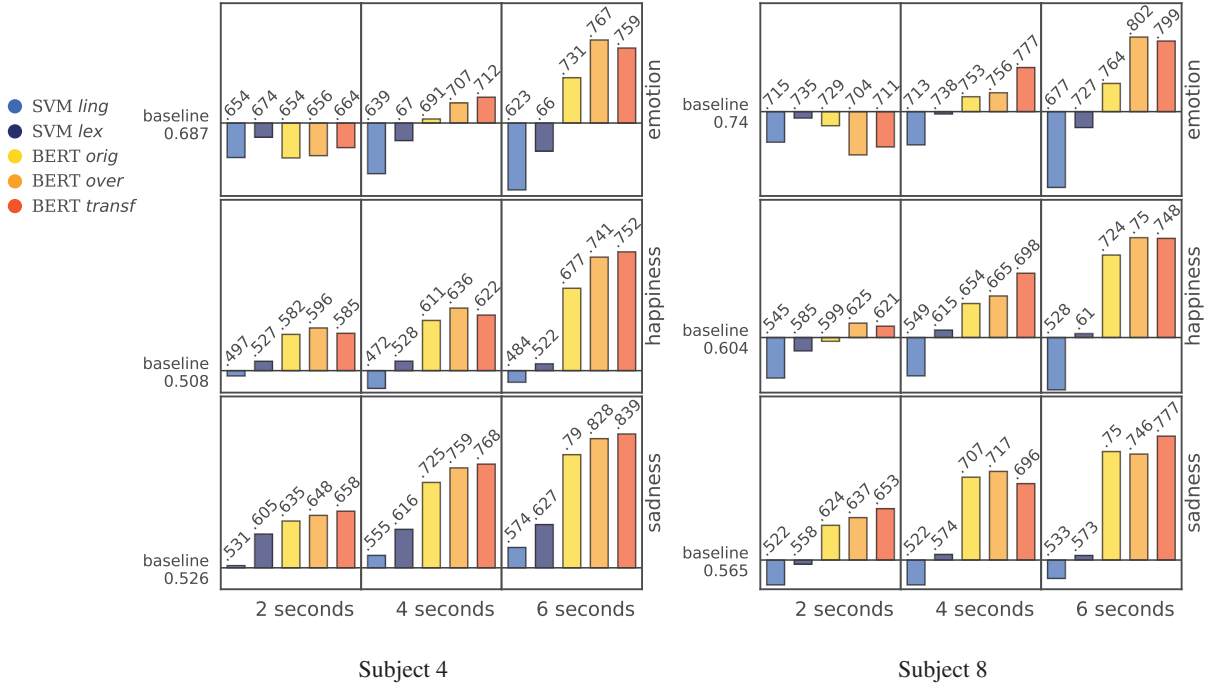


Figure 1: Performances (accuracy) of SVM and BERT models in the prediction of emotion, happiness of sadness, for every timespan window, and for both subject 4 and subject 8.

	subject 4			subject 8		
	2sec	4sec	6sec	2sec	4sec	6sec
emo.	82.75	83.63	85.82	82.03	87.8	90.44
hap.	70.77	72.64	79.78	76.26	72.31	79.67
sad.	82.53	85.93	87.47	80.44	79.45	85.05

Table 5: Agreement (%) between *BERTover* and *BERTtransf* predictions.

and *BERTtransf*. We defined agreement as the percentage of sentences for which the models gave the same output during the classification task. Table 5 reports the results for emotion, happiness and sadness, for every timespan window, and for both subjects 4 and subject 8. The agreement is quite high in all cases, and it tends to get stronger with the amount of text on which models are trained (i.e. 6 seconds). A higher level of agreement indicates that the models have similar behaviour, thus making the same mistakes in the classification task. The lowest levels of agreement are encountered on the classification of happiness, showing that the two models work differently in this part of the task. Indeed, both *BERTover* and *BERTtransf* obtain high performances in predicting happiness, but the fact that their agreement is lower suggests that they differ in the mistakes they make in the classification. We may exploit this information to create systems that combine different classifiers, actually enhancing the classification accuracy. By doing this, it

is possible to compare the cases in which two or more classifiers agree and the cases in which they make mistakes, thus choosing the best classification output accordingly.

5 Conclusion

In this paper, we presented a dataset of sentences extracted from the movie *Forrest Gump*, annotated with the emotions that a group of subjects perceived while watching the movie, and we studied how to predict these emotions. To do so, we retrieved different kinds of features from the sentences pronounced by the characters of the movie. We showed that contextual embeddings extracted from the sentences can accurately predict specific emotions, even if the amount of text used for the prediction is very little. Instead, when predicting generic emotional elicitation, a larger amount of text is required for an accurate prediction. We also show that lexical, morpho-syntactic and syntactic aspects of the sentences cannot be used to infer emotional elicitation during the view of the movie.

As emotional response is directly correlated with brain activity, we plan to add fMRI images recorded during the vision of the movie to the contextual embedding we extracted. In this way, we could verify if brain images can help to increase the accuracy in the prediction of perceived emotions.

Acknowledgments

We thank MoMiLab research group of IMT Lucca for having shared with us the data they collected on human-perceived emotions. Furthermore, we are grateful to the *studyforrest* project and all its contributors.

References

- [Acheampong et al.2020] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, page e12189.
- [Barrett2006] Lisa Feldman Barrett. 2006. Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40(1):35–55.
- [Brunato et al.2020] Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.
- [Calefato et al.2017] Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. Emotxt: a toolkit for emotion recognition from text. In *2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE.
- [Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- [Goleman2006] Daniel Goleman. 2006. *Emotional intelligence*. Bantam.
- [Krakovsky2018] Marina Krakovsky. 2018. Artificial (emotional) intelligence.
- [Lettieri et al.2019] Giada Lettieri, Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Paolo Papale, Monica Betta, Pietro Pietrini, and Luca Cecchetti. 2019. Emotionotopy in the human right temporo-parietal cortex. *Nature communications*, 10(1):1–13.
- [Mohammad et al.2018] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.
- [Salovey and Mayer1990] Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.
- [Straka and Straková2017] Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Point Break: Surfing Heterogeneous Data for Subtitle Segmentation

Alina Karakanta^{1,2}, Matteo Negri¹, Marco Turchi¹

¹ Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento - Italy

² University of Trento, Italy

{akarakanta, negri, turchi}@fbk.eu

Abstract

Subtitles, in order to achieve their purpose of transmitting information, need to be easily readable. The segmentation of subtitles into phrases or linguistic units is key to their readability and comprehension. However, automatically segmenting a sentence into subtitles is a challenging task and data containing reliable human segmentation decisions are often scarce. In this paper, we leverage data with noisy segmentation from large subtitle corpora and combine them with smaller amounts of high-quality data in order to train models which perform automatic segmentation of a sentence into subtitles. We show that even a minimum amount of reliable data can lead to readable subtitles and that quality is more important than quantity for the task of subtitle segmentation.¹

1 Introduction

In a world dominated by screens, subtitles are a vital means for facilitating access to information for diverse audiences. Subtitles are classified as interlingual (subtitles in a different language as the original video) and intralingual (of the same language as the original video) (Bartoll, 2004). Viewers normally resort to interlingual subtitles because they do not speak the language of the original video, while intralingual subtitles (also called captions) are used by people who cannot rely solely on the original audio for comprehension. Such viewers are, for example, the deaf and hard of hearing and language learners. Apart from creating a bridge towards information, entertainment and education, subtitles are a means to im-

proving the reading skills of children and immigrants (Gottlieb, 2004). Having such a large pool of users and covering a wide variety of functions, subtitling is probably the most dominant form of Audiovisual Translation.

Subtitles, however, in order to fulfil their purposes as described above, need to be presented on the screen in a way that facilitates readability and comprehension. Bartoll and Tejerina (2010) claim that subtitles which cannot be read or can be read only with difficulty ‘*are almost as bad as no subtitles at all*’. Creating readable subtitles comes with several challenges. The difficulty imposed by the transition to a different semiotic means, which takes place when transcribing or translating the original audio into text, is further exacerbated by the limitations of the medium (time and space on screen). Subtitles should not exceed a maximum length, usually ranging between 35-46 characters, depending on screen size and audience age or preferences. They should also be presented at a comfortable reading speed for the viewer. Moreover, chunking or segmentation, i.e. the way a subtitle is split across the screen, has a great impact on comprehension. Studies have shown that a proper segmentation can balance gazing behaviour and subtitle reading (Perego, 2008; Rajendran et al., 2013). Each subtitle should – if possible – have a logical completion. This is equivalent to a segmentation by phrase, sentence or unit of information. Where and if to insert a subtitle break depends on several factors such as speech rhythm, pauses but also semantic and syntactic properties. This all makes segmenting a full sentence into subtitles a complex and challenging problem.

Developing automatic solutions for subtitle segmentation has long been impeded by the lack of representative data. Line breaks are the new lines inside a subtitle block, which are used to split a long subtitle into two shorter lines. This type of breaks is not present in the subtitle files used

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to create large subtitling corpora such as OpenSubtitles (Lison and Tiedemann, 2016) and corpora based on TED Talks (Cettolo et al., 2012; Di Gangi et al., 2019), possibly because of encoding issues and the pre-processing of the subtitles into parallel sentences (Karakanta et al., 2019). Recently, MuST-Cinema (Karakanta et al., 2020b), a corpus based on TED Talks, was released, which added the missing line breaks from the subtitle files (.srt²) using an automatic annotation procedure. This makes MuST-Cinema a high-quality resource for the task of subtitle segmentation. However, the size of MuST-Cinema (about 270k sentences) might not be sufficient for developing automatic solutions based on data-hungry neural-network approaches, and its language coverage is so far limited to 7 languages. On the other hand, the OpenSubtitles corpus, despite being rather noisy, constitutes a large resource of subtitling data.

In this work, we leverage available subtitling resources in different resource conditions to train models which automatically segment sentences into readable subtitles. The goal is to exploit the advantages of the available resources, i.e. size for OpenSubtitles and quality for MuST-Cinema, for maximising segmentation performance, but also taking into account training efficiency and cost. We experiment with a sequence-to-sequence model, which we train and fine-tune on different amounts of data. More specifically, we hypothesise the condition where data containing high-quality segmentation decisions is scarce or non-existent and we resort to existing resources (OpenSubtitles). We show that high-quality data, representative of the task, even in small amounts, are a key to finding the break points for readable subtitles.

2 Related work

Automatically segmenting text into subtitles has long been addressed as a post-processing step in a translation/transcription pipeline. In industry, language-specific rules and simple algorithms are employed for this purpose. Most academic approaches on subtitle segmentation make use of a classifier which predicts subtitle breaks. One of these approaches used Support Vector Machine and Logistic Regression classifiers on correctly/incorrectly segmented subtitles to deter-

mine subtitle breaks (Álvarez et al., 2014). Extending this work, Álvarez et al. (2017) trained a Conditional Random Field (CRF) classifier for the same task, but in this case making a distinction between line breaks (next subtitle line) and subtitle breaks (next subtitle block). A more recent, neural-based approach (Song et al., 2019) employed a Long-Short Term Memory Network (LSTM) to predict the position of the period in order to improve the readability of automatically generated Youtube captions, but without focusing specifically on the segmentation of subtitles. Focusing on the length constraint, Liu et al. (2020) proposed adapting an Automatic Speech Recognition (ASR) system to incorporate transcription and text compression, with a view to generating more readable subtitles.

A recent line of works has paved the way for Neural Machine Translation systems which generate translations segmented into subtitles, here in a bilingual scenario. Matusov et al. (2019) customised an NMT system to subtitles and introduced a segmentation module based on human segmentation decisions trained on OpenSubtitles and penalties well established in the subtitling industry. Karakanta et al. (2020a) were the first to propose an end-to-end solution for Speech Translation into subtitles. Their findings indicated the importance of prosody, and more specifically pauses, to achieving subtitle segmentation in line with the speech rhythm. They further confirmed the different roles of line breaks (new line inside a subtitle block) and subtitle block breaks (the next subtitle appears on a new screen); while block breaks depend on speech rhythm, line breaks follow syntactic patterns. All this shows that subtitle segmentation is a complex and dynamic process and depends on several and varied factors.

3 Methodology

This section describes the data processing, model and evaluation used for the experiments. All experiments are run for English, as the language with the largest amount of available resources, but the approach is easily extended to all languages. Note that here we are focusing on a monolingual scenario, where subtitle segmentation is seen as a sequence-to-sequence task of passing from English sentences without break symbols to English sentences containing break symbols.

²<http://zuggu.wz.cz/>

3.1 Data

As training data we use MuST-Cinema and OpenSubtitles. MuST-Cinema contains special symbols to indicate the breaks: `<eob>` for subtitle breaks and `<eol>` for line breaks inside a subtitle block. We train models using all data (*MC-all*) and only 100k sentences (*MC-100*).³

The monolingual files for OpenSubtitles come in XML format, where each subtitle block forming a sentence is wrapped in XML tags. We are therefore able to insert the `<eob>` symbols for determining the end of a subtitle block. However, we mentioned that line breaks are not present in OpenSubtitles. We hence proceed to creating artificial annotations for `<eol>`. We filter all sentences for which all subtitles have a maximum length of 42 characters (*OpenSubs-42*). Then, for each `<eob>`, we substitute it with `<eol>` with a probability of 0.25, making sure to avoid having two consecutive `<eol>`, as this would lead to a subtitle of three lines, which occupies too much space on the screen. Since this length constraint results in filtering out a lot of data, we also relax the length constraint by allowing sentences with subtitles with up to 48 characters (*OpenSubs-48*). The motivation for this relaxation is that, if a sequence-to-sequence model is not able to learn the constraint of length from the data but instead learns segmentation decisions based on patterns of neighbouring words, having more data will increase the amount and variety of segmentation decisions observed by the model. This may result in more plausible segmentation, possibly though to the expense of length conformity. Dataset sizes are reported in Table 1.

We are interested in the real application scenario where high-quality data containing human segmentation decisions are not available or scarce. According to our hypothesis, a relatively limited size of high-quality data can be compensated by OpenSubtitles. Therefore, we fine-tune each of the OpenSubtitle models on 10k and 100k sentences from MuST-Cinema, which contain high-quality break annotations.

OpenSubtitles and TED Talks have been shown to have large differences and to constitute a sub-classification of the subtitling genre (Müller and Volk, 2013). For this reason, we experiment with 2 test sets for cross-domain evaluation. The first

³Training a model with 10k data did not bring good results.

Data	Sents
MuST-Cinema	275,085
OpenSubs-42	185,758
OpenSubs-48	13,713,708

Table 1: Dataset sizes in sentences.

set is the English test set released with MuST-Cinema, containing 10 single-speaker TED Talks (545 sentences). The second test set (782 sentences) is much more diverse. In order to create it, we have selected a mix of public and proprietary data, more specifically, excerpts from a TV series, a documentary, two short interviews and one advertising video. The subtitling was performed by professional translators and the .srt files were processed to insert the break symbols in the positions where subtitle and line breaks occur.

3.2 Model

The model is a sequence-to-sequence model based on the Transformer architecture (Vaswani et al., 2017), trained using fairseq (Ott et al., 2019) with the same settings as in Karakanta et al. (2020b). It takes as input a full sentence and returns the same sentence annotated with subtitle and line breaks. We process the data into sub-word units with SentencePiece (Kudo and Richardson, 2018) with 8K vocabulary size. The special symbols are kept as a single sub-word. Models were trained until convergence, on 1 Nvidia GeForce GTX1080Ti GPU.

As baseline, we use a simple segmentation approach inserting a break symbol at the first space before every 42 characters. From the two types of symbols, `<eol>` is selected with a 0.25 probability, but we avoid inserting two consecutive `<eol>`, since this would lead to a subtitle of three lines.

3.3 Evaluation

Evaluating the subtitle segmentation is performed with the following metrics. First, we compute the precision, recall and F1-score between the output of the segmenter and the human generated subtitles in order to test the model’s performance at inserting a sufficient number of breaks and at the right positions in the sentence. Additionally, we compute the BLEU score (Papineni et al., 2002) between the output of the segmenter and the human reference. Higher values for BLEU indicate a high similarity between the model’s and desired output.

Model	BLEU	Prec	Rec	F1	CPL	Time
baseline	55.30	50	47	48	100	-
MC-all	84.00	85	85	85	96	305
MC-100	81.77	84	83	83	94	210
OpenSubs-42	72.24	86	66	73	74	270
MC-10	77.99	83	76	79	88	+26
MC-100	80.09	87	78	81	88	+250
OpenSubs-48	76.00	77	67	68	72	6980
MC-10	82.46	86	80	82	91	+240

Table 2: Results for the MuST-Cinema test set. Training time in minutes.

Model	BLEU	Prec	Rec	F1	CPL	Time
baseline	51.45	46	43	44	100	-
MC-all	66.38	72	64	69	97	305
MC-100	65.38	76	64	68	96	210
OpenSubs-42	61.41	84	56	65	79	270
MC-10	63.53	76	60	66	93	+26
MC-100	65.3	77	62	67	94	+250
OpenSubs-48	63.37	63	56	59	81	6980
MC-10	65.66	78	61	67	94	+240

Table 3: Results for the second test set. Training time in minutes.

Finally, we want to check the performance of the system in generating readable subtitles, therefore, we use an intrinsic, task-specific metric. We compute the number of subtitles with a length of ≤ 42 characters (Characters per Line - CPL), according to the TED subtitling guidelines. This shows the ability of the system to segment the sentences into readable subtitles, by producing subtitles that are not too long to appear on the screen. We additionally report training time, as efficiency and cost are important factors for scaling such methods to tens of languages.

4 Results

Tables 2 and 3 show the results for the MuST-Cinema and the second test set respectively. As expected, the simple baseline achieves a 100% conformity to the length constraint, it is however not accurate in inserting the breaks at the right positions, as shown by the very low BLEU (55.30 and 51.45) and F1 scores (48 and 44). The best performance for all metrics and both test sets is achieved when using all available MuST-Cinema data (*MC-all*). For the in-domain test set, BLEU and F1 are higher than for the out-of-domain test set, however the number of subtitles conforming to the length constraint is consistently high (96% and 97%). This suggests that the systems trained on high-quality segmentation are able to produce

readable subtitles in terms of length in diverse testing conditions even without massive amounts of data. Even with 100k of training data (*MC-100*) the performance of the model, which is the fastest model to train, drops only slightly, with -2% for all metrics on the MuST-Cinema test set and -1% on the second test set. This shows that high efficiency can be achieved without dramatically sacrificing quality. This is particularly important for industry applications where tens of languages are involved and training data for a domain might not be vast.

The models trained only on OpenSubtitles show a great drop in performance for the MuST-Cinema test, which is to be expected because of the different nature of the data. However, the drop is present also for the second test set, which shows that these models are not robust to different domains. Surprisingly, the larger model (*OpenSubs-48*) does not perform much better than the model with less data (*OpenSubs-42*) even though it is trained on almost 10 times as much data. This could be an indication of a trade-off between data quality and data size. *OpenSubs-48* with more noisy data has similar recall to *OpenSubs-42*, but it is much less accurate in the position of the breaks, as shown by the drop in precision (86 vs. 77 and 84 vs. 63). We conjecture that the procedure of artificially inserting `<eol>` symbols by changing the existing `<eob>` does not reflect the distribution of the type of breaks in real data. Interestingly, the *OpenSubs-42* model, despite containing only subtitles of a maximum length of 42, is not able to generate subtitles which respect the length constraint (74% and 79%). It is therefore possible that the segmenter does not learn to take into consideration the constraint of length, but the segmentation decisions are based on lexical patterns in the data, as also suggested by Karakanta et al. (2020a).

Fine-tuning, even on a minimum amount of real data, as shown when fine-tuning on 10k of MuST-Cinema, can significantly boost the performance compared to the OpenSubtitles models and is a viable and fast solution towards readable subtitles. This corroborates the claim in favour of creating datasets which are representative of the task at hand. Surprisingly though, fine-tuning the *OpenSubs-42* model on MC-100 does not improve over training the model from scratch on MC-100 for neither test set. For the case when only a small amount of MuST-Cinema data is available (MC-

10), having a larger base model on which to fine-tune (*OpenSubs-48*) is beneficial, since there is an improvement for all metrics and in both testing conditions compared to all other models trained on OpenSubtitles or fine-tuned on them. Therefore, we conclude that, in the presence of little data containing human segmentation decisions, a model trained on more data, even though possibly noisier, is a more robust base on which to fine-tune using the high-quality data. One considerable drawback is that the improvement comes at a training time of x25 over the other base model (*OpenSubs-42*), which raises significant considerations for cost and efficiency. Such a model however, once trained, could be re-used for fine-tuning on several domains and for different client specifications.

5 Analysis and Discussion

We further perform a manual inspection to identify issues related to the models. We hypothesise that low precision is connected to over-splitting or splitting in wrong positions, while low recall suggests under-splitting (not inserting a sufficient number of breaks). Indeed, we observe that the OpenSubtitle models tend to over-segment short sentences, but under-segment longer sentences:

Reference:

Let's turn our attention to the hows. <eob>
(37 characters)

OpenSubs-42:

Let's turn our attention <eol>
to the hows. <eob> (25 + 12 characters)

Reference:

My family's traditions <eol>
and expectations for a woman <eob>
wouldn't allow me to own a mobile <eol>
phone until I was married. <eob>
(22 + 28 + 39 + 20 characters)

OpenSubs-42:

My family's traditions and expectations
<eol>
for a woman wouldn't allow me to own a mobile
phone until I was married. <eob>
(39+72 characters)

In the following example, fine-tuning on MC increases length conformity, splitting the first subti-

tle in two, while MC-100k succeeds in segmenting all subtitles exceeding 42 characters, matching the reference segmentation.

Reference:

Meditation is a technique <eol>
of finding well-being <eob>
in the present moment <eol>
before anything happens. <eob>

OpenSubs-42:

Meditation is a technique of finding well-being <eob>
in the present moment before anything happens. <eob>
(47+46 characters)

OpenSubs-42 + MC 10K:

Meditation is a technique <eol>
of finding well-being <eob>
in the present moment before anything happens. <eob>
(25+21+46 characters)

MC-100K: Meditation is a technique <eol>
of finding well-being <eob>
in the present moment <eol>
before anything happens. <eob>

The examples above confirm our results which showed that the models do not explicitly learn the constraint of length, but rather patterns of segmentation. From a syntactic point of view, the break symbols are inserted after a noun (e.g. attention, expectations) and before a preposition/conjunction (to, for, in, before), regardless of the model. The break symbols, even though do not overlap with the human segmentation decisions, are inserted at plausible positions. This leads in subtitles that present logical completion, *i.e.* each subtitle is formed by a phrase or syntactic unit, even though they do not respect the constraint of length. The conformity to the length constraint seems to be forced only with the high-quality MuST-Cinema data. It is possible that the artificial break symbols in OpenSubtitles clash with the real break symbols in MuST-Cinema, which creates confusion for the model. Replacing some <eob> with <eol> symbols in OpenSubtitles to simulate data where human-annotated line breaks exist means that the models trained on OpenSubtitles observe a line break at positions where normally a subtitle break is present. Given the different functions of the two types of breaks, this is a possible

explanation why fine-tuning OpenSubtitles-42 on MC-100 performs worse than training on MC-100 from scratch and provides us with insights on future design of artificial segmentation decisions to augment subtitling data.

6 Conclusion

We have presented methods to combine heterogeneous subtitling data in order to improve automatic segmentation of subtitles. We leverage large data containing noisy segmentation decisions from OpenSubtitles and combine them with smaller amounts of high-quality data from MuST-Cinema to generate readable subtitles from full sentences. We found that even limited data with reliable segmentation can improve performance. We conclude that quality matters more than size for determining the break points between subtitles.

Acknowledgments

This work is part of the “End-to-end Spoken Language Translation in Rich Data Conditions” project,⁴ which is financially supported by an Amazon AWS ML Grant.

References

- Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. 2014. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238, Cham. Springer International Publishing.
- Aitor Álvarez, Carlos-D. Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. In *Speech Communication*, volume 88, pages 83–95. Elsevier BV.
- E. Bartoll and A. Martínez Tejerina. 2010. The positioning of subtitles for the deaf and hard of hearing. *Listening to Subtitles. Subtitles for the Deaf and Hard of Hearing*, pages 69–86.
- Eduard Bartoll. 2004. Parameters for the classification of subtitles. *Topics in Audiovisual Translation*, 9:53–60.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit³: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.
- Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June.
- Henrik Gottlieb. 2004. Language-political implications of subtitling. *Topics in Audiovisual Translation*, 9:83–100.
- Alina Karakanta, Matteo Negri, and Marco Turchi. 2019. Are Subtitling Corpora really Subtitle-like? In *Sixth Italian Conference on Computational Linguistics, CLiC-It*.
- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online, July. Association for Computational Linguistics.
- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020b. Must-cinema: a speech-to-subtitles corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 13-15.
- Taku Kudo and John Richardson. 2018. Sentence-Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*.
- Danni Liu, Jan Niehues, and Gerasimos Spanakis. 2020. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256, Online, July. Association for Computational Linguistics.
- Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy, August. Association for Computational Linguistics.
- Mathias Müller and Martin Volk. 2013. Statistical machine translation of subtitles: From opensubtitles to ted. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, pages 132–138, Berlin, Heidelberg. Springer Berlin Heidelberg.

⁴<https://ict.fbk.eu/units-hlt-mt-e2eslt/>

- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Elisa Perego. 2008. Subtitles and line-breaks: Towards improved readability. *Between Text and Image: Updating research in screen translation*, 78(1):211–223.
- Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.
- Hye-Jeong Song, Hong-Ki Kim, Jong-Dae Kim, Chan-Young Park, and Yu-Seop Kim. 2019. Inter-sentence segmentation of YouTube subtitles using long-short term memory (LSTM). 9:1504.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

How granularity of orthography-phonology mappings affect reading development: Evidence from a computational model of English word reading and spelling

Alfred Lim and Beth A. O'Brien

Nanyang Technological University, Singapore

alfred.lim@nie.edu.sg

beth.obrien@nie.edu.sg

Luca Onnis

University of Genoa, Italy

luca.onnis@unige.it

Abstract

It is widely held that children implicitly learn the structure of their writing system through statistical learning of spelling-to-sound mappings. Yet an unresolved question is how to sequence reading experience so that children can 'pick up' the structure optimally. We tackle this question here using a computational model of encoding and decoding. The order of presentation of words was manipulated so that they exhibited two distinct progressions of granularity of spelling-to-sound mappings. We found that under a training regime that introduced written words progressively from small-to-large granularity, the network exhibited an early advantage in reading acquisition as compared to a regime introducing written words from large-to-small granularity. Our results thus provide support for the grain size theory (Ziegler and Goswami, 2005) and demonstrate that the order of learning can influence learning trajectories of literacy skills.

1 Introduction

Reading science provides evidence of the developmental path to acquiring reading for alphabetic languages (Ehri, 2005; Rayner et al., 2001). From parsing the speech stream into words in infancy (Christiansen et al., 2006; Saffran et al., 1997), to familiarizing with print in the preschool years (Thompson, 2009) — these activities lead to the accrual of key knowledge for learning to read. Knowledge about the language's phonotactic and

graphotactic properties and symbolic representations with abstract letter units is necessary for the forthcoming insight that print represents spoken language (the alphabetic principle). Subsequent to this insight, children are ready to take on the process of learning the precise mapping of print-to-speech.

At its basis, learning to read involves learning to *decode* a script into oral language representations. The question arises as to the optimal input for learning this orthography-to-phonology mapping in an alphabetic system, especially for languages that have deep orthographies, such as English. Shallow orthographies (e.g., Finnish, Spanish) have a more precise match between letters and sounds; whereas deep orthographies match phonemes to graphemes (one or more letters) in an inconsistent way — with multiple spellings per phoneme, or multiple pronunciations per grapheme — and, thus, have a greater number of GPCs (grapheme-phoneme correspondences). Therefore, reading acquisition is found to occur at a comparatively slower rate for readers in deep as compared to shallow orthographies (Ellis et al., 2004; Georgiou et al., 2008; Florit and Cain, 2011).

The deep orthographic complexity of English also partly results from variation in the functional units of the writing system — graphemes — which may consist of a single letter (e.g., *a*), or multiple letters (e.g., *ay*, *aye*). While skilled adult readers have unitized these subword patterns (Rey et al., 2000), beginning readers need to acquire these patterns of graphemes and their mappings to phonemes. Here we consider this mapping problem along two dimensions: (1) the *granularity* of the units of analysis to be picked up at any given time; and (2) the ordering of learning such units and types.

A fruitful approach to examining the GPC learning process is through computational mod-

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

elling (Monaghan and Ellis, 2010; Perry et al., 2019; Pritchard et al., 2016). Specifically, connectionist models are sensitive to the timing and ordering of learning events, in that they learn incrementally. This feature is particularly apt for modeling reading development, as it affords simulating the incremental nature of a child learning to read new words daily, as schooling progresses. Order effects as well as frequency trajectory effects have been documented in previous connectionist models (Mermillod et al., 2012), and here we are interested in comparing learning trajectories for particular training orderings for reading development.

To this end, we present connectionist networks with small batches of words, which we test regularly for accuracy until a given criterion across the batch is achieved — in essence, an adaptive training regime. Using this approach, we can address long-standing issues in the area of reading education with a more systematic approach to understanding how print-to-speech mappings are learned (Rueckl, 2016).

Below we briefly review why print-to-speech decoding can be a hard problem, both for learners and for researchers trying to understand its mechanisms. Then, we discuss dimensions of granularity derived from the literature, and offer a first set of connectionist simulations of the order of reading acquisition of American English.

1.1 Is there an optimal reading experience?

The psycholinguistic grain size theory (Ziegler and Goswami, 2005) has generated much research on reading acquisition, including across different alphabetic languages. It espouses that granularity for oral and written language development proceed in different directions — from larger to smaller, vs. from smaller to larger units. Thus, the mismatch in unit or “grain” sizes available over development introduces a disparity in learning the mapping between orthography and phonology.

This learning challenge has led to investigations of behavioral interventions for teaching reading at either whole word or subword levels (National Reading Panel, 2001; McArthur et al., 2015), showing an advantage for subword approaches emphasizing letter, grapheme or larger (subsyllable onset-rime) units (Rayner et al., 2001; Ehri et al., 2001; Torgerson et al., 2006; Olson and Wise, 1992; Ecalle et al., 2009). At the same time, the optimal subword grain size has been debated. De-

velopmentally, Treiman et al. (2006) reported that children appear to initially attend to small units (graphemes), before gradually showing an influence of surrounding graphemes when confronted with inconsistencies in pronunciation.

Thus, in the current study we focus on single grapheme to phonemes, or single phoneme to grapheme mappings in our inquiry of granularity and learning to read. In this way, we make no assumptions about a beginning reader’s knowledge of subword units or syllable structure, instead assuming all letters are created equal (whether vowels or consonants) and that the reading system must initially acquire knowledge of print patterns for GPC on its own, through experience with the print input. **Granularity** was, therefore, operationalised for each word as the difference between the number of letters (N_{letter}) and phonemes (N_{phon} ; i.e., $N_{letter} - N_{phon}$). For example, the granularity of the word *mince* ($N_{letter} = 5$, $N_{phon} = 4$) is 1, and the granularity of *thought* ($N_{letter} = 7$, $N_{phon} = 3$) is 4. A granularity of 0, hence, indicates that the word comprises of no multi-letter grapheme (e.g., *held*, *storm*).

The aim of this study was to systematically examine granularity related to the learning of GPC and word decoding. Theoretical accounts of the best representational units for learning to read have not been explicitly tested in the modelling literature to our knowledge. This, in turn, may inform instructional practices as to the best approaches for optimizing the learning curve, and results can be interpreted in terms of optimal child developmental trajectories and reading curricula (McKeown et al., 2017).

2 Method

2.1 Model Architecture

The model had four types of layers: orthographic, phonological, hidden, and clean-up (see Figure 1). The orthographic and phonological layers were each connected to a clean-up layer that mediated connections within the respective units, creating an attractor network that settles into a stable pattern over time (Harm and Seidenberg, 1999).

The orthographic layer was composed of 260 units, corresponding to 10 positions \times 26 possible letters. Words were coded as vowel-centred, such that the fourth slot was filled with the left-most vowel of a word (e.g., *mince* \rightarrow _ _ m i n c e _ _ _).

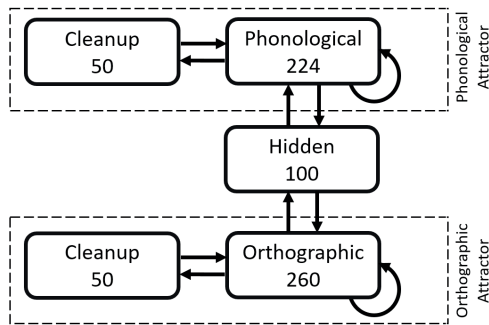


Figure 1: Architecture of the implemented reading model.

A word’s phonology was represented with nodes coding features of phonemes (8 positions \times 28 possible phonological features = 224 units). Pronunciation of each word was positioned with the vowel at the fourth slot (e.g., *mince* \rightarrow / _ _ m i n s _ _/). Each phoneme was encoded by a binary vector of 28 phonological features taken from PHOIBLE (Moran and McCloy, 2019), an online repository of cross-lingual phonological data. A list of phonemes and their respective phonological features used in the present work can be found in the Open Science Framework (OSF) repository for this project (<https://osf.io/hj96x/>).

While traditionally the problem of learning to read is conceptualized in terms of decoding unidirectionally from orthography to phonology, research suggests that children engage in spelling words simultaneously as they learn to decode. In addition, *feedback* sound-to-spelling relations are also informative in establishing mappings for reading. Thus, we implemented a new model with a bidirectional network architecture that connects orthographic-to-phonological and phonological-to-orthographic layers via the hidden units.

2.2 Training Procedure

The model was trained with a learning rate of 0.05 using a back-propagation through time (BPTT) algorithm with input integration and a time constant of 0.5 (Harm and Seidenberg, 1999; Plaut et al., 1996). Each word item was clamped and presented for six time ticks, and then in an additional six time ticks, the model was required to reproduce the target pattern of the word by the final 12th tick. The weight connections are updated based on cross-entropy error computed between the target and the actual activation of the output units.

Training proceeded in two distinct stages re-

flecting naturalistic child language development: (1) a pre-literacy training stage, in which the model was trained to learn the phonology-to-phonology mappings with an accuracy of 99%; and (2) a literacy training stage, in which the model was trained on both **decoding** (orthography-to-phonology) and **encoding** (phonology-to-orthography) tasks in a sequential manner. The pre-literacy stage of training was intended to mimic the fact that children develop oral skills through hearing and speaking long before learning to read.

Models were trained with a cumulative process of learning to encode and decode, whereby words with different granularity were introduced to the model in either an ascending or descending sequence. These two models were referred to as **small-to-large (SL)** and **large-to-small (LS)** from here onwards.

Words were first sorted with regard to their granularity, followed by a second-level sorting criterion to arrange words with the same granularity in order of decreasing frequency. The first batch of words in each training regime, therefore, comprises of high frequency words that are of either the smallest (e.g., *fix*, *lynx*) or largest (e.g., *bought*, *should*) granularity in the corpus. During training, words were sampled according to their frequency from the Word Frequency Guide (WFG) corpus (Zeno et al., 1995), and the resulting probability values were normalized over all words in the training set. Correspondingly, low frequency words had a lower probability of being presented to the model during training as compared to high frequency words [e.g., $P(yules) = 0.05$ vs. $P(of) = 0.97$].

2.2.1 Adaptive training

Teachers introduce written words progressively to their pupils, and regularly assess progress before introducing new words. Likewise, our model training introduced batches of 45 new words at a time. Importantly, a new batch of words was introduced only after model performance exceeded a criterion threshold of 70% combined accuracy for the decoding and encoding tasks on trained words — which included only words that the model had been trained on cumulatively up to the last training epoch. This tested the network success at reproducing the training set to which it had been progressively exposed, and allowed us to compare the rates of learning under different training regimes.

2.3 Testing Procedure

Two complementary tests are carried out every 100 training epochs: (1) a *total vocabulary test* which uses words from the entire corpus, regardless of whether they have been presented to the model in previous training phases; and (2) an *untrained pseudo-words test* which uses a fixed set of pronounceable and spellable monosyllabic non-words. This pseudo-word set is derived from previous empirical studies on developmental reading skills (Torgesen et al., 1999). Thus, with these tests we assess the network's (1) transfer and (2) decoding abilities.

Because no learning occurs during testing, the same set of test words and non-words can be used routinely as novel testing items after 100 training epochs. This represents a considerable advantage with respect to behavioural longitudinal experiments, where successive test sessions can suffer from previous exposure effects.

Each test was administered twice, once in a decoding task and again in an encoding task. The **decoding task** activated the orthographic pattern for a given test word on the orthographic layer, say, *eye*, and measured the accuracy of the network to reproduced the corresponding target phonological word (/aɪ/) on the phonological layer. Conversely, the **encoding task** activated the phonological pattern for a given word on the phonology layer, say, /aɪ/, and measured the accuracy of the network to reproduced the corresponding target orthographic word (*eye*) on the orthographic layer.

Similar to the training procedure, each test word item was clamped and presented for six time samples, and then in an additional six time samples, the model was required to produce the target phonological/orthographic pattern of the word. An output was scored as correct when the target nodes were active with a value ≥ 0.75 , and concurrently the other nodes were inactive (≤ 0.25). Intermediate values were considered incorrect.

2.4 Corpus

All stimuli were monosyllabic American English words. The CHILDES database (MacWhinney, 2000), WFG corpus (Zeno et al., 1995), and the Phonemic Decoding Efficiency sub-test of the TOWRE (Torgesen et al., 1999) were used for pre-literacy training ($N = 5032$), literacy training ($N = 4394$), and pseudo-words testing ($N = 163$), respectively. The full list of words and their respec-

tive granularity and batch number can be found on OSF (<https://osf.io/hj96x/>).

To check whether frequency covaries with grain size, and may therefore confound the order effect, we conducted Spearman's correlations across the training regime between batch number and mean log frequency per batch. This was done for each training order: **small-to-large (SL)** and **large-to-small (LS)**. Importantly, while batch number was significantly correlated with frequency for both training orders [SL: $r_s(96) = -0.43$, $p < .001$; LS: $r_s(96) = -0.30$, $p = .003$], the relation was in the same, negative direction in both cases — ensuring that frequency was not systematically tied to grain size. Rather, the result was from the second-level sorting by frequency in descending order.

To identify the possible relationship between the granularity and consistency of the mapping for the units to be learned, we calculated the decoding and encoding consistency measures to reflect how often the orthographic/phonological unit was spelled/pronounced in the same way as it was across all words (Berndt et al., 1987). The procedure required the conditional probabilities of GPCs and PGCs to be computed as they occur in the corpus [e.g., the probability of the grapheme *ew* being pronounced as /o/ is, $P(/o/|ew) = 0.057$].

We then derived a *composite consistency score* to account for the two measures (decoding and encoding), with a higher score representing higher overall bi-directional word consistency. Consistency was found to correlate negatively with granularity increases [SL: $r_s(96) = -0.69$, $p < .001$; LS: $r_s(96) = 0.71$, $p < .001$], indicating that words with smaller granularity were more consistent.

3 Results

At the time of writing, each model had been trained on 67 out of 98 batches of words (or 3015 out of 4394 unique words). While incomplete, our preliminary observations suggest a clear difference in rate of learning across the two training regimes.

Results are summarised in Figure 2, and show that under a training regime that introduces written words in batches progressively from small-to-large granularity, the network exhibited an early advantage in reading acquisition as compared to a regime introducing written words from large-to-

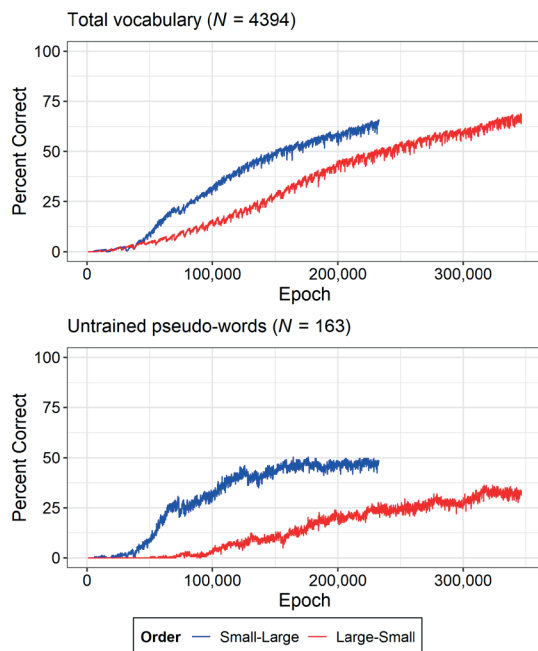


Figure 2: Models’ accuracy in the encoding task when tested against the full vocabulary and a set of pseudo-words. Test results compare models trained on two ordering regimes based on granularity of the orthography-phonology mappings.

small granularity.

The two types of repeated tests served to evaluate the accuracy of phonological output for: (a) total vocabulary (including trained and untrained words) and (b) pseudo-words. Both tests measured the ability of the networks to generalize to unseen but orthographically legal strings (see Figure 2). Specifically, the SL and LS models took 232800 and 346400 epochs, respectively, to reach the criterion threshold of 70% accuracy for all 67 batches of words that were introduced cumulatively over time. Apart from reaching the criterion threshold earlier, the SL model also performed better than the LS model in pseudo-words test (47.85% vs. 33.13%) at the end of preliminary training.

4 Discussion

As the process of learning to read requires picking up and internalizing representational units of print associated with sound, the ordering of training input to the reading system becomes paramount. How best to order input and maximize learning efficiency has been debated in the literacy education field. This study capitalizes on a computational modelling approach to this issue, using a highly

controlled context without the ethical concerns of human learning studies. Directly contrasting the effects of two literacy training regimes differing in granularity order, the simulation results support better learning with smaller, less complex orthographic units, as predicted from corpus-based research (Vousden, 2008). At training stages comprising of 3015 words, we found that the model initially trained with words of smaller granularity performed and generalized to pseudo-words better than the model trained with larger granularity. The LS model did require significantly more training epochs to reach the same performance as the SL model.

Essentially, when children learn to read, they must navigate the structure of their language and its writing system. Granularity and consistency are important aspects of this structure, and both impact reading performance. Adult readers are slower to identify letters within a multi-letter grapheme (Smith and Monaghan, 2011; Rey et al., 2000), suggesting that graphemes are functional reading units. Furthermore, Rastle and Coltheart (1998) found that naming latencies were slower for pseudo-words with, as compared to without, multi-letter graphemes. Adult word naming and lexical decision are also faster for consistent words (Andrews, 1982; Jared, 1997; Jared, 2002), and consistent words are more accurately read and spelled by children (Alegria and Mousty, 1996; L  t   et al., 2008; Weekes et al., 2006).

Granularity and consistency have been regarded to be associated (Treiman et al., 1995), and our corpus analysis revealed this as well — monosyllabic English words of smaller granularity tend to be more consistent than words with larger granularity. This relationship indicates that granularity and consistency may not be entirely disentangled, at least for English. With this in mind, the SL model was first exposed to words of smaller granularity that were also more consistent in their GPC and PGC (phoneme-grapheme correspondence) mappings. Thus consistency and granularity may be two sides of the same coin, and when manipulated they could lead to faster or slower rates of convergence. Importantly, the current model included bidirectional links between orthographic and phonological units, simulating the real-world scenario that children acquire decoding and encoding skills simultaneously.

These findings have implications for educa-

tional planning for early literacy. In particular, our pilot simulation provides preliminary evidence on the potential utility of manipulating the order of training in terms of word granularity to unveil facilitative effects on literacy acquisition. Reading instruction can consider the early acquisition of words with smaller granularity, or more consistency. However, we note that the present findings are based on the analysis of monosyllabic words only and should not be generalized to multisyllabic words directly. Future work can consider using models that are capable of reading multisyllabic words (Perry et al., 2010), or explore the link between granularity and consistency across languages that are either less or more orthographically transparent.

Acknowledgments

Support comes from the Education Research Funding Programme of the National Institute of Education (NIE), Nanyang Technological University, Singapore, grant #OER0417OBA.

References

- Jesus Alegria and Philippe Mousty. 1996. The development of spelling procedures in french-speaking, normal and reading-disabled children: Effects of frequency and lexicality. *Journal of experimental child psychology*, 63(2):312–338.
- Sally Andrews. 1982. Phonological recoding: Is the regularity effect consistent? *Memory & Cognition*, 10(6):565–575.
- Rita S. Berndt, James A. Reggia, and Charlotte C. Mitchum. 1987. Empirically derived probabilities for grapheme-to-phoneme correspondences in english. *Behavior Research Methods, Instruments, & Computers*, 19(1):1–9.
- Morten H. Christiansen, Stephen A. Hockema, and Luca Onnis. 2006. Using phoneme distributions to discover words and lexical categories in unsegmented speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- Jean Ecalte, Annie Magnan, and Caroline Calmus. 2009. Lasting effects on literacy skills with a computer-assisted learning using syllabic units in low-progress readers. *Computers & Education*, 52(3):554–561.
- Linnea C. Ehri, Simone R. Nunes, Steven A. Stahl, and Dale M. Willows. 2001. Systematic phonics instruction helps students learn to read: Evidence from the national reading panel’s meta-analysis. *Review of Educational Research*, 71(3):393–447.
- Linnea C. Ehri. 2005. Learning to read words: Theory, findings, and issues. *Scientific Studies of reading*, 9(2):167–188.
- Nick C. Ellis, Miwa Natsume, Katerina Stavropoulou, Lorenc Hoxhallari, Victor H. P. Van Daal, Nicoletta Polyzoe, Maria-Louisa Tsipa, and Michalis Petalas. 2004. The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading research quarterly*, 39(4):438–468.
- Elena Florit and Kate Cain. 2011. The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, 23(4):553–576.
- George K. Georgiou, Rauno Parrila, and Timothy C. Papadopoulos. 2008. Predictors of word decoding and reading fluency across languages varying in orthographic consistency. *Journal of Educational Psychology*, 100(3):566.
- Michael W. Harm and Mark S. Seidenberg. 1999. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, 106(3):491.
- Debra Jared. 1997. Spelling–sound consistency affects the naming of high-frequency words. *Journal of Memory and Language*, 36(4):505–529.
- Debra Jared. 2002. Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language*, 46(4):723–750.
- Bernard Lété, Ronald Peereman, and Michel Fayol. 2008. Consistency and word-frequency effects on spelling among first-to fifth-grade french children: A regression-based study. *Journal of Memory and Language*, 58(4):952–977.
- Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.
- Genevieve McArthur, Saskia Kohnen, Kristy Jones, Philippa Eve, Erin Banales, Linda Larsen, and Anne Castles. 2015. Replicability of sight word training and phonics training in poor readers: a randomised controlled trial. *PeerJ*, 3:e922.
- Margaret G. McKeown, Paul D. Deane, and Ren R. Lawless. 2017. *Vocabulary assessment to support instruction: Building rich word-learning experiences*. Guilford Publications.
- Martial Mermillod, Patrick Bonin, Alain Méot, Ludovic Ferrand, and Michel Paindavoine. 2012. Computational evidence that frequency trajectory theory does not oppose but emerges from age-of-acquisition theory. *Cognitive Science*, 36(8):1499–1531.
- Padraic Monaghan and Andrew W. Ellis. 2010. Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language*, 63(4):506–525.

- Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.
- National Reading Panel. 2001. *Report of the National Reading Panel: Teaching Children to Read (Report NIH Pub. No. 00-4769)*. National Institutes of Health.
- Richard K. Olson and Barbara W. Wise. 1992. Reading on the computer with orthographic and speech feedback. *Reading and Writing*, 4(2):107–144.
- Conrad Perry, Johannes C. Ziegler, and Marco Zorzi. 2010. Beyond single syllables: Large-scale modeling of reading aloud with the connectionist dual process (cdp++) model. *Cognitive psychology*, 61(2):106–151.
- Conrad Perry, Marco Zorzi, and Johannes C. Ziegler. 2019. Understanding dyslexia through personalized large-scale computational models. *Psychological science*, 30(3):386–395.
- David C. Plaut, James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56.
- Stephen C. Pritchard, Max Coltheart, Eva Marinus, and Anne Castles. 2016. Modelling the implicit learning of phonological decoding from training on whole-word spellings and pronunciations. *Scientific studies of reading*, 20(1):49–63.
- Kathleen Rastle and Max Coltheart. 1998. Whammy and double whammy: Length effects in nonword naming. *Psychonomic Bulletin and Review*, 5:277–282.
- Keith Rayner, Barbara R. Foorman, Charles A. Perfetti, David Pesetsky, and Mark S Seidenberg. 2001. How psychological science informs the teaching of reading. *Psychological science in the public interest*, 2(2):31–74.
- Arnaud Rey, Johannes C. Ziegler, and Arthur M. Jacobs. 2000. Graphemes are perceptual reading units. *Cognition*, 75(1):B1–B12.
- Jay G. Rueckl. 2016. Toward a theory of variation in the organization of the word reading system. *Scientific Studies of Reading*, 20(1):86–97.
- Jenny R. Saffran, Elissa L. Newport, Richard N. Aslin, Rachel A. Tunick, and Sandra Barrueco. 1997. Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, 8(2):101–105.
- Alastair C. Smith and Padraic Monaghan, 2011. *What are the functional units in reading? Evidence for statistical variation influencing word processing*, volume 20, pages 159–172. World Scientific. Proceedings of the 12th Neural Computation and Psychology Workshop.
- Ross A. Thompson. 2009. Doing what doesn't come naturally. *Zero to Three Journal*, 30(2):33–39.
- Carole Torgerson, Greg Brooks, and Jill Hall. 2006. *A systematic review of the research literature on the use of phonics in the teaching of reading and spelling*. DfES Publications Nottingham.
- Joseph K. Torgesen, Richard K. Wagner, Carol A. Rashotte, Elaine Rose, Patricia Lindamood, Tim Conway, and Cyndi Garvan. 1999. Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, 91(4):579.
- Rebecca Treiman, John Mullennix, Ranka Bijeljac-Babic, and E. Daylene Richmond-Welty. 1995. The special role of rimes in the description, use, and acquisition of english orthography. *Journal of Experimental Psychology: General*, 124(2):107.
- Rebecca Treiman, Brett Kessler, Jason D. Zevin, Suzanne Bick, and Melissa Davis. 2006. Influence of consonantal context on the reading of vowels: Evidence from children. *Journal of Experimental Child Psychology*, 93(1):1–24.
- Janet I. Vouden. 2008. Units of english spelling-to-sound mapping: a rational approach to reading instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(2):247–272.
- Brendan S. Weekes, Anne E. Castles, and Robert A. Davies. 2006. Effects of consistency and age of acquisition on reading and spelling among developing readers. *Reading and Writing*, 19(2):133–169.
- Susan Zeno, Stephen H. Ivens, Robert T. Millard, and Raj Duvvuri. 1995. *The educator's word frequency guide*. Touchstone Applied Science Associates.
- Johannes C. Ziegler and Usha Goswami. 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, 131(1):3.

Simple Data Augmentation for Multilingual NLU in Task Oriented Dialogue Systems

Samuel Louvan
University of Trento
Fondazione Bruno Kessler
slouvan@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

Data augmentation has shown potential in alleviating data scarcity for Natural Language Understanding (e.g. slot filling and intent classification) in task-oriented dialogue systems. As prior work has been mostly experimented on English datasets, we focus on five different languages, and consider a setting where limited data are available. We investigate the effectiveness of non-gradient based augmentation methods, involving simple text span substitutions and syntactic manipulations. Our experiments show that (i) augmentation is effective in all cases, particularly for slot filling; and (ii) it is beneficial for a joint intent-slot model based on multilingual BERT, both for limited data settings and when full training data is used.

1 Introduction

Natural Language Understanding (NLU) in task-oriented dialogue systems is responsible for parsing user utterances to extract the intent of the user and the arguments of the intent (i.e. *slots*) into a semantic representation, typically a semantic frame (Tur and De Mori, 2011). For example, the utterance “*Play Jeff Pilson on Youtube*” has the intent PLAYMUSIC and “*Youtube*” as value for the slot SERVICE. As more skills are added to the dialogue system, the NLU model frequently needs to be updated to scale to new domains and languages, a situation which typically becomes problematic when labeled data are limited (*data scarcity*).

One way to combat data scarcity is through data augmentation (DA) techniques performing *label preserving* operations to produce auxiliary training data. Recently, DA has shown potential in tasks such as machine translation (Fadaee et al., 2017), constituency and dependency parsing

(Şahin and Steedman, 2018; Vania et al., 2019), and text classification (Wei and Zou, 2019; Kumar et al., 2020). As for slot filling (SF) and intent classification (IC), a number of DA methods have been proposed to generate synthetic utterances using sequence to sequence models (Hou et al., 2018; Zhao et al., 2019), Conditional Variational Auto Encoder (Yoo et al., 2019), or pre-trained NLG models (Peng et al., 2020). To date, most of the DA methods are evaluated on English and it is not clear whether the same finding apply to other languages.

In this paper, we study the effectiveness of DA on several non-English datasets for NLU in task-oriented dialogue systems. We experiment with existing lightweight, non-gradient based, DA methods from Louvan and Magnini (2020) that produces varying slot values through substitution and sentence structure manipulation by leveraging syntactic information from a dependency parser. We evaluate the DA methods on NLU datasets from five languages: Italian, Hindi, Turkish, Spanish, and Thai. The contributions of our paper are as follows:

1. We assess the applicability of DA methods for NLU in task-oriented dialogue systems in five languages.
2. We demonstrate that simple DA can improve performance on all languages despite different characteristic of the languages.
3. We show that a large pre-trained multilingual BERT (M-BERT) (Devlin et al., 2019) can still benefit from DA, in particular for slot filling.

2 Slot Filling and Intent Classification

The NLU component of a task-oriented dialogue system is responsible in a parsing user utterance into a semantic representation, such as semantic

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

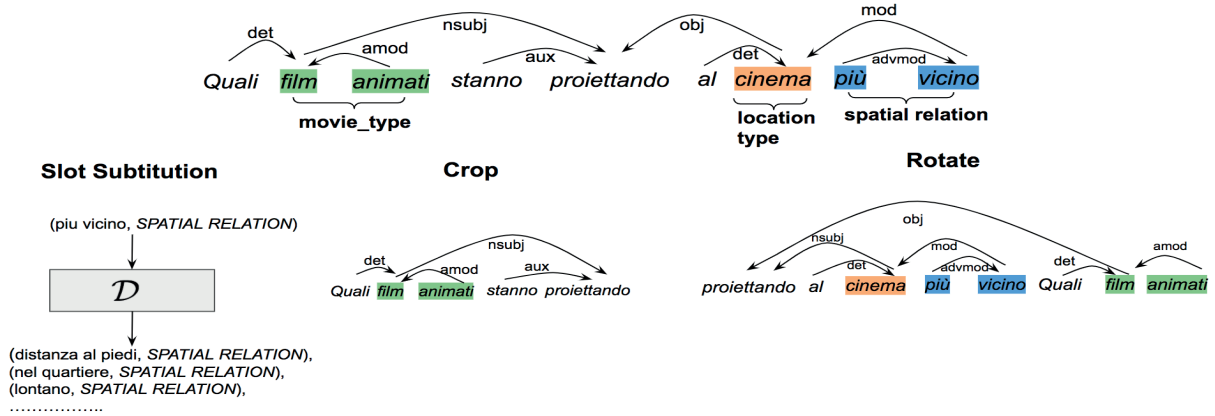


Figure 1: Augmentation operations performed on an utterance, “Quali film animati stanno proiettando al cinema più vicino” (“Which animated films are showing at the nearest cinema”). The utterance is taken from the Italian SNIPS dataset.

frame. The semantic frame conveys information, namely the user intent and the corresponding arguments of the intent. Extracting such information involves slot filling (SF) and intent classification (IC) tasks.

Given an input utterance of n tokens, $x = (x_1, x_2, \dots, x_n)$, the system needs to assign a particular intent y^{intent} for the whole utterance x and the corresponding slots that are mentioned in the utterance $y^{slot} = (y_1^{slot}, y_2^{slot}, \dots, y_n^{slot})$. In practice, IC is typically modeled as text classification and SF as a sequence tagging problem. As an example, for the utterance “Play Jeff Pilson on Youtube”, y^{intent} is PLAYMUSIC, as the intent of the user is to ask the system to play a song from a musician and $y^{slot} = (O, B-ARTIST, I-ARTIST, O, B-SERVICE)$ in which the artist is “Jeff Pilson” and the service is “Youtube”. Slot labels are in BIO format: B indicates the start of a slot span, I the inside of a span while O denotes that the word does not belong to any slot. Recent approaches for SF and IC are based on neural network methods that models SF and IC jointly (Goo et al., 2018; Chen et al., 2019) by sharing model parameter among both tasks.

3 Data Augmentation (DA) Methods

DA aims to perform *semantically preserving* transformations on the training data \mathcal{D} to produce auxiliary data \mathcal{D}' . The union of \mathcal{D} and \mathcal{D}' is then used to train a particular NLU model. For each utterance in \mathcal{D} , we produce N augmented utterances by applying a specific augmentation operation. We adopt a subset of existing augmentation

methods from Louvan and Magnini (2020), that has shown promising results on English datasets. We describe the augmentation operations in the following sections.

3.1 Slot Substitution (SLOT-SUB)

SLOT-SUB (Figure 1 left) performs augmentation by substituting a particular text span (*slot-value pair*) in an utterance with a different text span that is semantically consistent i.e., the slot label is the same. For example, in the utterance “Quali film animati stanno proiettando al cinema più vicino”, one of the spans that can be substituted is the slot value pair (*più vicino*, SPATIAL RELATION). Then, we collect other spans in \mathcal{D} in which the slot values are different, but the slot label is the same. For instance, we found the substitute candidates $SP' = \{(\text{“distanza a piedi”}, \text{SPATIAL RELATION}), (\text{“lontano”}, \text{SPATIAL RELATION}), (\text{“nel quartiere”}, \text{SPATIAL RELATION}), \dots\}$, and then we sample one span to replace the original span in the utterance.

3.2 CROP and ROTATE

In order to produce sentence variations, we apply the crop and rotate operations proposed in Şahin and Steedman (2018), which manipulate the sentence structure through its dependency parse tree. The goal of CROP (Figure 1 middle) is to simplify the sentence so that it focuses on a particular *fragment* (e.g. subject/object) by removing other fragments in the sentence. CROP uses the dependency tree to identify the fragment and then remove it and its children from the dependency tree.

Dataset	Language	#Label		#Utterances (\mathcal{D})			#Augmented Utterances (\mathcal{D}')		
		#slot	#intent	#train	#dev	#test	#SLOT-SUB	#CROP	#ROTATE
SNIPS-IT	Italian	39	7	574	700	698	5,404	1,431	1,889
ATIS-HI	Hindi	73	17	176	440	893	1,286	460	472
ATIS-TR	Turkish	70	17	99	248	715	144	161	194
FB-ES	Spanish	11	12	361	1,983	3,043	1,455	769	1,028
FB-TH	Thai	8	10	215	1,235	1,692	781	-	-

Table 1: Statistics on the datasets. #train indicates our limited training data setup (10% of full training data). \mathcal{D}' is produced by tuning the number of augmentations per utterance (N) on the dev set.

Model	DA	SNIPS-IT		ATIS-HI		ATIS-TR		FB-ES		FB-TH	
		Slot	Intent	Slot	Intent	Slot	Intent	Slot	Intent	Slot	Intent
M-BERT	None	78.25	94.99	69.57	86.57	64.36	78.98	84.13	97.68	56.06	89.80
	SLOT-SUB	81.97[†]	94.93	72.44[†]	87.29	66.60 [†]	79.85	84.27	97.72	59.68[†]	91.42[†]
	CROP	80.12 [†]	94.60	70.04	86.92	65.11	79.48	83.85	98.08 [†]	-	-
	ROTATE	79.24 [†]	95.37	70.69	87.60[†]	65.20	80.06	83.28	98.20[†]	-	-
	COMBINE	81.27 [†]	95.00	72.13 [†]	86.93	66.68[†]	81.12[†]	83.67	97.94	-	-

Table 2: Performance comparison of the baseline and augmentation methods on the test set. F1 score is used for slot filling and accuracy for intent classification. Scores are the average of 10 different runs. [†] indicates statistically significant improvement over the baseline (p -value < 0.05 according to Wilcoxon signed rank test).

The ROTATE (Figure 1 right) operation is performed by moving a particular fragment (including subject/object) around the root of the tree, typically the verb in the sentence. For each operation, all possible combinations are generated, and one of them is picked randomly as the augmented sentence. Both CROP and ROTATE rely on the universal dependency labels (Nivre et al., 2017) to identify relevant fragments, such as NSUBJ (nominal subject), DOBJ (direct object), OBJ (object), IOBJ (indirect object).

4 Experiments

Our primary goal is to verify the effectiveness of data augmentation on Italian, Hindi, Turkish, Spanish and Thai NLU datasets with limited labeled data. To this end, we compare the performance of a baseline NLU model trained on the original training data (\mathcal{D}) with a NLU model that incorporates the augmented data as additional training instances ($\mathcal{D} + \mathcal{D}'$). To simulate the limited labeled data situation we randomly sample 10% of the training data for each dataset.

Baseline and Data Augmentation (DA) Methods. We use the state of the art BERT-based joint intent slot filling model (Chen et al., 2019) as the baseline model. We leverage the pre-trained

multilingual BERT (M-BERT), which is trained on 104 languages. During training, M-BERT is fine tuned on the slot filling and intent classification tasks. Given a sentence representation $x = ([CLS] t_1 t_2 \dots t_L)$, we use the hidden state $h_{[CLS]}$ to predict the intent, and h_{t_i} to predict the slot label. As for DA methods, in addition to the methods described in Section 2, we add one configuration COMBINE, which combines the result of SLOT-SUB and ROTATE, as ROTATE obtains better results than CROP on the development set.

Settings. The model is trained with the BertAdam optimizer for 30 epochs with early stopping. The learning rate is set to 10^{-5} and batch size is 16. All the hyperparameters are listed in Appendix A. For SLOT-SUB the number of augmentation per sentence N is tuned on the development set. To produce the dependency tree, we parse the sentence using Stanza (Qi et al., 2020). For both CROP and ROTATE we follow the default hyperparameters from Şahin and Steedman (2018). We did not experiment with Thai for CROP and ROTATE as Thai is not supported by Stanza. The number of augmented sentences (\mathcal{D}') for each method is listed in Table 1. For evaluation metric, we use the standard CoNLL script to compute F1 score for slot filling and accuracy for intent classification.

Datasets. For the Italian language, we use the data from Bellomaria et al. (2019), translated from the English SNIPS dataset (Coucke et al., 2018). SNIPS has been widely used for evaluating NLU models and consists of utterances in multiple domains. As for Hindi and Turkish, we use the ATIS dataset from Upadhyay et al. (2018), derived from Hemphill et al. (1990). ATIS is a well known NLU dataset on flight domain. As for Spanish and Thai we use the FB dataset from Schuster et al. (2019) that contains utterances in alarm, weather, and reminder domains. The overall statistics of the datasets are shown in Table 1.

5 Results

The overall results reported in Table 2 show that applying DA improves performance on slot filling and intent classification across all languages. In particular, for SF, the SLOT-SUB method yields the best result, while for IC, ROTATE obtains better performance compared to CROP in most cases. These results are consistent with the finding from Louvan and Magnini (2020) on the English dataset, where SLOT-SUB improves SF and CROP or ROTATE improve IC. In general, ROTATE is better than CROP for most cases on IC, and we think this is because CROP may change the intent of the original sentence. Intents typically depend on the occurrence of specific slots, so when the cropped part is a slot-value, it may change the sentence’s overall semantics.

We can see that languages with different typological features (e.g. subject/verb/object ordering)¹ benefit from ROTATE operation for IC. This result suggests that augmentation can produce useful noise (regularization) for the model to alleviate overfitting when labeled data is limited. When we use COMBINE, it still helps the performance of both SF and IC, although the improvements are not as high as when only one of the augmentation method is applied. The only language that gets the benefits the most from COMBINE is Turkish. We hypothesize that as Turkish has a more flexible word order than the other languages it benefits the most when ROTATE is performed.

Performance on varying data size. To better understand the effectiveness of SLOT-SUB, we perform further analysis on different training data size (see Figure 2). Overall, we observe that as we

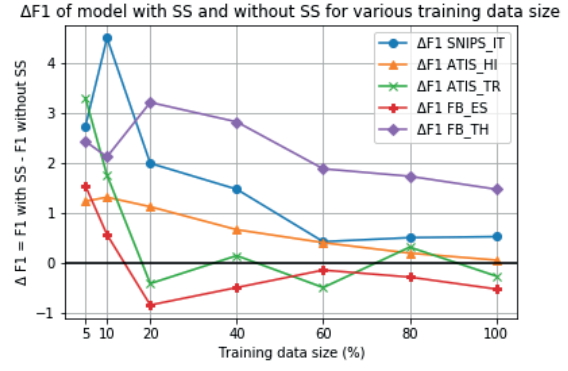


Figure 2: Improvement ($\Delta F1$) obtained by SLOT-SUB (SS) on different training data size. Positive numbers mean that the model with SS yields gain.

increase the training size, the benefit of SLOT-SUB is decreasing for all datasets. For some datasets, namely ATIS-HI and FB-ES, SLOT-SUB can cause performance drop for larger data size, although it is reasonably small (less than 1 F1 point). FB-TH consistently benefits from SLOT-SUB even when full training data is used. Until which training data size the improvement is significant vary across datasets². For SNIPS-IT, improvement is clear for all training data size and they are statistically significant up until the training data size is 80%. For ATIS-HI improvements are significant until data size of 40%. As for FB datasets, improvements are significant only until the training data size is 10%. Overall, we can see that SLOT-SUB is effective for cases where data is scarce (5%, 10%), while it is still relatively robust for larger data size on all datasets.

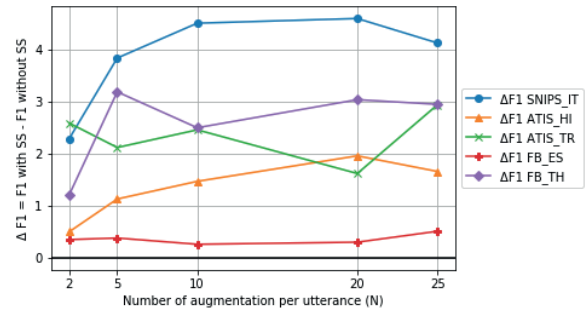


Figure 3: Gain ($\Delta F1$) obtained by SLOT-SUB (SS) on various number of augmented sentence (N). Positive numbers mean that the model with SS yields gain.

¹Italian, Spanish, and Thai are SVO languages while Hindi and Turkish are SOV languages.

²For more details of the p-value of the statistical tests please refer to Appendix B

Performance on different numbers of augmentation per utterance (N). We examine the effect of a larger number of augmentations per utterance (N) to the model performance, specifically for SF (see Figure 3). For FB-ES, similarly to the results in Table 2, increasing N does not affect the performance. For the other datasets, increasing N brings performance improvement. For ATIS-HI, SNIPS-IT, and FB-TH the trend is that, as we increase N , performance goes up and plateau. For ATIS-TR, changing N does not really affect the gain of the performance as the performance trend is quite steady across number of augmentations. For most combinations of N in each dataset (except FB-ES), the difference between the performance of model that using SLOT-SUB and the model that does not use SLOT-SUB is significant³.

6 Related Work

Data augmentation methods that has been proposed in NLP aims to automatically produce additional training data through different kinds of methods ranging from simple word substitution (Wei and Zou, 2019) to more complex methods that aims to produce semantically preserving sentence generation (Hou et al., 2018; Gao et al., 2020). In the context of slot filling and intent classification, recent augmentation methods typically apply deep learning models to produce augmented utterances.

Hou et al. (2018) proposes a two-stages methods to produce the delexicalized utterances generation and slot values realization. Their method is based on a sequence to sequence based model (Sutskever et al., 2014) to produce a paraphrase of an utterance with its slot values placeholder (delexicalized) for a given intent. For the slot values lexicalization, they use the slot values in the training data that occur in similar contexts. Zhao et al. (2019) trains a sequence to sequence model with training instances that consist of a pair of atomic templates of dialogue acts and its sentence realization. Yoo et al. (2019) proposes a solution by extending Variational Auto Encoder (VAE) (Kingma and Welling, 2014) into a Conditional VAE (CVAE) to generate synthetic utterances. The CVAE controls the utterance generation by conditioning on the intent and slot labels

³For more details of the p-value of the statistical tests please refer to Appendix B

during model training. Recent work from Peng et al. (2020) make use of Transformer (Vaswani et al., 2017) based pre-trained NLG namely GPT-2 (Radford et al., 2019), and fine-tune it to slot filling dataset to produce synthetic utterances. We consider these deep learning based approaches as *heavyweight* as they often require several stages in the augmentation process namely generating augmentation candidates, ranking and filtering the candidates before producing the final augmented data. Consequently, the computation time of these approaches is generally more expensive as separate training is required to train the augmentation and joint SF-IC models. Recent work from Louvan and Magnini (2020) apply a set of *lightweight* methods in which most of the augmentation methods do not require model training. The augmentation methods focus on varying the slot values through substitution mechanisms and varying sentence structure through dependency tree manipulation. While the methods are relatively simple it obtains competitive results with deep learning based approaches on the standard English slot filling benchmark datasets namely ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018), and FB (Schuster et al., 2019) datasets.

Existing methods mostly evaluate their approaches on English datasets, and little work has been done on other languages. Our work focuses on investigating the effect of data augmentation on five non-English languages. We apply a subset of *lightweight* augmentation methods from Louvan and Magnini (2020) that do not require separate model training to produce augmentation data.

7 Conclusion

We evaluate the effectiveness of data augmentation for slot filling and intent classification tasks in five typologically diverse languages. Our results show that by applying simple augmentation, namely slot values substitutions and dependency tree manipulations, we can obtain substantial improvement in most cases when only small amount of training data is available. We also show that a large pre-trained multilingual BERT benefits from data augmentation.

Acknowledgments

We thank Valentina Bellomaria for providing the Italian SNIPS dataset. We thank Clara Vania for the feedback on the early draft of the paper.

References

- Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almwavslu: A new dataset for SLU in italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573. Association for Computational Linguistics.
- Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 639–649. Association for Computational Linguistics.
- Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, USA, June 24-27, 1990*. Morgan Kaufmann.
- Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.
- Samuel Louvan and Bernardo Magnini. 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. *arXiv preprint https://arxiv.org/abs/2009.03695*. PACLIC 2020 - The 34th Pacific Asia Conference on Language, Information and Computation.
- Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1.
- Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models. *CoRR*, abs/2004.13952.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

- (*Long and Short Papers*), pages 3795–3805, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.
- Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.
- Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1105–1116. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.
- Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7402–7409. AAAI Press.
- Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3635–3641. Association for Computational Linguistics.

Appendix A. Hyperparameters

Hyperparameter	Value
Learning rate	10^{-5}
Dropout	0.1
Mini-batch size	16
Optimizer	BertAdam
Number of epoch	30
Early stopping	10
N	Tuned on $\{2, 5, 10\}$
Max rotation	3
Max crop	3

Table 3: List of hyperparameters used for the BERT model and data augmentation methods

Appendix B. Statistical Significance

Dataset	Nb Aug	p-value
ATIS-TR	2	0.005062032126
	5	0.01251531869
	10	0.006910429808
	20	0.5001842571
	25	0.07961580146
ATIS-HI	2	0.1097446387
	5	0.005062032126
	10	0.005062032126
	20	0.04311444678
	25	0.04311444678
SNIPS-IT	2	0.005062032126
	5	0.005062032126
	10	0.005062032126
	20	0.04311444678
	25	0.04311444678
FB-ES	2	0.0663160313
	5	0.02831405495
	10	0.09260069782
	20	0.3452310718
	25	0.07961580146
FB-TH	2	0.03665792867
	5	0.005062032126
	10	0.005062032126
	20	0.04311444678
	25	0.04311444678

Table 5: The p-values of statistical tests on the experiments on Figure 3

Dataset	Training Size (%)	p-value
ATIS-HI	5	0.04311444678
	10	0.005062032126
	20	0.04311444678
	40	0.04311444678
	80	0.1380107376
	100	0.2733216783
ATIS-TR	5	0.224915884
	10	0.005062032126
	20	0.7150006547
	40	0.1797124949
	80	0.1797124949
	100	0.1797124949
SNIPS-IT	5	0.04311444678
	10	0.005062032126
	20	0.04311444678
	40	0.04311444678
	80	0.04311444678
	100	0.04311444678
FB-ES	5	0.04311444678
	10	0.02831405495
	20	0.1797124949
	40	0.1755543028
	80	0.1380107376
	100	0.1797124949
FB-TH	5	0.04311444678
	10	0.005062032126
	20	0.1797124949
	40	0.1797124949
	80	0.1797124949
	100	0.10880943

Table 4: The p-values of statistical tests on the experiments on Figure 2.

The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Begoña Altuna
Fondazione Bruno Kessler
University of the Basque Country
altuna@fbk.eu

Alberto Lavelli
Fondazione Bruno Kessler
lavelli@fbk.eu

Manuela Speranza
Fondazione Bruno Kessler
manspera@fbk.eu

Roberto Zanolli
Fondazione Bruno Kessler
zanoli@fbk.eu

Abstract

English. We present the European Clinical Case Corpus (E3C) project, aimed at collecting and annotating a large corpus of clinical cases in five European languages (Italian, English, French, Spanish, and Basque). Project results include: (i) a freely available collection of multilingual clinical cases; and (ii) a two-level annotation scheme based on temporal relations (derived from THYME), whose purpose is to allow the construction of clinical timelines, and taxonomy relations based on medical taxonomies, to be used for semantic reasoning over clinical cases.

1 Introduction

Identifying clinically relevant events and anchoring them to a chronology is very important in clinical information processing, as the ability to access an ordered sequence of events can help to understand the evolution of clinical conditions in patients. However, although interest in information extraction from clinical narratives has increased in recent decades, attention has been focused on clinical entity extraction and classification (Schulz et al., 2020; Grabar et al., 2019; Dreisbach et al., 2019; Luo et al., 2017) rather than on temporal information. If temporal information is extracted from clinical free text, it can be added to structured data collections, e.g. MIMIC III (Johnson et al., 2016), to train clinical prediction systems. Despite some effort on the organization of clinical narratives processing challenges, e.g. CLEF eHealth (Kelly et al., 2019), few shared training and test data sets have been created, and thus developing tools for this task is still difficult.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In fact, the amount of freely available annotated corpora for any of the clinical information extraction tasks has not grown at the same rate as interest in the field, mainly due to patient privacy and data protection issues. In addition, most datasets consist of English texts, which makes research focus on that language.

In an attempt to overcome these problems, we present the European Clinical Case Corpus (E3C)¹, a project aimed at offering a freely available multilingual corpus of semantically annotated clinical narratives. The project will build a 5-language (Italian, English, French, Spanish, and Basque) clinical narrative corpus to allow for the linguistic analysis, benchmarking, and training of information extraction systems. We build upon available resources and collect new data when necessary, with the goal to harmonize current annotations, introduce new annotation layers, and provide baselines for information extraction tasks.

We foresee two types of annotations: (i) temporal information and factuality: events (including attributes expressing factuality-related information), time expressions, and temporal relations according to the THYME standard; and (ii) clinical entities: pathologies, symptoms, procedures, body parts, etc., according to standard clinical taxonomies (e.g. SNOMED-CT² (Donnelly, 2006) and ICD-10³ (WHO, 2015)).

The E3C corpus is organized into three layers, with different purposes:

Layer 1: about 25K tokens per language of clinical narratives with full manual or manually

¹E3C is a one-year pilot project, started in July 2020. The E3C website is available at <https://e3c.fbk.eu>. The project is funded by the European Language Grid (ELG), an initiative aimed at developing a cloud platform that provides access to Language Technologies (i.e. running tools and services, data sets and resources) for all European languages.

²<http://www.snomed.org/>

³<https://icd.who.int/browse10/2019/en>

checked annotation of clinical entities, temporal information and factuality, for benchmarking and linguistic analysis.

Layer 2: 50-100K tokens per language of clinical narratives with automatic annotation of clinical entities and manual check of a small sample (about 10%) of this annotation.

Layer 3: about 1M tokens per language of non-annotated medical documents (not necessarily clinical narratives) to be exploited by semi-supervised approaches.

In this paper we present our data collection effort, focused on clinical cases (Section 3), and we describe our annotation scheme (Section 4).

2 Clinical Cases

A clinical case is a statement of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. We focus on clinical cases because they are often de-identified, overcoming privacy issues, and are rich in clinical entities as well as temporal information, which is almost absent in other clinical documents (e.g., radiological reports).

A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk. Continuous superficial ulcers and spontaneous bleeding were observed under colonoscopy. Subsequent gastroscopy revealed mucosa with diffuse edema, ulcers, errhysis, and granular and friable changes in the stomach and duodenal bulb, which were similar to the appearance of the rectum. After ruling out other possibilities according to a series of examinations, a diagnosis of GDUC was considered. The patient hesitated about intravenous corticosteroids, so he received a standardized treatment with pentasa of 3.2 g/d. After 0.5 mo of treatment, the patient's symptoms achieved complete remission. Follow-up endoscopy and imaging findings showed no evidence of recurrence for 26 mo.

Here we present a sample case extracted from our collection. It is about a patient presenting gastric symptoms (mucopurulent bloody stool and epigastric persistent colic pain), who is finally diagnosed with gastroduodenitis associated with ul-

cerative colitis (GDUC). To reach the diagnosis, two consecutive medical tests (colonoscopy and gastroscopy) were performed. Treatment (treatment with pentasa of 3.2 g/d), outcome (complete remission) and follow-up (no evidence of recurrence) are also present in the text. Symptoms, tests, observations, treatments and diseases are relevant events for the history of a patient, and it is relevant to place them in chronological order, so as to understand the evolution of the health situation of the patient. For example, we know that the symptoms started 2 weeks prior to the hospital visit, that the colonoscopy was performed before the gastroscopy, that the treatment lasted for half a month and that the patient had no recurrence in the following 26 months.

Since precision in symptom description and diagnosis is utterly important in the clinical field, the clinical findings, body structures, medicines, etc., have to be uniquely identified. This can be done through international coding standards, which allow to assign a unique code to every clinically relevant element in the text.

3 Data Collection

When building the E3C corpus, a big concern has been ensuring its reusability and shareability, which forced us to use anonymised and freely redistributable clinical cases. We deal with three types of clinical narratives: discharge summaries, clinical cases published in journals, and clinical cases from medical training resources. The clinical cases in the E3C corpus contain narratives such as the excerpt presented here.

2020-09-01. The patient enters the ER due to abdominal pains. He reports chest pain 5 days ago.

The state of the data collection efforts for the five languages addressed by the project vary depending on their online presence and the number of publications available. For Spanish, a large dataset of clinical narratives and other clinical text collections already exist; for English and French, a significant amount of published material is publicly available. Corpus collection for Italian and Basque, on the other hand, has been more demanding, as we have had to manually extract clinical cases from a number of different sources.

This is shown by the data in Table 1, where we report statistics about the clinical cases col-

Language	Clinical cases	Tokens	Tok./doc
Italian	1,323	73K	55.1
English	9,533	928K	97.2
French	1615	548K	339.1
Spanish	1,400	531K	379.27
Basque	122	26K	214.2

Table 1: Statistics on the clinical cases collected for each language.

lected so far for each language⁴. The collection of clinical cases has been completed for all languages with respect to Layer 1 and for most languages with respect to Layer 2. Layer 3 of English, French and Spanish is also totally or partially filled with clinical cases.

Italian. The clinical cases come from two main sources, either cases described in public examinations (*test di abilitazione* and *test di specializzazione*) (1276 cases, 56,496 tokens) or clinical cases presented in clinical journals distributed under CC licenses (47 cases, 16,412 tokens). Apart from the clinical cases, we have also collected 8,087 patient information leaflets for medicines (13M tokens).

English. The dataset for English consists of 63,515 abstracts extracted from PubMed with the ‘clinical case’ query (9.7M tokens). From those, we identified automatically 9,533 clinical case descriptions (928,554 tokens). We first downloaded all abstracts through the PubMed API and then selected only those coming from CC-licensed journals, in order to ensure their redistribution.

French. We used the same strategy to build the French corpus. We downloaded the abstracts from PubMed and selected those from CC-licensed journals. In total, we obtained almost 12,000 abstracts (around 1.5M tokens) out of which we have automatically recognised 199 clinical case descriptions (21,485 tokens). In addition, we have also automatically extracted 1416 clinical cases (547,644 tokens) from CC-BY licensed medical journals. Apart from those, we have also collected circa 8,000 patient information leaflets for medicines (13M tokens).

Spanish. The SPACCC corpus (Intxaurreondo et al., 2018) contains 1000 clinical cases (350,761 tokens) extracted from SciELO⁵ and distributed un-

⁴For English and French, the numbers are approximate.

⁵Scientific Electronic Library Online <http://www.scielo.org>

Language	Tokens	L1 (25K)	L2 (50K)	L3 (1M)
Italian	13.2M	100%	96%	100%
English	9.7M	100%	100%	100%
French	13.7M	100%	100%	100%
Spanish	1.1M	100%	100%	100%
Basque	74K	100%	2.27%	4.76%

Table 2: Statistics on the layer coverage for each language.

der a CC license. We have also collected an additional dataset of clinical cases extracted from SciELO (400 documents, 180,216 tokens). In addition, two datasets that contain sentences extracted from clinical cases have been added to our corpus: NUBes (518,068 tokens) and IULA+ (38,208 tokens) (Lima López et al., 2020).

Basque. The Basque dataset consists of model discharge summaries (43 documents, 14,239 tokens), clinical cases presented in teaching materials (16 cases, 3,116 tokens), journals and clinical symposia (63 cases, 8,781 tokens) and a dataset of Wikipedia articles on the biomedical domain (47,613 tokens) used in other NLP tasks⁶. Some of the clinical cases are under a CC license, while explicit authorization from the owners has been obtained for the rest.

Taking into account those numbers and the types of documents we have collected for each language, we can say that we have been able to collect enough data to complete Layer 1 in all the languages. For Layer 2, instead, we have only been able to collect enough clinical cases for English, French and Spanish. Reaching the million tokens in Layer 3 is not as complicated as it may seem, as the documents in it do not necessarily need to be clinical cases, although not as many data is available for Basque. The total amount of collected tokens and the layer coverage for each language can be seen in Table 2.

Corpus collection is in a very advanced stage, but new data will be added in the near future. The whole E3C corpus, including core metadata (i.e. language, source, date, length, etc.), will be made available.

3.1 Data Protection in the E3C Corpus

As mentioned, there are two main types of documents in the E3C corpus: clinical narratives and descriptive clinical documents. The latter and

⁶<http://www.statmt.org/wmt20/biomedical-translation-task.html>

even some of the clinical cases (the ones that describe model situations) do not contain any personal data and are out of the scope of data protection regulations. Personal data protection issues, instead, regard the reports that have been written after an actual clinical case. These often contain sensitive patient information and it is the researchers' duty to disseminate them respecting data protection rules (e.g. European Union General Data Protection Regulation) and to address other ethical issues such as achieving informed consent from the patients prior to publication.

All the clinical cases in the E3C corpus have been previously published in other sources, and furthermore, they have been published under licenses that allow redistribution. As a consequence, we consider that all data protection and ethical issues were addressed at the time of first publication and that the documents already comply with the patient data protection policies.

While preparing the E3C dataset, we have also contributed to the protection of personal data, only getting the relevant information for our corpus, responding to the principle of data minimization. For example, many clinical case reports provide illustrative images that have not been considered, as image processing is out of the scope of our project.

In addition, we have also contributed to the reduction of patient traceability, as the article publication date (or an approximate one) has been established as the day the clinical case was written.

4 Annotation Scheme

E3C annotation consists of two levels that provide complementary information. On one hand, annotation of temporal information and factuality follows a mostly language-independent annotation scheme consisting of the THYME guidelines and their extensions (described in more detail in (Speranza and Altuna, 2020)). Annotation and classification of clinical entities, on the other hand, is based on two comprehensive medical taxonomies, SNOMED-CT and ICD-10.

The THYME-driven annotation focuses mainly on clinically relevant events and on the temporal relations between them, with the end goal of coding the information needed to build complete timelines, while the taxonomy-driven annotation provides semantic information and domain-specific knowledge. Looking at the sample clinical case in Section 3, the taxonomy-driven annotation might

allow one to infer, for instance, that *abdominal pains* in the first sentence and *chest pain* in the last sentence are closely related, as they are siblings in the hierarchy (in fact, they are both children of [pain of truncal structure] in SNOMED-CT). From the THYME-driven annotation, instead, one might infer the chronological order in which the two events happened.

4.1 THYME-driven Annotation

THYME offers guidelines for the annotation of clinically relevant events, time expressions and the relations between them.

Events are all actions, states, and circumstances that are relevant to the clinical history of a patient (for example, we have pathologies and symptoms such as *pain*, but also more general events such as *enters*, *reports*, and *continue*). The annotation of events also includes a number of attributes, some of which focus on factuality-related information (the contextual modality attribute, for instance, is used to mark non-factual, either generic or hypothetical, events).

Time expressions are all references to time, such as dates (both absolute like *2020-09-01* and relative like *5 days ago*), intervals (*last three days*), etc.

THYME also provides guidelines for the annotation of relations between events and/or time expressions. By expressing precedence, overlap, containment, initiation or ending between two events and/or time expressions, **TLINKs** allow for chronologically ordering them. **ALINKs** are relations that link aspectual events, i.e. events indicating a specific phase (beginning, end, continuation, etc.) of an event, to the event itself.

To obtain annotations that will allow more descriptive timelines, we have expanded the THYME annotation scheme.

Anatomical parts are not annotated in THYME even if noun phrases whose head is a body part can be clinically very relevant (as in *He had a swollen eye*). To annotate them, we have created the new **BODY PART** tag. In addition, a new **ACTOR** tag is used to mark the actors (patients, health professionals, etc.) mentioned in the narratives. Finally, **RML** is a tag we have created to mark test results, results of laboratory analyses, formulaic measurements, and measure values (which are not marked in THYME), as we think that they offer relevant insights into the health status of a patient.

Table 3 represents the annotated version of the clinical case in Section 3. The first column contains the original text (one token per line). The second column shows the span of the THYME-driven annotated elements (specifically, examples of time expressions, actors, events, and body parts) in the IOB2 format, where B-LABEL marks the first token of an element of type LABEL, I-LABEL is used for the subsequent tokens (if any), and O is used for tokens that do not belong to an annotated element. The last two columns represent the taxonomy-driven annotation (see below).

4.2 Taxonomy-driven Annotation

Clinical coding is widely spread in clinical practice; either doctors add the codes for findings, procedures, treatments, etc. to the patients' clinical histories, or large amounts of raw clinical data are automatically coded for the development of clinical prediction systems. The coded concepts are hierarchically classified in taxonomies such as SNOMED-CT and ICD-10.

SNOMED-CT is considered to be the most comprehensive clinical healthcare taxonomy, and is available for most of the languages of the E3C project, i.e. English, French, Spanish, and Basque. There is a validated SNOMED-CT version for the first three languages, while for Basque a partial version has been used (Perez de Viñaspre and Oronoz, 2015). SNOMED-CT offers 19 main categories (and a wide set of subcategories) that range from clinical findings and body structures to social contexts. On the other hand, ICD-10 (International Classification of Diseases, 10th revision) is a classification of diagnoses and procedures. The diseases are classified in 22 categories.

Taxonomy-driven annotation consists of marking in the texts all mentions of clinical entities and mapping them to a code from both international standards.

Table 3 represents the annotated version of the clinical case in Section 3. The third and fourth columns show the span of the annotated clinical entities in the IOB2 format, with respect to SNOMED-CT and ICD-10 respectively.

The taxonomy-driven annotation is based, for each concept, on the specific linguistic realization that is coded in the taxonomy, whereas in texts we can find a number of different textual realizations of the same concept. Variability may relate to the alternation between singular and plural

and between similar prepositions, or to the presence/omission of a preposition or article. In E3C we have devised a set of rules to account for the variability of linguistic expressions. For instance, looking at the excerpt in Section 3, the textual realization *abdominal pains* is associated with the singular SNOMED-CT concept [abdominal pain]. In addition, if overlapping portions of text match different concepts, we select the most specific one; for instance, [chest pain] is preferred over [pain].

The E3C guidelines for taxonomy-driven annotation are based on both the ShARe (Elhadad et al., 2012) and the ASSESS CT annotation guidelines⁷ (Miñarro-Giménez et al., 2018).

4.3 Language-dependent Decisions

Semantic annotation of the E3C corpus is largely language-independent. However, as we are dealing with morpho-syntactically diverse languages, we have added additional annotation guidelines for each language. These guidelines respond mainly to the annotation of the extent of the temporal and clinical entities, since their semantic features are not altered by the morpho-syntactic features.

Both the THYME-driven and the taxonomy-driven annotation schemes were originally developed for English, a language whose morphology is not particularly rich compared to the other languages of the E3C corpus (especially the Basque language). For all these, it was therefore necessary to define language specific guidelines handling the annotation of semantically complex tokens resulting from the combination of different elements (e.g., a preposition and an article)⁸.

In the case of romance languages (Italian, French and Spanish), we have taken decisions on the annotation of preposition+article contractions. The article may be part of the extent of time expressions, RML, actors and body parts, whereas the preposition should not be included. When a contraction is present, though, we have decided to capture it inside the extent (1–3).

- (1) [Nel condotto uditivo esterno] si evidenziava una lesione. ([In the external ear canal] an injury was observed.)

⁷The ASSESS CT annotation guidelines can be found at <https://user.medunigraz.at/jose.minarro-gimenez/docs/assessct/AnnotationGuidelines.pdf>

⁸It is to be remembered that the annotations in the E3C corpus are performed at token-level.

	THYME	Taxonomy	
		SNOMED-CT	ICD-10
2020-09-01	B-TIMEX3	O	O
The	B-ACTOR	O	O
patient	I-ACTOR	O	O
enters	B-EVENT	O	O
the	O	O	O
ER	O	O	O
due	O	O	O
to	O	O	O
abdominal	O	B-ENTITY	B-ENTITY
pains	B-EVENT	I-ENTITY	I-ENTITY
.	O	O	O
He	B-ACTOR	O	O
reports	B-EVENT	O	O
chest	B-BPART	B-ENTITY	B-ENTITY
pain	B-EVENT	I-ENTITY	I-ENTITY
5	B-TIMEX3	O	O
days	I-TIMEX3	O	O
ago	I-TIMEX3	O	O
.	O	O	O

Table 3: Annotation of the excerpt in Section 3 in IOB2 format.

- (2) Nous recommandons un suivi [des malades guéris du COVID-19]. (*We recommend a follow up [of the patients cured from Covid-19].*)
- (3) El drenaje [del flanco izquierdo] se retiró [al día 16]. (*The drainage [of the left side] was withdrawn [at day 16].*)

Basque, on the other hand, is a highly agglutinative language in which information expressed by prepositions in Indo-European languages is expressed by postpositions. Most of those postpositions appear attached to the nouns, adjectives, verbs and adverbs they refer to, while there is also a small set of free postpositions. The attached postpositions are taken inside the extent of the tags in E3C (4), while the free postpositions are left unmarked (5).

- (4) [Larriagotzeetan] infekzio estrep-tokozikoa izana zuen. (*[In the wors-enings] s/he had also had streptococcal infections.*)
- (5) Tik motoreak zeuzkan, [bizpahiru urtez] **geroztik**. (*S/he had motor tics, **after** [two-three years].*)

4.4 Discussion

The two annotation levels can be mapped to address specific tasks, or to develop applications that need to exploit both. Within the E3C project, we are exploring the main issues that emerge when

trying to exploit the two annotation levels at the same time. Our future aim within the project is to select a specific task and implement a mapping tailored to that task.

The main mapping issue is determined by non-matching annotated spans. Given that more specific (typically longer) taxonomy concepts are preferred to more generic ones, and that in THYME only the syntactic head of events is marked, in many cases the span of the concept is longer than the span of the event. Compare, for example, the SNOMED-CT concept associated with *abdominal pains* and the THYME event *pains* in Table 3.

More interestingly, in some cases, we can have two separate THYME annotations within the span of a single taxonomic concept. Back to our example, the SNOMED-CT concept [chest pain] overlaps with the two separate THYME annotations *pain* and *chest*.

Another issue is the inevitably different classification criteria in medical taxonomies and THYME. For instance, only a minimal part of what is marked as an event in THYME is a child concept of [event] in SNOMED-ct (e.g., *abuse* and *death*); in most cases what is marked as an event in THYME belongs to a different subpart of the SNOMED-CT hierarchy (for instance, *pain* is part of the [finding] subhierarchy, not of [event]).

5 Conclusions and Future Work

We presented the E3C project, which aims to become a reference European corpus of annotated clinical cases. We focused on two initial achievements: (i) a freely available collection of clinical cases in five languages; and (ii) a comprehensive annotation scheme based both on temporal information and on medical taxonomies.

Our next steps include the extensive manual annotation of the clinical cases in all five languages, and the definition of tasks and baselines on top of the annotated data, taking advantage of neural models derived from training data. More specifically, we plan to target the automatic construction of clinical timelines and question answering over clinical cases.

Acknowledgements

This work has been partially funded by the ELG project (EU grant no. 825627) and by the Basque Government post-doctoral grant POS_2019_1.0030.

References

- Kevin Donnelly. 2006. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. In Lodewijk Bos, Laura Roa, Kanagasingam Yogesan, Brian O’Connell, Andy Marsh, and Bernd Blobel, editors, *Medical and Care Compunetics 3*, volume 121 of *Studies in Health Technology and Informatics*, chapter 31, pages 279–290. IOS Press, Amsterdam The Netherlands.
- Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, 125:37–46.
- Noémie Elhadad, Guergana Savova, Wendy Chapman, Glenn Zaramba, David Harris, and Amy Vogel. 2012. ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes. Technical report, Columbia University.
- Natalia Grabar, Cyril Grouin, Thierry Hamon, and Vincent Claveau. 2019. Recherche et extraction d’information dans des cas cliniques. Présentation de la campagne d’évaluation DEFT 2019. In *Actes du Défi Fouille de Textes 2019*, pages 7–16, Toulouse, France. Actes DEFT 2019.
- Ander Intxaurreondo, Montserrat Marimón, Aitor González-Agirre, José Antonio López-Martín, Heidy Rodríguez, Jesús Santamaría, Marta Villegas, and Martin Krallinger. 2018. Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 280–289, Seville, Spain. Spanish Society for Natural Language Processing.
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.
- Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrison Scells, and João Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 322–339, Cham. Springer International Publishing.
- Salvador Lima López, Naiara Pérez, Montse Cuadros, and German Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France, May. European Language Resources Association.
- Yuan Luo, William K. Thompson, Timothy M. Herr, Zexian Zeng, Mark A. Berendsen, Siddhartha R. Jonnalagadda, Matthew B. Carson, and Justin Starren. 2017. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety*, 40:1075–1089.
- José Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. 2018. Qualitative analysis of manual annotations of clinical text with SNOMED CT. *PLoS ONE*, 13(12).
- Olatz Perez de Viñaspre and Maite Oronoz. 2015. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC medical informatics and decision making*, 15 Suppl 2.
- Sarah Schulz, Jurica Ševa, Samuel Rodríguez, Malte Ostendorff, and Georg Rehm. 2020. Named Entities in Medical Case Reports: Corpus and Experiments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4495–4500, Marseille, France. European Language Resources Association.
- Manuela Speranza and Begoña Altuna. 2020. E3C annotation guidelines. Technical report, Fondazione Bruno Kessler.
- World Health Organization WHO. 2015. *International statistical classification of diseases and related health problems*. World Health Organization, 10th revision, fifth edition, 2016 edition.

Monitoring Social Media to Identify Environmental Crimes through NLP A Preliminary Study

Raffaele Manna, Antonio Pascucci, Wanda Punzi Zarino, Vincenzo Simoniello, Johanna Monti

UNIOR NLP Research Group

University “L’Orientale”

Naples, Italy

[rmanna, apascucci, jmonti]@unior.it

[w.zarino, vincenzosimoniello]@gmail.com

Abstract

This paper presents the results of research carried out on the *UNIOR Eye* corpus, a corpus which has been built by downloading tweets related to environmental crimes. The corpus is made up of 228,412 tweets organized into four different subsections, each one concerning a specific environmental crime. For the current study we focused on the subsection of waste crimes, composed of 86,206 tweets which were tagged according to the two labels *alert* and *no alert*. The aim is to build a model able to detect which class a tweet belongs to.

1 Introduction

In the current era, social media represent the most common means of communication, especially thanks to the speed with which a post can go viral and reach in no time every corner of the globe. The speed with which information is produced creates an abundance of (linguistic) data, which can be monitored and handled with the use of hashtags (#). Hashtags are user-generated labels, which allow other users to track posts with a specific theme on Twitter. Moreover, social media such as Twitter can be powerful tools for identifying a variety of information sources related to people’s actions, decisions and opinions before, during and after broad scope events, such as environmental disasters like earthquakes, typhoons, volcanic eruptions, floods, droughts, forest fires, landslides (Imran et al., 2015; Maldonado et al., 2016; Corvey et al., 2010). In light of the above,

our aim is to monitor social media in order to detect environmental crimes.

Our research is guided by the following question: can Natural Language Processing (NLP) represent a valuable ally to identify these kinds of crimes through the monitoring of social media? For this purpose, we compiled a corpus of tweets starting from a list of 41 terms related to environmental crimes, e.g. *combustione illecita* (*illicit combustion*), *rifiuti radioattivi* (*radioactive waste*), *discarica abusiva* (*illegal dumping*), and we used the Twitter API to download all the tweets (specifically 228,412) related to these terms introduced by hashtag. In this research, a special focus is dedicated to the tweets related to *La terra dei fuochi* (literally *the Land of Fires*) (Peluso, 2015), a large area located between Naples and Caserta (in the South of Italy) victim of illegal toxic wastes dumped by organized crime for about fifty years and routinely burned to make space for new toxic wastes.

In order to achieve our purpose, we trained different machine learning algorithms to classify report emergency text and user-generated reports. The paper is organized as follows: in Section 2 we discuss Related Work, in Section 3 we present the *UNIOR Earth your Estate* (*UNIOR Eye*) corpus. The case study is described in Section 4 and Results are discussed in Section 5. Conclusions are in Section 6 along with directions for Future Work.

2 Related Work

As previously mentioned, hashtags are one of the most important resources - if not the most important - in text data such as those of Twitter. The possibility to aggregate data according to their content allows users to monitor all the discussion about a specific subject in real-time (an emblematic case is the hashtag *#Covid_19*).

Concerning the topic of our research, namely environmental issues, the most representative and

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

productive hashtags have proved to be *#terradeifuochi* and *#rifiuti* (respectively with a frequency of 92,322 and 62,750 occurrences), that directly refer to circumstances that have a strong impact on the environment and people's health. The use of hashtags proved to be useful in monitoring natural disasters, such as earthquakes, flood and hurricane.

For a survey on information processing and management of social media contents to study natural disasters, see (Imran et al., 2016). (Neubig et al., 2011) focused on the 2011 East Japan earthquake. The scholars built a system able to extract the status of people involved in the disaster (e.g. if they declared to be alive, they request for help, their information requests, information about missing people). About one hundred scholars participated spontaneously in the project *ANPI-NLP* (ANPI means Safety in Japanese) and the results show convincing performances by the classifier they built. (Maldonado et al., 2016) investigated natural disasters in Ecuador, monitoring Twitter to filter contents according to four different categories: volcanic, telluric, fires and climatological. The filtering process is based on keywords related to the four categories. The scholars released a web application that graphically shows the database evolution. The efficiency of the tweet filtering algorithm that they developed is expressed in terms of precision (%93.55). (Tarasconi et al., 2017) investigated tweets related to eight different event types (floods, wildfires, storms, extreme weather conditions, earthquakes, landslides, drought and snow) in Italian, English and Spanish. The corpus is composed of 9,695 tweets and can be extremely useful to perform information extraction in the aforementioned three languages. (Sit et al., 2019) used the Hurricane Irma, which devastated Caribbean Islands and Florida in September 2017, as a case-study: the scholars demonstrate that by monitoring tweets it is possible to detect potential areas with high density of affected individuals and infrastructure damage throughout the temporal progression of the disaster. By focusing on tweets generated before, during, and after Hurricane Sandy, a superstorm which severely impacted New York in 2012, (Stowe et al., 2016) proposed an annotation schema to identify relevant Twitter data (within a corpus of 22.2M unique tweets from 8M unique Twitter users), categorizing these tweets into fine-grained categories,

such as *preparation* and *evacuation*. (Imran et al., 2016) presented Twitter corpora composed of over 52 million crisis-related tweets, collected during 19 different crises that took place from 2013 to 2015. These corpora were manually-annotated by volunteers and crowd-sourced workers providing two types of annotations, the first one related to a set of categories, the second one concerning out-of-vocabulary words (e.g. slangs, places names, abbreviations, misspellings). The scholars then built machine-learning classifiers in order to demonstrate the effectiveness of the annotated datasets, also publishing *word2vec* word embeddings trained on more than 52 million messages. The preliminary results of this study posit that a classification with a high precision of tweets relevant to the disaster is possible to assist crisis managers and first responders. Our study is not devoted to monitor natural disasters but to monitor natural human-caused disasters. More specifically, the aim is to exploit NLP techniques to contribute to the identification of intentional environmental crimes through social media analysis. To the best of our knowledge, this perspective of investigation is rather novel in the field.

3 The *UNIOR Eye* Corpus

This section outlines the way the *UNIOR Eye* corpus was created and how it is internally structured. The research has been carried out in the framework of the C4E - Crowd for the Environment (Progetto PON Ricerca e Innovazione 2014-2020) project².

The *UNIOR Eye* corpus is made up of 228,412 tweets related to environmental crimes downloaded through Twitter API, covering the period from 01 January 2013 to 06 August 2020. The compilation phase of the corpus was divided into two steps: the creation of a vocabulary containing keywords related to environmental crimes and the creation of the corpus. During this work phase, the data was structured and organized according to the different keywords, obtained from glossaries and documents specific to the topic.

Precisely, the following resources

- *Glossario di termini sull'ambiente* (FIMP, 2017) (a guide from A to Z concerning the complex issue of environmental pollution);

²<http://www.unior.it/ateneo/20574/1/c4e-crowd-for-the-environment-progetto-pon-ricerca-e-innovazione-2014-2020.html>

- *Glossario dinamico per l'Ambiente ed il Paesaggio* (IS-PRA, 2012) (a glossary supplied by the Italian Institute for Environmental Protection and Research);
- *Glossario ambientale*³ (a glossary supplied by the national agency for the environmental protection of Tuscany);
- *BeSafeNet*⁴ (a glossary based on the *Glossary on Emergency Management*, which has been developed in 2001 by European Centre of Technological Safety (TESEC) of Euro-Mediterranean network of Centres EUR-OPA Major Hazard Agreement of Council of Europe in collaboration with other centres of network);
- *HERAmbiente*⁵ (a glossary provided by Herambiente, the largest company in the waste management sector);
- *Enciclopediaambiente*⁶ (the first freely available online Encyclopedia on the Environment, designed by a group of four engineers with the aim of spreading “environmental knowledge”)

and the following two web sources

- a dossier containing important provisions aimed at dealing with environmental and industrial emergencies and encouraging the development of the affected areas⁷;
- a document on environmental crimes and environmental protection⁸.

were consulted. All of these language resources contain information and definitions of the basic terms related to environmental disasters and crimes, e.g. *Rifiuti pericolosi* (hazardous waste): waste products which can generate potential/substantial risk to human health/the environment if handled improperly. Hazardous waste contains at least one of these characteristics: flammability, corrosivity, or toxicity,⁹ and is included in special lists. Here are some examples.

- *HASHTAG HASHTAG Fiumicino: eternit e rifiuti pericolosi al Passo della Sentinella* URL *HASHTAG HASHTAG Fiumicino: eternit and hazardous waste in Passo della Sentinella* URL *HASHTAG*);
- *Cani in gabbia in discarica abusiva: Due animali tra rifiuti pericolosi, amianto e bombole gas* URL (Caged

³<http://www.arp.at.toscana.it/glossario-ambientale>

⁴<http://www.besafenet.net/it-it/glossary>

⁵http://ha.gruppohera.it/glossario_ambiente/

⁶<http://www.enciclopediaambiente.com>

⁷https://www.senato.it/japp/bgt/showdoc/17/DOSSIER/0/740667/index.html?part=dossier_dossier1-sezione_sezione12-h2_h28

⁸https://scuola21.fermi.mn.it/documenti/reati_ambientali.pdf

⁹<https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2018:124:FULL&from=IT>

dogs in illegal dump: two pets among hazardous waste, asbestos and gas cylinders URL)

After this phase it was possible to create the corpus by downloading from Twitter all the tweets containing these keywords preceded by the hashtag. These hashtags helped us to gather the information needed to detect crimes against the environment. More specifically, the corpus is internally divided into semantic areas, each one concerning a specific environmental crime: *rifiuti e terra dei fuochi* (waste and Terra dei fuochi); *reati contro le acque* (water-related crimes); *materiali e sostanze pericolose* (hazardous substances and materials); *incendi e roghi ambientali* (environmental fires). These sets are further divided into more specific subsets, e.g. the folder *reati contro le acque* (water-related crimes) contains the subsets *acque di scarico*, *acque reflue*, *fiumi inquinati*, *liquami* (sewage, wastewater, polluted rivers, slurry). The resulting corpus contains, therefore, a total of 228,412 tweets, 22,780,746 tokens, 569,905 types with a type/token ratio (TTR) of 0.025.

4 Case Study

This section describes the steps taken to perform the preliminary experiments on a selected part of the *UNIOR Eye* corpus. First, the dataset on which the experiments and data preparation operations were carried out is presented, then the pre-processing steps are listed and, finally, the different machine learning approach used are described.

4.1 Dataset

As described in Section 3, the *UNIOR Eye* corpus is divided into four semantic areas related to the most common crimes against the environment. Among the four semantic areas, we decided to use the waste crimes subsection to test a specific use case: whether an NLP system can understand and report emergency text and user-generated reports. Therefore, for the experiments described in this paper, we focus our investigation on a sub-section of the *UNIOR Eye* corpus, namely tweets about waste related crimes and tweets with the hashtag *#terradeifuochi* contained in the corresponding semantic area: *waste and Terra dei fuochi*. This subsection of the corpus contains 86,206 tweets. First, for the total number of tweets, hashtags, mentions and URLs are replaced with placeholder words. Then tweets were annotated by the paper authors on the basis of two labels: i) alert and ii) no alert,

i.e. if the tweet contains or not a message aimed at reporting and locating a waste related crime. Below, we provide a sample of annotated tweets following our two labels, **alert - no alert**:

- *Ore 11:40 autostrada A1 altezza Afragola Acerra direzione Roma. Roghi Tossici indisturbati, la HASHTAG... URL HASHTAG HASHTAG (11:40 am A1 motorway near Afragola Acerra towards Rome. Undisturbed toxic fires, the HASHTAG ... URL HASHTAG HASHTAG) — **ALERT***
- *MENTION ministro, piuttosto che pensare alla HASHTAG pensi ai continui roghi MENTION (MENTION Minister, rather than thinking about the HASHTAG think about the continuous fires MENTION) — **NO ALERT***

During the annotation phase, we noted that the **no alert** class is the one which contains the majority of tweets and includes examples of hate speech, satirical texts, news about emergency actions as well as politically oriented texts. Consequently, our dataset built in this way is unbalanced for the two classes, counting 81,235 tweets for the **no alert** class and 4,970 **alert** tweets. In order to visualize alert tweets, we exploit Carto¹⁰, a cloud computing platform that provides a geographic information system, web mapping, and spatial data science tools¹¹.

4.2 Inter-annotator Agreement

When different annotators label a corpus, it is important to calculate the inter-annotator agreement (IAA) with a twofold objective: i) make sure that annotators agree and ii) test the clarity of guidelines. As previously mentioned, the dataset (composed of 86,206 tweets) has been annotated by four of the paper authors on the basis of two labels: i) alert and ii) no alert. This implies that each author annotated about 21,000 tweets. Then, to calculate inter-annotator agreement we randomly selected 10% of the tweets (i.e. 8,620) which were tagged by all annotators.

The agreement among the four annotators is measured using Krippendorff's α coefficient; instead, to estimate the agreement between pairs of annotators, we use Cohen's κ coefficient (Artstein and Poesio, 2008). Taking into account the recommendations set out in (Artstein and Poesio, 2008; Krippendorff, 2004), we interpret the κ values obtained

in IAA according to the strength of agreement criteria described in (Landis and Koch, 1977) for each pair of annotators; whereas, for agreement among four annotators, we follow the suggested standard in (Krippendorff, 2004). The calculated value of Krippendorff's α is **0.706**. Considering the standard value in (Krippendorff, 2004), our value of $\alpha=0.706$ is considered as acceptable and expressing a good data reliability. In Table 1 we show the results for pairs of annotators.

Pair of annotators	Value of κ
a1 - a2	0.691
a1 - a3	0.742
a1 - a4	0.841
a2 - a3	0.676
a2 - a4	0.644
a3 - a4	0.641

Table 1: Cohen's κ values for pairs of annotators.

According to (Landis and Koch, 1977), five out of six Cohen's κ values show a "substantial" strength of agreement for each pair; while a pair (a1-a4) show a κ value considered "almost perfect" in the research cited.

4.3 Preprocessing

Before feeding the machine learning algorithms, some pre-processing steps are performed. Since the majority of mentions and hashtags are shared by both alert and no-alert samples, we focus on the tweet itself, by removing any reference to people, entities and organizations conveyed through hashtags and mentions. Therefore, the placeholder words related to hashtags, URLs and mentions are removed. Then, punctuation is removed from the tweets along with a custom list of function words such as determiners, prepositions and conjunctions. Finally, the tweets are lower-cased and the tokenization is performed.

4.4 Machine Learning Approaches

We set the problem of tweets related to waste crimes as a supervised binary classification problem between different textual content.

To tackle the problem as first task within the *C4E Project*, we select a machine learning approach using Support Vector Machines (SVM) with linear kernel and $C=1$ and Multinomial Naive Bayes (MNB) as classification algorithms (Imran et al., 2015). Since the task concerns the classification

¹⁰carto.com

¹¹A map showing toxic fires alert tweets in the UNIOR Eye corpus is available at this link <https://uniornlp.carto.com/builder/04f2cca9-08cd-4b9f-90cd-79fc0d93af42/embed>

of tweets belonging to the **alert** class, to deal with the unbalanced dataset, we use the undersampling technique by automatically reducing the number of samples for the majority class (**no alert**) (Li et al., 2009), until they were balanced with the samples of the **alert** class. We used the tf-idf technique to extract the features used by both algorithms. To build algorithms and extract features, we used the Python scikit-learn library.

In addition to the MNB and SVM with tf-idf technique, we built two models with sentence embeddings as features and SVM with the tuning of C parameter as a classification algorithm. In the first model (FT-SVM), we used the Italian pre-trained word vectors from *fastText*¹²(Bojanowski et al., 2017) to build our sentence embeddings by averaging word embeddings for all tokens for each tweet; then, $C=10$ is found as the best C parameter value using GridSearchCV¹³ instance. In the second model (mDB-SVM), we generated sentence embeddings using the pretrained multilingual DistilBERT (Sanh et al., 2019) model from *Transformers*¹⁴. To accomplish this, each tweet is represented as a list of tokens and, then, each list is padded to the same size ($max_len = 94$). The attention mask is used. Before fitting the sentence embeddings thus constructed in the SVM classifier, it is searched for the best value of the C parameter set to $C=0.1$. For both models (FT-SVM and mDB-SVM) the pre-processing steps described above are performed.

5 Results

In this section, we show the results obtained by our models in terms of Precision, Recall, F-Measure and Accuracy. For all models, the results are obtained on 30% of the dataset set aside as a test set, keeping the samples balanced between the two classes. Furthermore our models were evaluated using a 10-Fold Cross-Validation¹⁵.

As a baseline to compare with, we used Dummy classifier which achieves an accuracy of 0.501. On the test set, the SVM classifier achieves an accuracy of 0.870, while for the MNB classifier it is 0.839. Regarding the evaluation by 10-fold cross

validation, our SVM reaches an accuracy of 0.868 with the mean and standard deviation of 0.008, instead the accuracy of the MNB is 0.841 with the mean and standard deviation of 0.010. In Table 2 we show the performances achieved by both models.

MNB	Precision	Recall	F-Measure
<i>alert</i>	0.871	0.816	0.843
<i>no alert</i>	0.807	0.864	0.835
SVM	Precision	Recall	F-Measure
<i>alert</i>	0.857	0.878	0.867
<i>no alert</i>	0.883	0.862	0.873

Table 2: Results in terms of Precision, Recall and F-Measure.

Both classifiers with tf-idf achieve good accuracy and seem to have a good ability to classify a considerable amount of tweets providing good results in terms of precision and recall. One of the reasons for these performances may be ascribed to a discriminating lexical composition regarding the samples belonging to the **alert** and **no alert** classes.

Regarding the accuracy of sentence embeddings models on the test set, FT-SVM reaches an accuracy of 0.822, while mDB 0.774. By evaluating the predictive performance of the two models with 10-fold cross-validation, FT-SVM achieves an accuracy of 0.825 with the mean and standard deviation of 0.011, while mDB-SVM reaches the accuracy of 0.773 with the mean and standard deviation of 0.013. In Table 3, the results in terms of Precision, Recall and F-Measure are shown.

FT-SVM	Precision	Recall	F-Measure
<i>alert</i>	0.826	0.817	0.821
<i>no alert</i>	0.818	0.827	0.822
mDB-SVM	Precision	Recall	F-Measure
<i>alert</i>	0.785	0.766	0.775
<i>no alert</i>	0.765	0.783	0.774

Table 3: Classification Reports for FT-SVM and mDB-SVM.

Both models fed with sentence embeddings constructed with different techniques, seem to perform well in this classification task. In particular, the FT-SVM model based on sentence embeddings built with *FastText* seems to have better scores in terms of Precision and F-measure than those achieved by the mDB-SVM model. One

¹²<https://fasttext.cc/docs/en/pretrained-vectors.html>

¹³https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#

¹⁴https://huggingface.co/transformers/pretrained_models.html

¹⁵https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

of the reasons could be that sentence embeddings built with *FastText* benefit from a resource tailored on the Italian language compared to a multilingual one used in DBert-SVM. Specifically, mDB-SVM achieved good results in terms of precision and f-measure for the alert class. Instead, in terms of Recall, both models have a high proportion of relevant instances for the no alert class.

5.1 Confusion Matrices

In this section we show the four confusion matrices in order to graphically display the performances achieved by the different models. In Figure 1 we show the confusion matrix of the MNB model, while in Figure 2 that of the SVM model.

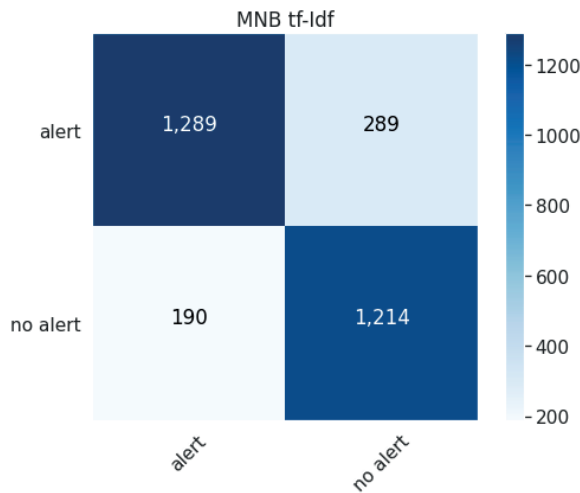


Figure 1: MNB model confusion matrix.

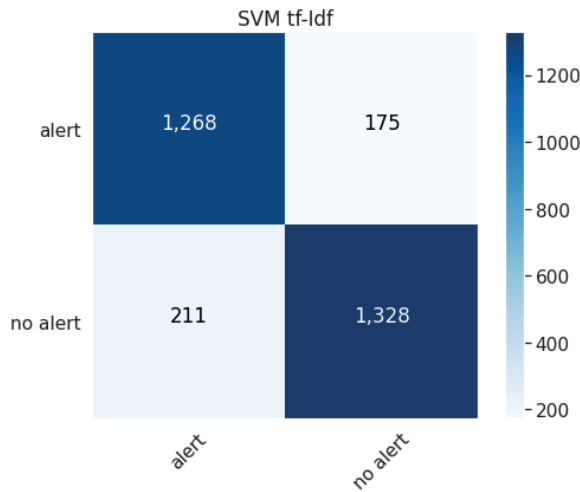


Figure 2: SVM model confusion matrix.

The confusion matrices of the FT-SVM and the mDB-SVM model are shown respectively in Figure 3 and Figure 4.

Figure 3 and Figure 4.

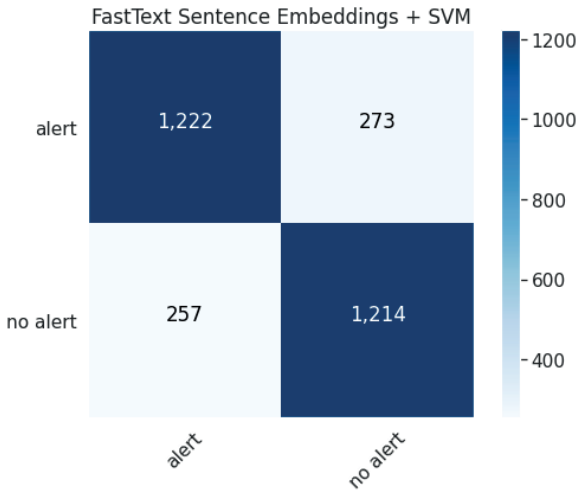


Figure 3: FT-SVM model confusion matrix.

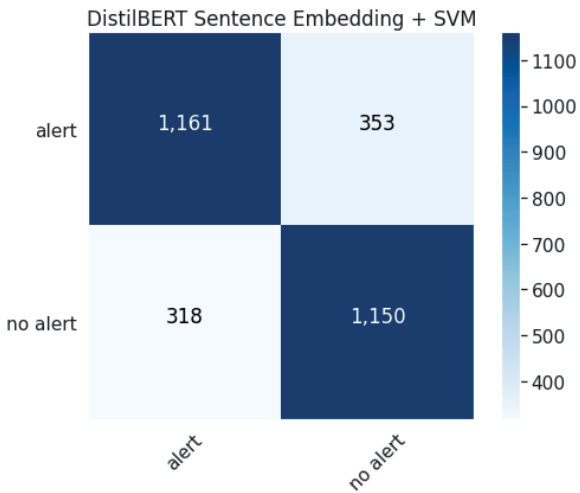


Figure 4: mDB-SVM model confusion matrix.

6 Conclusions and Future Work

We presented a case study within the C4E project aimed at monitoring social media to provide support against environmental crimes. In particular, we described the *UNIOR Eye* corpus, in some sections still in progress, on which we tested four models with three different features extraction and construction techniques on a part of the corpus. We proposed two classifiers, namely SVM and MNB, with tf-idf features as the first experiment; then, SVM with C parameter tuning fed with sentence embeddings. These embeddings were built both using Italian pre-trained *fastText* model and using pre-trained DistilBert multilingual model. Our purpose was to classify alert tweets related

to waste crimes vs no alert tweets. Future research will include the enlargement of the corpus, applications of NLP in the field of environmental protection as well as the analysis of contextual features related to environmental issues used as a medium to polarize public opinion (Karol, 2018).

Acknowledgements

This research has been carried out within the framework of two Innovative Industrial PhD projects supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds and two research grants supported by the PON Ricerca e Innovazione 2014/20 in the context of the C4E project.

Authorship contribution is as follows: Raffaele Manna is author of section 4. Section 2 is by Antonio Pascucci. Section 5 is by Raffaele Manna and Antonio Pascucci. Sections 1, 3 and 6 are by Wanda Punzi Zarino and Vincenzo Simoniello. We are grateful to Prof. Johanna Monti for supervising the research.

References

- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- William J Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. 2010. Twitter in mass emergency: What nlp can contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 23–24.
- FIMP. 2017. Fimp ambiente - federazione italiana medici pediatri glossario di termini sull'ambiente. una guida dalla a alla z per orientarsi nel complesso tema dell'inquinamento ambientale.
- Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38.
- Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.
- ISPRA. 2012. Ispra – l'istituto superiore per la protezione e la ricerca ambientale, glossario dinamico per l'ambiente ed il paesaggio.
- David Karol. 2018. Party polarization on environmental issues: toward prospects for change. *Research Paper. Niskanen Center, Washington, DC*.
- Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. 2009. Adapting svm for data sparseness and imbalance: a case study in information extraction. *Natural Language Engineering*, 15(2):241–271.
- Miguel Maldonado, Darwin Alulema, Derlin Morocho, and Marida Proaño. 2016. System for monitoring natural disasters using natural language processing in the social network twitter. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–6. IEEE.
- Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining—what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 965–973.
- Pasquale Peluso. 2015. Dalla terra dei fuochi alle terre avvelenate: lo smaltimento illecito dei rifiuti in italia. *Rivista di Criminologia, Vittimologia e Sicurezza*, 9(2):13–30.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Muhammed Ali Sit, Caglar Koylu, and Ibrahim Demir. 2019. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma. *International Journal of Digital Earth*, 12(11):1205–1229.
- Kevin Stowe, Michael Paul, Martha Palmer, Leysia Palen, and Kenneth M Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The fourth international workshop on natural language processing for social media*, pages 1–6.
- Francesco Tarasconi, Michela Farina, Antonio Mazzei, and Alessio Bosca. 2017. The role of unstructured data in real-time disaster-related social media monitoring. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3769–3778. IEEE.

Does finger-tracking point to child reading strategies?

Claudia Marzi

ILC - CNR Pisa

`claudia.marzi@ilc.cnr.it`

Anna Rodella

Università di Pisa

`a.rodella@studenti.unipi.it`

Andrea Nadalini

ILC - CNR Pisa

`andrea.nadalini@ilc.cnr.it`

Loukia Taxitari

ILC - CNR Pisa

`loukia.taxitari@ilc.cnr.it`

Vito Pirrelli

ILC - CNR Pisa

`vito.pirrelli@ilc.cnr.it`

Abstract

The movement of a child's index finger that points to a printed text while (s)he is reading may provide a proxy for the child's eye movements and attention focus. We validated this correlation by showing a quantitative analysis of patterns of "finger-tracking" of Italian early graders engaged in reading a text displayed on a tablet. A web application interfaced with the tablet monitors the reading behaviour by modelling the way the child points to the text while reading. The analysis found significant developmental trends in reading strategies, marking an interesting contrast between typically developing and atypically developing readers.

1 Introduction

Recent experimental evidence in visual perception analysis (Lio et al., 2019) shows that eye movements and finger movements strongly correlate during scene exploration, at both individual and group levels. In Lio *et al.*'s (2019) experiment, subjects are invited to explore a blurred image displayed on a touchscreen by moving their fingers on the display. Picture areas that are located immediately above the touch point of the subject's finger on the screen are automatically shown in high resolution, thereby simulating the subject's central (foveal) vision. The experiment proves that the subjects' image-exploring patterns in the two modalities (optical and tactile) are highly congruent. The result is not surprising. A familiar context which exploits this synergistic behaviour is when

children are learning to read. Despite the undoubtedly different dynamics of the two types of text exploration, finger-pointing to text helps children learn to look at print, and supports critical early reading behaviours: directional movement, attention focus, and voice-print match (Mesmer and Lake, 2010; Uhry, 2002).

ReadLet (Ferro et al., 2018a; Ferro et al., 2018b) is a web application with a tablet front-end, designed to support online monitoring of silent and oral reading abilities through *finger-tracking*. Finger-tracking consists of recording the time series of touch events on the tablet screen where a child is reading a short story, while the child is pointing to the text with the index finger of her dominant hand.² Preliminary analyses of our finger-tracking data (Pirrelli et al., 2020) highlighted a diminishing influence of word frequency and word length on reading time as readers get older and more proficient (from 3rd to 6th grades). With increasing exposure to written words, differences in tracking time between high and low frequency words gradually tend to decrease, suggesting a ceiling effect in the entrenchment of both high- and low-frequency lexical representations in long-term memory (Zoccolotti et al., 2009). Similarly, word length was found to significantly interact across grades. Younger readers show increasing difficulty with longer words, with a steeper time increment for word length > 6, while older readers are slowed down when words are longer than 8 letters. This integrates previous evidence (De Luca et al., 2008), confirming that not even the most experienced readers can avoid the slowing down effect of word length.

The two-fold interaction of word frequency and word length with grade levels strongly suggests that Italian children use a *lexical route* to decod-

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.readlet.it/>: see section *publications* for materials related to this paper.

ing even at early stages of their reading development, despite the transparency of Italian orthography (Bates et al., 2001; Davies et al., 2013). It also suggests that young readers can make use of sublexical information, whenever they are confronted with words that are not contained in their orthographic lexicon. This developmental dynamic would account for the stronger sensitivity of less skilled readers to both lexical frequency and word length. As lexical information increases with age, with rarer and longer words finding their way into the reader's orthographic lexicon, the reader makes an increasingly prominent use of lexical information and an increasingly sparser use of sublexical information.

In this paper, we provide a finer-grained quantitative analysis of the finger-tracking profile of typically developing readers, offering further evidence that their reading strategy results from an optimal, interactive combination of both lexical and sublexical information. The evidence is compared with the finger-tracking profile of difficult readers. To provide a more realistic developmental profile of these effects, we restrict our focus on nouns only, which are less likely to be skipped while finger-tracking, and present a narrower range of variability in both length and frequency.

2 ReadLet

ReadLet is a tablet-based application that combines an objective assessment of a child's reading fluency and comprehension skills with careful collection of large-scale behavioural data, and quantitative modelling of the specific factors affecting reading development. It leverages an ICT infrastructure with a cloud-based back-end exposing a battery of web services acting as an interface between the central repository and the users. The ReadLet front-end is an ordinary tablet, where short stories are displayed for children to read, either silently or aloud. In both cases, the child is asked to finger-point to the text while reading. Texts are displayed on a 10" tablet screen in Lato font 17pt in black against a white background. During each reading session, the behaviour of the child is captured through large streams of time-aligned signals including voice recording, time-stamped finger-tracking patterns, reading time and question-answering time. Data are automatically captured and sent to a centralised server for post-

processing, where audio and finger-tracking time series are aligned with the text. Recorded and post-processed data are exposed through a set of web services offered by the cloud server.

3 The Data

For our present goal, we focus on reading data of 237 children, sampled from entire classes ranging from 3rd through to 6th school grades, in Italian and Italian-speaking Swiss schools.³ Participants included both *typically* developing readers (N=214) and children screened and reported in schools as *atypically developing readers* (N=23), but who did not receive a clinical diagnosis. Eight short stories were created for the pilot study, one for each of the four school grades, and for each experimental condition (silent and aloud reading). Children were asked to read a story while finger-tracking the text. After reading in the silent condition only, children were asked a few multiple-choice questions, to ascertain they actually carried out the task.

Texts were automatically annotated for part-of-speech, word token frequency, and word typicality (measured as either the size of the word's lexical neighbourhood, or *N*-size, or the mean Levenshtein distance from its top 20 neighbouring words, or *Old20* (Yarkoni et al., 2008)).

For each child, in both reading conditions, we calculated the *token tracking time* as the total time spent in finger-tracking each word token while reading. To ensure reliability and precision in the alignment of finger-tracking data with the text being read, we selected reading trials with $\geq 75\%$ of finger-tracked text pages. From the original set of tokens making up the 8 short stories, we selected 97 lemmas for 109 noun tokens, by intersecting our data with age of acquisition and imageability assessments by Italian speakers (Montefinese et al., 2014; Montefinese et al., 2019). In the resulting data sample, word frequency⁴ is observed

³Data come from a pilot data collection carried out by Claudia Cappa and Sara Giulivi in schools of Southern Tuscany and Canton of Ticino in 2018 and 2019 for the *AEREST* project, funded by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland. Data were encrypted and pseudonymized locally, to then be delivered through a *https* protocol to a centralized server, where they are accessed with authentication credentials and dedicated web services. A background history of children including information about any clinical and/or psychological disorders was provided by the schools.

⁴Token frequency is given as the natural logarithm of occurrences in the Subtlex corpus (<http://crr.ugent.be/subtlex->

to vary between min=5.61 and max=11.77, and word length between 4 and 10 letters (median=5, mean=5.62, sd=1.40).

4 Typical and atypical reading development

The main goal of the *ReadLet* project is to propose and validate an ICT methodology for assessing the typical reading development of children in Italian schools. In this section we focus on the finger-tracking behaviour of typically developing children engaged in reading a short text. The idea is to provide evidence that finger-tracking patterns exhibit lexical effects that are well-established in the reading literature: namely word frequency, word length and word similarity (or *N*-size).⁵

Figure 1 shows the effects of word frequency across grade levels, in both aloud (left panel) and silent (right panel) reading, for typically developing readers. A linear mixed model fitting token tracking time as a function of reading type, word frequency and grade levels shows shorter tracking times in reading more frequent words. The model also highlights a significant interaction between years of schooling and word frequency, with facilitation effects getting smaller for older graders, particularly in silent reading. The difference in facilitation rate between the two reading tasks is not statistically significant.

Figure 2 compares the developmental patterns of token tracking time of typically (left panel) and atypically (right panel) developing readers, modelled as a linear function of word token frequency and grade level. The two patterns exhibit a clear facilitatory effect of token frequency on reading speed, confirming that frequency makes reading consistently easier for both typical and atypical populations of young readers, who appear to entertain the same *lexical* reading strategy. However, only in typically developing children the effect tends to diminish across grade levels, with slopes getting less steep as grade levels increase (Figure 2, left panel).⁶

A similar overall pattern is shown in Figure 3, where the sensitivity to word length of typical

it/) plus one. For our set of noun data the mean frequency is 9.45 (sd=1.61).

⁵All figures in the section show regression plots of the interaction of main effects, using the *ggplot* function.

⁶Regression slopes for 4th and 5th grades are not statistically different from 3rd grade, but there is a significant difference when comparing slopes for 3rd and 6th grades.

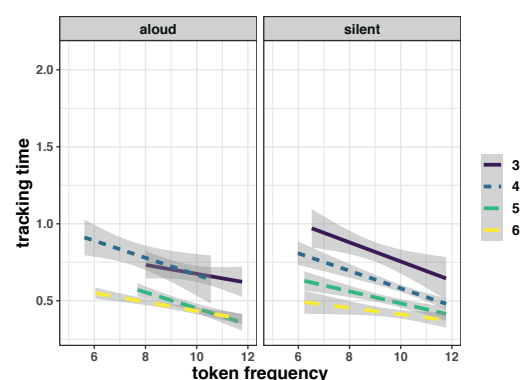


Figure 1: Linear interaction of token frequency and grade level (from 3rd to 6th level), for typically developing readers in both aloud and silent reading. Shaded areas indicate 95% confidence intervals.

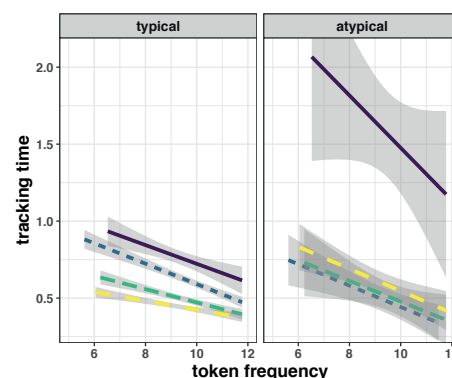


Figure 2: Linear interaction of token frequency and grade level (from 3rd to 6th level), for typically and atypically developing readers.

readers is contrasted with the same effect in atypical readers. In both groups, children take more time to read longer words, but only typically developing children exhibit a less strong sensitivity to word length as grade level increases. The statistical significance of this interaction disappears in atypical readers, with the only exception of 3rd graders, compared with all remaining graders.

Figure 4 shows how grade levels interact with *N*-size in affecting aloud (left panel) and silent (right panel) reading time. The dominant effect is facilitatory, with a clear incremental advantage in reading times for words with a high number of neighbours. Words are finger-tracked more quickly when they belong to more dense neighbourhoods, and this facilitatory effect is stronger for younger (3rd and 4th grade) than older (5th

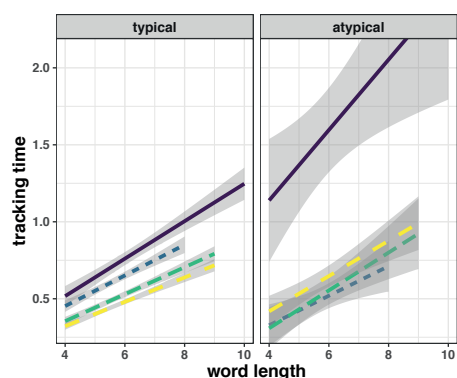


Figure 3: Linear interaction of word length and grade level (from 3rd to 6th) for typically and atypically developing readers.

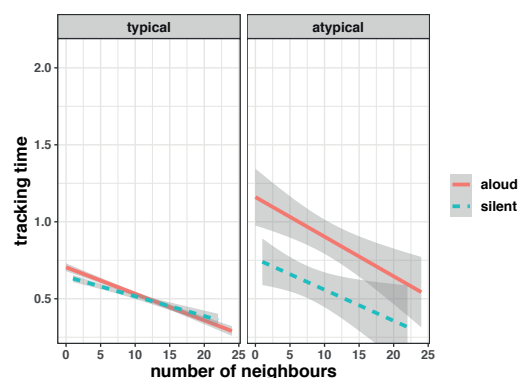


Figure 5: Linear interaction of N -size and reading task (aloud and silent) for typical and atypical readers.

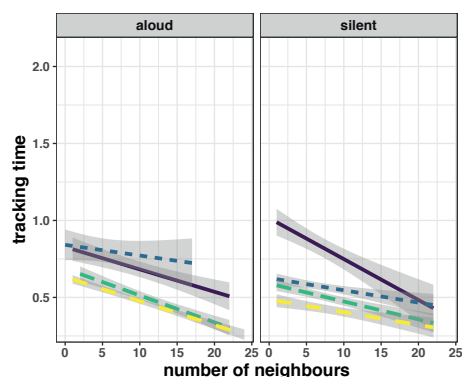


Figure 4: Linear interaction of N -size and grade level (from 3rd to 6th) for aloud and silent reading of typical readers.

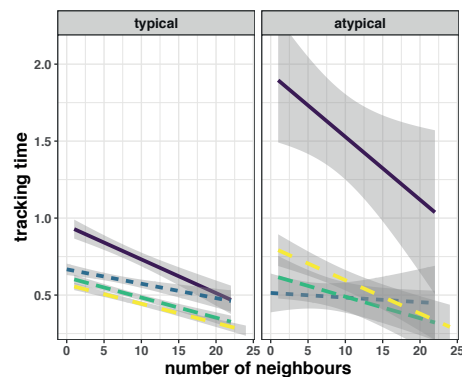


Figure 6: Linear interaction of N -size and grade level (from 3rd to 6th) for typically and atypically developing readers.

and 6th grade) readers. No significant difference is found in the interaction between reading type and N -size in typical readers (Figure 5, left panel). Atypical readers show equal slopes in both aloud and silent reading, but different intercepts, which capture the additional processing demands of concurrent articulation (Figure 5, right panel). This evidence suggests a *sublexical* reading strategy that relies on orthographic similarity: words that are not read lexically (because they are too long or less frequent), are read by decoding and combining the smaller parts they share with other neighbouring words. Fitting a mixed model with N -size, frequency ranges, and grade levels, as variables predicting the token tracking time, and with subjects as random effect, we find that all predictors and interactions are highly significant for typically developing children (Figure 6).

The behaviour of atypically developing read-

ers does not replicate the trend of typical readers. First, the tracking time of atypical readers is more strongly - and significantly - affected by token frequency, when compared with the typical tracking time of their age-matched peers. This is especially true for the youngest readers in our sample (3rd graders in the right plot of Figure 2). In addition, sensitivity to frequency appears to persist with age, as there are no significant differences in the facilitatory effect of frequency across later grade levels. This suggests a delay in developing and integrating lexical information. A nearly identical developmental pattern is replicated with N -size effects (Figure 6, right panel): younger children read words in denser neighbourhoods more easily, taking advantage of the recurrent sublexical parts shared by neighbouring words. Once more, no significant developmental pattern is observed across grade levels, as atypical readers do

not appear to be able to increasingly rely on lexical reading as they get more experienced (Figure 2). Finally, their reading time is persistently slowed down by longer words, suggesting a difficulty in memorizing and making them accessible through the lexical route (Figure 3, right panel).

5 General discussion

Facilitatory effects of lexical frequency on reading reaction time have been reported for Italian children (Barca et al., 2006; Burani et al., 2002) as well as adults (Barca et al., 2002; Burani et al., 2007). The effects are argued to reflect the working of the lexical route in dual-route models of reading (Coltheart et al., 2001): word items are accessed in the reader's orthographic lexicon, to then be pronounced after their full phonological code is retrieved. The faster reading of high-frequency lexical items thus reflects the well-established sensitivity of lexical access to word frequency. In Italian, the systematic nature of letter-to-sound mapping rules makes the operation of a sublexical reading strategy a reliable alternative to the lexical route. However effective, sublexical reading is nonetheless less efficient, since it requires the online, serial decoding of a word by its parts (e.g. *n*-grams or syllables). We conjecture that Italian children optimize reading efficiency at early stages of their reading practice, through dynamic integration of sublexical and lexical reading. Whenever possible, they resort to word-sized orthographic information in their lexicon (e.g. short and frequent words), and make it up for missing orthographic items through sublexical information.

Such an opportunistic strategy is in keeping with the idea that early readers strive, through reading practice, to "chunk" letter *n*-grams into longer orthographic units. Chunked units are stored and made accessible in the readers' lexicon, where they are associated with their fully specified phonological code. The length of stored items is a function of their frequency, and the reader's processing efficiency, reading practice and age. Our data confirm that this strategy is consistently adopted by typically developing readers in both silent and aloud reading, suggesting that the influence of lexical frequency is not confined to the retrieval and planning stage of the word phonological code, but appears to extend beyond response initiation, to affect full articulation of the code (Balota and Yap, 2006).

This strategy remains in operation through reading development, as shown by the decreasing tracking times of 6th graders as a function of lexical frequency (Figure 2, left panel). Nonetheless, the impact of word frequency is less strong in older readers, whose orthographic lexicon makes room for increasingly rarer (and longer) words. Also atypical readers appear to use a similar "chunking" strategy, but their developmental pattern fails to show a clear interaction between grade level and frequency. In Figure 2 (right panel), 3rd graders show a robust word frequency effect, but the diminishing role of frequency on tracking time across grades turns out not to be significant. This suggests that atypical readers have problems with developing orthographic representations for rarer (and longer) words, and they are not quite as successful as typical readers in optimally integrating lexical and sublexical information.

This interpretation is supported by the analysis of two other lexical effects on child reading development: word length and neighbourhood size (*N*-size). As expected, longer words elicit longer response latency and reading duration, but the effect is bigger for younger, typically developing readers compared to older ones (Figure 3, left panel), and for atypical readers compared to their age-matched peers (Figure 3, right panel). The use of sublexical information and serial *n*-gram decoding appears to be more prominent in younger and atypical readers than in the older and more skilled group of readers. Once more, the effect can be argued to reflect the absence of fully specified orthographic representations for longer words in the lexicon of less skilled readers, and a related difficulty in building up complex orthographic chunks.

Facilitatory effects of *N*-size on reading time are reported for atypical Italian readers by Marinelli et al. (2013), who, however, found no significant facilitation in age-matched typical readers. They argue that atypical readers over-rely on co-activation of word neighbours during reading to make it up for their poorly entrenched lexical representations. Conversely, access to individual lexical representations by typical readers is fast enough to make *N*-size effects hardly detectable. Our data are consistent with Marinelli et al.'s evidence, but integrate it in two important respects. First, the speeding-up influence of *N*-size is detected in both aloud and (for the first time to our knowledge) silent reading of Italian, with no

significant difference between the two (Figure 4 and Figure 5, left panel). This supports an interpretation of the *N*-size effect as having an impact on both phonological planning and overt articulation. Secondly, our data show that the effect is not limited to the reading pace of younger and atypical readers, as observed by Marinelli et al., but it also holds for typically developing readers (Figure 6), with an interesting modulation by grade level. This is mainly due to our focus on nouns, which include longer and less frequent words, for which *N*-size effects are known to be stronger and easier to detect (Davies et al., 2013). Finally, the diminishing impact of *N*-size for increasing grade levels confirms a sparser use of the sublexical route by more skilled readers, who are equipped with a richer and more efficient orthographic lexicon.

To sum up, typical and atypical readers alike strive to optimally integrate lexical and sublexical input patterns while reading, using the former whenever possible for efficient decoding, and the latter as a fall-back strategy, whenever the lexical route fails. This dynamic, however straightforward, has non-trivial consequences. In a developmental perspective, the orthographic lexicon gets richer with practice, boosted by an age-driven improvement of children's global ability in information processing (Zoccolotti et al., 2009), which makes longer and rarer words easier to store. As a result, the dynamic balance is shifted towards lexical reading. Conversely, atypical readers find it more difficult to develop and store detailed mappings between orthographic and phonological sequences, as confirmed by their greater sensitivity to frequency and length effects (Figures 2 and 3) and by a prolonged, larger effect of *N*-size on finger-tracking (Figure 5, right panel).

6 Concluding remarks

We provided evidence that finger-tracking data of reading children can highlight congruent developmental patterns in the acquisition of literacy skills. We only replicated established benchmark effects reported in the psycholinguistic literature on decoding transparent orthographies. Nonetheless, to our knowledge, this is the first time that finger-tracking patterns are shown to significantly correlate with more established reading data.

Unsurprisingly, typically developing readers were shown to read at a faster rate than atypical readers. Our comparative analysis shows that both

groups of readers are sensitive to the same lexical effects, but that atypical readers rely on an impoverished lexicon. We take this evidence to show that although the two groups adopt the same strategy, they differ in their global ability in serial information processing, which has a boosting influence on lexical development and reading speed.

Despite our promising results, one could legitimately wonder why we propose using finger-tracking as a proxy of a more established technology such as eye-tracking. Portability and task ecology are our strongest arguments. ReadLet can be used in almost any environment with no data-acquisition specialist or invasive, anxiety-provoking equipment. This has practical consequences for research in education, computer science, human cognition and medical sciences. Our architecture supports highly parallel and distributed processes of data acquisition, which can be delivered in real time to research, clinical and education centers as terminals for data harvesting and quantitative analysis. Large-scale studies can be conducted, paving the way to more generalizable results than ever in the past. In addition, the possibility to take single-subject measurements on more occasions and in different environments makes finger-tracking evidence usable not only in group studies but also for individual diagnostic purposes. Furthermore, the fine-grained, multimodal evidence of different signal streams which are aligned with time and with linguistically annotated texts provides invaluable training data for artificial neural networks and classification algorithms designed to solve engineering problems or simulate neurophysiological correlates of cognitive tasks. Last but not least, we know that reading probes are a commonly used for monitoring progress in reading fluency and text comprehension (Miura Wayman et al., 2007), but take huge time and effort to collect. The use of a tablet for extended reading enables deriving this information unobtrusively and continuously, wherever the child fancies reading, even at home.

Acknowledgments

This work is supported by PRIN grant 2017W8HFRX *ReadLet: reading to understand. An ICT driven, large-scale investigation of early grade children's reading strategies* (2020-22), from the Italian Ministry of University and Research.

References

- David A Balota and Melvin J Yap. 2006. Attentional control and the flexible lexical processor: Explorations of the magic moment of word recognition. *From inkmarks to ideas: Current issues in lexical processing*, 229.
- Laura Barca, Cristina Burani, and Lisa S Arduino. 2002. Word naming times and psycholinguistic norms for italian nouns. *Behavior Research Methods, Instruments, & Computers*, 34(3):424–434.
- Laura Barca, Cristina Burani, Gloria Di Filippo, and Pierluigi Zoccolotti. 2006. Italian developmental dyslexic and proficient readers: Where are the differences? *Brain and Language*, 98(3):347–351.
- Elizabeth Bates, Cristina Burani, Simona D’Amico, and Laura Barca. 2001. Word reading and picture naming in italian. *Memory & Cognition*, 29(7):986–999.
- Cristina Burani, Stefania Marcolini, and Giacomo Stella. 2002. How early does morpholexical reading develop in readers of a shallow orthography? *Brain and language*, 81(1-3):568–586.
- Cristina Burani, Lisa S Arduino, and Laura Barca. 2007. Frequency, not age of acquisition, affects italian word naming. *European Journal of Cognitive Psychology*, 19(6):828–866.
- Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.
- Robert Davies, Javier Rodríguez-Ferreiro, Paz Suárez, and Fernando Cuetos. 2013. Lexical and sub-lexical effects on accuracy, reaction time and response duration: impaired and typical word and pseudoword reading in a transparent orthography. *Reading and Writing*, 26(5):721–738.
- Maria De Luca, Laura Barca, Cristina Burani, and Pierluigi Zoccolotti. 2008. The effect of word length and other sublexical, lexical, and semantic variables on developmental reading deficits. *Cognitive and Behavioral Neurology*, 21(4):227–235.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Franco Alberto Cardillo, and Vito Pirrelli. 2018a. ReadLet: an ICT platform for the assessment of reading efficiency in early graders. page 61, Edmonton, Alberta (Canada), 25-29 September, 2018. 11th International Conference on the Mental Lexicon.
- Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouaphae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018b. Readlet: Reading for understanding. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 1–6.
- Guillaume Lio, Roberta Fadda, Giuseppe Doneddu, Jean-René Duhamel, and Angela Sirigu. 2019. Digit-tracking as a new tactile interface for visual perception analysis. *Nature Communications*, 10(5392):1–13.
- Chiara Valeria Marinelli, Daniela Traficante, Pierluigi Zoccolotti, and Cristina Burani. 2013. Orthographic neighborhood-size effects on the reading aloud of italian children with and without dyslexia. *Scientific Studies of Reading*, 17(5):333–349.
- Heidi Anne E. Mesmer and Karen Lake. 2010. The role of syllable awareness and syllable-controlled text in the development of finger-point reading. *Reading Psychology*, 31(2):176–201.
- Miya Miura Wayman, Teri Wallace, Hilda Ives Wiley, Renáta Tichá, and Christine A Espin. 2007. Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41(2):85–120.
- Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. The adaptation of the affective norms for english words (anew) for italian. *Behavior research methods*, 46(3):887–903.
- Maria Montefinese, David Vinson, Gabriella Vigliocco, and Ettore Ambrosini. 2019. Italian age of acquisition norms for a large set of words (itaoa). *Frontiers in psychology*, 10:278.
- Vito Pirrelli, Claudia Cappa, Davide Crepaldi, Viola Del Pinto, Marcello Ferro, Sara Giulivi, Claudia Marzi, Andrea Nadalini, and Loukia Taxitari. 2020. Tracking the pace of reading with finger movements. Canada, 16-18 October, 2020. Words in the World International Conference.
- Joanna K. Uhry. 2002. Finger-point reading in kindergarten: The role of phonemic awareness, one-to-one correspondence, and rapid serial naming. *Scientific Studies of Reading*, 6(4):319–342.
- Tal Yarkoni, David Balota, and Melvin Yap. 2008. Moving beyond coltheart’s N: A new measure of orthographic similarity. *Psychonomic bulletin & review*, 15(5):971–979.
- Pierluigi Zoccolotti, Maria De Luca, Gloria Di Filippo, Anna Judica, and Marialuisa Martelli. 2009. Reading development in an orthographically regular language: Effects of length, frequency, lexicality and global processing ability. *Reading and Writing*, 22(9):965–992.

Multiword expressions we live by: a validated usage-based dataset from corpora of written Italian

Francesca Masini¹, M. Silvia Micheli², Andrea Zaninello³, Sara Castagnoli⁴, Malvina Nissim⁵

¹Alma Mater Studiorum – University of Bologna, ²University of Milano-Bicocca

³Zanichelli editore, ⁴University of Macerata, ⁵CLCG, University of Groningen

francesca.masini@unibo.it, maria.micheli@unimib.it

azaninello@zanichelli.it, sara.castagnoli@unimc.it, m.nissim@rug.nl

Abstract

The paper describes the creation of a manually validated dataset of Italian multiword expressions, building on candidates automatically extracted from corpora of written Italian. The main features of the resource, such as POS-pattern and lemma distribution, are also discussed, together with possible applications.

1 Introduction

The computational treatment of multiword expressions (henceforth, MWEs) is notoriously a major challenge in NLP (Ramish, 2015; Villavicencio et al., 2005). In the last decades, the (computational) linguistics community has dedicated many efforts to the development of techniques for the (semi-)automatic identification and extraction of MWEs from corpora and the consequent creation of resources, such as gold standard lists of MWEs, which are needed for evaluation tasks or machine learning training. This notwithstanding, the availability of such resources is still quite limited compared with “the ubiquitous and pervasive nature of MWEs” (Ramish, 2015), especially for ‘non-mainstream’ languages like Italian.

With this work, we contribute to this line of research by providing a dataset of 1,682 validated Italian multiword expressions, obtained through the manual annotation of candidates automatically extracted from corpora of written Italian within the CombiNet project (Simone and Piunno, 2017b). The dataset is to be intended as a first release that will be enriched in the future. We describe our methodology in Section 2, while in Section 3 we

report on preliminary analyses carried out with respect to MWE features and distribution.

2 Methodology

For the creation of the dataset we built on data extracted within the CombiNet project, where the computational task of extracting candidate word combinations from corpora was aimed at supporting the creation of an online lexicographic resource for Italian (Simone and Piunno, 2017a). The notion of ‘word combination’ was large enough to encompass both MWEs (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008; Baldwin and Kim, 2010) – namely strings endowed with (different degrees of) fixedness, idiomaticity or simply conventionality – and more abstract distributional properties of a word, such as argument structures, subcategorization frames or selectional preferences (Lenci et al., 2017).

As a consequence, two different extraction methods – both based on the technique of searching corpora¹ with sets of patterns, and ranking retrieved candidates using frequency and association measures – were used.² More precisely, the search was performed using, in turn, shallow part-of-speech (POS) sequences and syntactic relations: the former method performs better with fixed and adjacent word combinations, whereas the latter is more efficient for syntactically flexible combinations. Since for the present work we focus more on MWEs proper rather than combinatorics in general, we opted to use the data previously gathered with the POS-based method.

Candidates were obtained by feeding the EXTra software (Passaro and Lenci, 2015) with a list of 122 POS-patterns deemed representative of Italian

¹The corpora used within CombiNet were *la Repubblica* (Baroni et al., 2004) and *PAISÀ* (Lyding et al., 2014).

²For a full description of the methods and their assessment, see (Lenci et al., 2017) In what follows we only provide information which is relevant for the current discussion.

MWEs, derived from both relevant literature and a corpus-driven identification task; the list includes adjectival, adverbial, nominal, prepositional and verbal patterns, up to five slots (see Lenci et al., 2017). The results were ranked by LogLikelihood.

As a first step, we selected top-ranked results by cutting at $LL \geq 7,500$, which we observed to be a good balance between precision (high chance of being a MWE) and recall (enough variety), yielding 7,045 candidates. Then we manually annotated this list of candidates to obtain the gold standard inventory of Italian MWEs released and described in the present paper. Each candidate was validated independently by two annotators, and a third annotator judged the conflicted cases,³ which amounted to 673 (less than 10%). We validated sequences that were deemed to display some type of conventionality (fixedness, idiomaticity, high familiarity of use). We included only MWEs in their ‘full’ form (e.g., *punto di partenza* ‘starting point’, *in breve tempo* ‘in a short time’), thus excluding sequences that were clearly part of incomplete MWEs (e.g. *scanso di equivoci*, lit. avoidance of misunderstandings, as part of the larger adverbial MWE *a scanso di equivoci*, lit. at avoidance of misunderstandings, ‘to avoid misunderstandings’).

3 The Resource

The final list of valid MWEs amounts to 1,682 (about 24% of the candidates), and is made available to the community.⁴ The resource contains the following information: (i) lemmatized MWE;⁵ (ii) corresponding POS-pattern;⁶ (iii) corpus/corpora where the MWE was found; (iv) LogLikelihood; (v) raw frequency.

3.1 Caveat

In order to make our resource re-usable on the very same corpora employed for the extraction,

³All annotations were performed by the authors.

⁴DOI: 10.6092/unibo/amsacta/6506.

<http://amsacta.unibo.it/id/eprint/6506>

⁵MWEs are lemmatized because the extraction was performed using lemmas. A consequence of this is that we may have two identical lemmatized sequences that however differ in POS-tagging. For instance, *cambio di guardia* (lit. change of guard) occurs twice: in one case *di* ‘of’ is tagged as a bare preposition, in the other as an articulated preposition (*della* ‘of the’), giving rise to two partially different MWEs (the latter may mean both ‘changing of the guard’ and ‘changeover of leaders’, whereas the former can refer only to the second of these meanings).

⁶The tagset is available here: http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

we kept all data in their original form. This means that lemmatization and POS-tagging were retained, even if erroneous.

Examples of errors and anomalies include:

(a) inconsistent lemmatization, especially for prepositions (e.g. *radere al suolo* ‘raze to the ground’ occurs twice, lemmatized as *radere a suolo* and *radere al suolo*, although the preposition is correctly tagged as an articulated preposition in both cases) and conjunctions (e.g. *carne e ossa* ‘flesh and blood’ and the almost identical *carne ed ossa*, with the euphonic *-d* on the conjunction *e* ‘and’, are two separate items);

(b) wrong lemmatization and tagging, especially for participial-like forms (e.g. *centro abitato* ‘residential area’, lit. center inhabited, lemmatized as *centro abitare*, lit. center to inhabit; or *posta elettronica* ‘electronic mail’ lemmatized as *porre elettronico*, lit. to put electronic, since *posta* is interpreted as the feminine past participle of *porre* ‘to put’ and not as the noun *posta* ‘mail’), but not only (e.g. *lavori di costruzione* ‘construction works’ lemmatized as *lavorio* [instead of *lavoro*] *di costruzione*; or *meccanica quantistica* ‘quantum mechanics’ where *meccanica* is tagged as an adjective);

(c) multiple tagging for the same form (*essere vero* ‘be true’ occurs twice because *vero* is tagged sometimes as an adjective, sometimes as an adverb).

Tricky cases also include lexicalized forms (*guarda caso* ‘strangely enough’, where *guarda* is – correctly, from the technical point of view – lemmatized as *guardare* ‘look’ and tagged as verb, although it is no longer a verb within that lexicalized expression) and pronominal verbs (like *sentirsi in dovere* ‘to feel obliged’, where the verb is lemmatized as *sentire* ‘to feel’, and not as its reflexive form *sentirsi*, although the MWE requires the reflexive form).

3.2 POS-patterns

The validated MWEs in this first release instantiate 82 POS patterns out of the 122 used for the extraction (cf. Section 2). Non-represented patterns (over 30% of the original set) include e.g. Prep-Adj-Verb (e.g. *per quieto vivere* ‘for a quiet life’) as well as more complex – and arguably less frequent – patterns such as N-Prep-ArtDef-N-Adj (e.g. *lotta contro la criminalità organizzata* ‘fight against organized crime’).

Table 1 shows the most attested patterns, while Table 2 the rarely attested ones (only one MWE in our dataset).

Pattern	Fq.	Example
N-Prep-N	165	<i>punto di vista</i> 'viewpoint'
V-ArtDef-N	152	<i>valere la pena</i> 'to be worth'
V-Prep-N	110	<i>scendere in campo</i> 'to take the field'
V-N	83	<i>avere paura</i> 'to be afraid'
V-ArtIndef-N	83	<i>correre un rischio</i> 'to run a risk'
N-A	80	<i>tavola rotonda</i> 'round table'
N-PrepArt-N	79	<i>vigile del fuoco</i> 'fireman'
Prep-N-Prep	77	<i>di fronte a</i> 'in front of'
PrepArt-N-Prep	75	<i>al fine di</i> 'with the aim of'
Prep-N	63	<i>di parte</i> 'biased'
V-Adv	62	<i>andare avanti</i> 'to go on'
N-N	62	<i>piano terra</i> 'ground floor'
V-Adj	55	<i>essere presente</i> 'to be there'
V-PrepArt-N	47	<i>entrare nel merito</i> 'to address'
Prep-ArtDef-N	35	<i>dietro le quinte</i> 'behind the scenes'

Table 1: Most attested POS-patterns

Overall, most attested patterns are 2- or 3-grams. The first 4-slot pattern V-Prep-ArtIndef-N only appears at rank 36, corresponding to 8 different MWEs (e.g. *rispondere a una domanda* 'to answer a question').

In terms of lexical categories, expectedly, most frequent patterns pertain to the nominal and verbal domains. The N-Prep(Art)-N type is the most common pattern for complex nominals, in agreement with theoretical literature (Masini, 2009, e.g.). Patterns headed by prepositions and giving rise to complex prepositions, conjunctions and modifiers are also numerous.

Pattern	Fq.	Example
Prep-Adj-Conj-Adj	1	<i>in bianco e nero</i> 'in black and white'
V-ArtDef-N-A	1	<i>dare il via libera</i> 'to give green light'
A-Prep-V	1	<i>difficile a dirsi</i> 'difficult to say'
V-Prep-Adj-N	1	<i>mettere a dura prova</i> 'to put a strain (on)'
Adj-Prep-N	1	<i>degno di nota</i> 'noteworthy'

Table 2: Least attested POS-patterns

3.3 Lemmas used to form MWEs

The single-word lemmas that concur to form the MWEs in our list amount to 1,235.

Not surprisingly, among the most used lemmas we find function words like prepositions (*di* 'of' fq.421; *in* 'in' fq.227; *al* 'at/to the' fq.124, *a* 'at/to' fq.55 and *ad* 'at/to' fq.10; *per* 'for' fq.50; *da* 'from' fq.34; *su* 'on' fq.24; *con* 'with' fq.20) and determiners (*il* 'the' fq.208; *un* 'a' fq.71 and *una* 'a' fq.41), which appear in many POS-patterns. Conjunctions are instead less frequent (*e* 'and' fq.21 and *ed* 'and' fq.4; *o* 'or' fq.4), like quantifiers (e.g. *ogni* 'each' fq.11).

Quite expectedly, top-ranked verbs (*essere* 'to be' fq.67; *fare* 'to do/make' fq.46; *avere* 'to have' fq.36; *mettere* 'to put' fq.35; *prendere* 'to take' fq.27; *andare* 'to go' fq.19; *dare* 'to give' fq.17) and top-ranked nouns (*tempo* 'time' fq. 32; *mano* 'hand' fq.26; *parte* 'part' fq.23; *posto* 'place' fq.17; *giorno* 'day' fq.16) are lexemes carrying a generic meaning, which favors their combinatory power. Among the mostly used words we also find numerals like *primo* 'first' (fq.30) or *secondo* 'second' (fq.18), and adverbs like *non* 'not' (fq.29).

A cursory comparison between the lemmas of the MWEs in our list and the *Vocabolario di Base* (De Mauro, 1980), which contains the 7,000 most common lemmas in Italian, shows a large convergence: well over 70% of our lemmas are included in the *Vocabolario di Base*. Thus, very frequent MWEs also feature very common lexical items.

3.4 Distribution in corpora

The distribution of MWEs in the two corpora used for the extraction is shown in Table 3.

We retrieved more MWEs from *la Repubblica*

Corpus	N. of MWEs
la Repubblica (total)	1354
PAISÀ (total)	700
la Repubblica (only)	982
PAISÀ (only)	328
Both	372

Table 3: Distribution of MWEs in the two corpora. “Only” indicates how many MWEs are specific to one corpus only and are not found in the other.

than *PAISÀ*, which is expected given that the latter is smaller in size (250M tokens vs. 380M). What is less expected is the rather low number of MWEs shared by the two corpora, amounting to 372, hence 22%. Although *la Repubblica* is a journalistic source and *PAISÀ* is a web corpus containing more varied text genres (especially from Wikimedia Foundation projects), we expected a larger convergence, considering that they both contain written (mid-)formal texts and that *PAISÀ* also contains texts from the news.

Some POS-patterns seem to be definitely more typical of one corpus over the other. As Table 4 illustrates, the N-Prep-N pattern, for instance, is much more typical of *la Repubblica*, whereas the N-Adj pattern is more attested in *PAISÀ*.

Corpus	N-Prep-N	N-Adj
la Repubblica (only)	120	36
PAISÀ (only)	27	44
Both	18	0

Table 4: Distribution of MWEs of two common POS patterns in the two corpora

Among top-ranked MWEs for both LogLikelihood and raw frequency we find *in grado di* ‘able to’ and *per la prima volta* ‘for the first time’, in both corpora. The highest ranked MWEs in *PAISÀ* is *voce correlata* ‘see also’, which is obviously due to the texts that form this resource. Generally, top-ranked MWEs for LogLikelihood also have high frequency, but not in all cases: *essere in essere* ‘to exist’, for instance, turns out to be highly significant in terms of LogLikelihood but has a very low frequency in both corpora.

4 Discussion

The sequences contained in this release are obviously quite heterogeneous.

Semantically speaking, some are very idiomatic in meaning (e.g. *braccio di ferro* ‘arm wrestling’, *colpo di scena* ‘coup de théâtre’, *mandare in onda* ‘to broadcast’), some other (much) less so (e.g. *prendere le distanze* ‘to distance (oneself)’, *andare in pensione* ‘to retire’, *di servizio* ‘service (adj.)’), their specialty lying more in their familiar, conventional status (e.g. *sapere benissimo* ‘to know (damn) well’, *essere favorevole* ‘to be in favour’, *nella storia* ‘in history’). Still others may have more than one meaning, with different degrees of figurativity (e.g. *mettere in scena*, which can mean both ‘to stage’ and ‘to enact’).

From a formal point of view, some look rather fixed and do not admit lexical insertion (e.g. *vero e proprio* ‘proper’) or inflection (e.g. *tra l’altro* ‘by the way’, *ordine del giorno* ‘agenda’), whereas others seem more flexible (e.g. *essere certo* ‘to be sure’, *andare bene* ‘to be OK, to go well’, *posto di lavoro* ‘workplace’). MWE variability is one aspect that we did not address here but definitely deserves to be investigated more thoroughly (cf. e.g. (Nissim and Zaninello, 2011)). In fact, some MWEs may exhibit different behaviour and even completely different meanings according to their grammatical form, like, for example, *a suo tempo* ‘in due course’ (lit. in his/her time) vs. *ai suoi tempi* ‘in his/her time’ (lit. in his/her times). Being based on lemmatized forms, our study does not currently account for such form differences. Moreover, our study is based on contiguous sequences, therefore discontinuous or topicalized occurrences are not accounted for.

We also aim at broadening this initial list by exploring more candidates from the CombiNet data, which are obviously still rich of relevant material. This first release, although limited, is meaningful since it is the first list of commonly used MWEs available for the Italian language, except for domain-specific resources such as PANACEA (Frontini et al., 2012). Although lexicographic material is now accessible for Italian lexical combinatorics (see e.g. (Lo Cascio, 2013)), usage-based and freely available lists of MWEs are still missing and much needed, both for computational tasks and for applied (lexicographic and language teaching related) purposes.

Acknowledgments

This research relies on data extracted within the CombiNet project (PRIN 2010-2011 Word Combinations in Italian, n. 20105B3HE8), coordinated by Raffaele Simone and Alessandro Lenci, and funded by the Italian Ministry of Education, University and Research (MIUR).

References

- Calzolari Nicoletta, Fillmore Charles J., Grishman Ralph, Ide Nancy, Lenci Alessandro, MacLeod Catherine and Zampolli Antonio. 2002. Towards best practice for multiword expressions in computational lexicons. In Rodríguez, M. G. and Araujo, C. P. S. (eds.), *Towards Best Practice for Multiword Expressions in Computational Lexicons*. LREC, 1934-40.
- Baldwin Timothy and Kim Su Nam. 2010. Multiword expressions. In Indurkha, N. and Damerau, F. J. (eds.), *Handbook of natural language processing*, 267-29. Taylor and Francis Group, Boca Raton (FL).
- Baroni Marco, Bernardini Silvia, Comastri Federica, Piccioni Lorenzo, Volpi Alessandra, Aston Guy and Mazzoleni Marco. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)-compliant Corpus of Newspaper Italian. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R. and Silva, R. (eds.), *Proceedings of the Third International Conference on Language Resources and evaluation (LREC)*, 1771-4. European Language Resources Association, Lisbon.
- De Mauro Tullio. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.
- Frontini Francesca, Quochi Valeria, and Rubino Francesco. 2012. Automatic Creation of quality Multi-word Lexica from noisy text data. In Kay, M, Boitet, C. (eds.), *Proceedings of COLING 2012: Sixth Workshop on Analytics for Noisy Unstructured Text Data: 24th International Conference on Computational Linguistics COLING 2012; 2012 December 8-15*. <http://hdl.handle.net/10230/20422>.
- Gries Stefan T. 2008. Phraseology and linguistic theory: A brief survey. In Granger, S. and Meunier, F. (eds.), *Phraseology: An interdisciplinary perspective*, 3-25. John Benjamins, Amsterdam/Philadelphia.
- Lenci Alessandro. 2014. Carving verb classes from corpora. In Simone, R. and Masini, F. (eds.), *Word Classes. Nature, typology and representation*, 17-36. John Benjamins, Amsterdam/Philadelphia.
- Lenci Alessandro, Masini Francesca, Nissim Malvina, Castagnoli Sara, Leboni Gianluca E., Passaro Lucia C. and Senaldi Marco S. G. 2017. How to harvest Word Combinations from corpora: Methods, evaluation and perspectives. *Studi e Saggi linguistici*, 55(2): 45-68.
- Lo Cascio Vincenzo. 2013. *Dizionario combinatorio italiano*. John Benjamins, Amsterdam/Philadelphia.
- Lyding Verena, Stemle Egon, Borghetti Claudia, Brunello Marco, Castagnoli Sara, Dell'Orletta Felice, Dittmann Henrik, Lenci Alessandro and Pirrelli Vito. 2014. The PAISA corpus of Italian web texts. *9th Web as Corpus Workshop (WaC-9)@EACL 2014*, 36-43. EACL (European chapter of the Association for Computational Linguistics).
- Masini Francesca. 2009. Phrasal lexemes, compounds and phrases. *Word Structure*, 2(2): 254-71.
- Nissim Malvina and Zaninello Andrea. 2011. A quantitative study on the morphology of Italian multiword expressions. *Lingue e linguaggio*, 10(2): 283-300.
- Passaro Lucia C. and Lenci Alessandro. 2015. Extracting Terms with EXTra. In Corpas Pastor, G. (ed.), *Computerised and Corpus-based Approaches to Phraseology. Monolingual and Multilingual Perspective*, 188-196. Editions Tradulex, Geneva.
- Ramisch Carlos. 2015. *Multiword Expressions Acquisition - A Generic and Open Framework*. Springer, Dordrecht.
- Sag Ivan A., Baldwin Timothy, Bond Francis, Copestake Ann and Flickinger Dan. 2002. Multiword expressions: A pain in the neck for NLP. *International conference on intelligent text processing and computational linguistics*, 1-15. Springer, Dordrecht.
- Simone Raffaele and Piuanno Valentina. 2017a. Entry word combination: lexicographical representation and lexicological aspects. *Studi e Saggi Linguistici*, 55(2): 13-44.
- Simone Raffaele and Piuanno Valentina, editors. 2017b. Word Combinations: phenomena, methods of extraction, tools, Special Issue of *Studi e Saggi Linguistici*, 55(2).
- Villavicencio Aline, Bond Francis, Korhonen Anna, McCarthy Diana. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech and Language*, 19(4): 365-377.

The Style of a Successful Story: a Computational Study on the Fanfiction Genre

Andrea Mattei*, Dominique Brunato[◊], Felice Dell’Orletta[◊]

• University of Pisa

a.mattei3@studenti.unipi.it

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC–CNR)

ItaliaNLP Lab - www.italianlp.it

{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

This paper presents a new corpus for the Italian language representative of the *fanfiction* genre. It comprises about 55k user-generated stories inspired to the original fantasy saga “Harry Potter” and published on a popular website. The corpus is large enough to support data-driven investigations in many directions, from more traditional studies on language variation aimed at characterizing this genre with respect to more traditional ones, to emerging topics in computational social science such as the identification of factors involved in the success of a story. The latter is the focus of the presented case-study, in which a wide set of multi-level linguistic features has been automatically extracted from a subset of the corpus and analysed in order to detect the ones which significantly discriminate successful from unsuccessful stories

1 Introduction

Computational Sociolinguistics is an emergent interdisciplinary field aimed at exploiting computational approaches to study the relationship between language and society (Nguyen et al., 2016). One of the primary factors driving its foundation is the widespread diffusion of social media and other user-generated data available online, which has promoted massive research on computer-mediated communication from several perspectives. For instance, scholars working in the field of genre and register variation have relied on quantitative approaches to inspect the peculiarities of social media language, with the purpose of providing

a characterization of this new genre with respect to more traditional ones (Paolillo, 2001; Herring and Androutsopoulos, 2015). In the NLP community, the writing style of user-generated data has been analyzed through computational stylometry approaches for addressing tasks broadly related to author profiling (Daelemans, 2013), such as gender and age detection (Peersman et al., 2011; Koppel et al., 2002). The vast majority of this work has taken into account contents published on few microblogging platforms considered as more representative of the contemporary user-generated mediascape, e.g. Twitter. More recently, the attention has been oriented to the language used by online communities whose members share a common interest towards an object, an activity – and more in general any area of human interest – allowing scholars to shed light on the growing phenomenon of fandom (Sindoni, 2015). One of the most prominent expressions of fandom is *fanfiction* (fanfic, fic or FF), i.e. fiction written by fans of a TV series, movie, book etc., using existing characters and situations to develop new plots. In many languages dedicated websites exist where users can publish their own literary works inspired to the original book they are fans of.

From a computational linguistics standpoint, one perspective from which fanfiction has been investigated aimed to infer the relationship between user-generated stories and their original source, e.g. comparing the representation of characters according to their gender, as well as to model reader reactions to stories (Smitha and Bamman, 2016). Inspired to that study, which was based on a large dataset of stories mainly in English, we collect a new corpus of fanfic stories¹, which, to our knowledge, is the first one for the Italian language. We rely on this corpus to carry out an investigation

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Terms of service forbid us to distribute this data. However, the tools used to gather it are available at <https://github.com/AndreMatte97/Fanfiction>

aimed at shedding light on the possibility of computationally modeling the expected success of a fanfic story, based on the assumptions of linguistic profiling and stylometry research.

2 Dataset collection

The corpus comprises texts collected from *efpfanfic.net*, a portal active since 2001 which allows users to publish stories and to comment on them. The website is made up of two sections: one for original stories and the other for fanfictions. We considered only the latter and we limited the collection to stories based on the fantasy saga by the British writer J.K. Rowling, “Harry Potter”. This choice was motivated by the main purpose of our analysis, i.e. characterizing the success of a novel with respect to its writing style rather than as an effect of the various subject matters it deals with. At the same time, the preference given to a very popular book allowed us to keep a consistent number of potential readers and reviewers across the corpus, still having a large sample of texts to analyze. The data collection was performed through web scraping, with two spiders written in Python using the open-source Scrapy framework. The first spider crawls the list of stories in the category of choice and extracts their first chapters together with some metadata, including the URLs of the subsequent chapters. The second spider takes these addresses as input and downloads texts and additional information about all the chapters after the firsts. In the dataset created this way, the record for each chapter includes: *ID* and *Reference ID*, combinations used by the website to identify the webpage of each chapter. We use the ID of the first chapter as a reference to group together records belonging to the same story; *Title*; *Rating*, an estimate given by the author about the rawness of themes and scenes contained in his story; *Date* of posting; Author’s *nickname*; *Number of chapters* in the story; *Text*; *Total number of reviews received by the story*, divided in positive, negative and neutral; *Number of reviews received by the single chapter*, as well as the text of the most recent ones. The crawlers downloaded 54,717 stories, for a total of 19,7310 chapters and a mean of approximately 3.6 chapter per story, which is consistent with the one calculated taking into account every entry on the website. The obtained corpus was divided into folders, each containing stories with the same number of chapters.

3 The success of a fanfiction story: an exploratory study

Based on the newly created dataset, we carried out a computational stylometric analysis aimed at studying whether there is a connection between the success of a fanfic story and its writing style. Such a connection has been demonstrated for more canonical literary works covering novel and movie domains (Ganjigunte et al., 2013; Solorio et al., 2017), showing that stylometry is a viable approach also in scenarios different from authorship attribution and verification.

The methodological framework of our investigation is *linguistic profiling* (Montemagni, 2013; van Halteren, 2004), a NLP-based approach in which a large set of linguistically-motivated features automatically extracted from text are used to obtain a vector-based representation of it. Such representations can be then compared across texts representative of different textual genres and varieties to identify the peculiarities of each. For the purpose of our analysis, we split the original dataset into two varieties corresponding to “successful” and “unsuccessful” stories. To define success we follow an approach similar to that used by Solorio et al. (2017), which is based on the number of reviews obtained by each story. In this regard, we decided to include all reviews, not only the positive ones, which can undoubtedly testify a favorable attitude by the reader for the story. Two main reasons motivated our choice: first, we noticed that the overwhelming majority of collected reviews are written to convey appreciation, with just 0.73% among a total of nearly 900k reviews being negative; therefore, from a statistical point of view, we can reasonably get rid of the distinction between various kinds of reviews and simply take into consideration the overall amount of feedback received. Secondly, also a negative feedback proves that a given story has been read and aroused some interest in the reader. With this in mind, we define as “unsuccessful” those stories that did not receive any reviews, thus being largely ignored by their readers. Conversely, the “successful” category includes all stories with the same number of chapters having received a review count higher than the average of all stories of that length. We also decided to limit the focus of this analysis to single-chapter fanfictions written before 2018, so as to avoid the inclusion of stories not yet concluded. The resulting classes comprise 2101 un-

successful texts and 14486 successful ones, with a threshold for success amounting to 5 reviews. Table 1 shows an example of stories classified in the two categories.

All texts were pre-processed by means of regular expressions, with the aim of removing errors and inconsistencies in the use of punctuation, capitalization and special characters, in order to increase the reliability of automatic linguistic annotation and the process of feature extraction, which were performed using the Profiling-UD tool (Brunato et al., 2020).

In what follows we first provide an overview of the linguistic features used for our statistical analysis and then we discuss the ones that turned out to be more prominent in successful writing.

3.1 Linguistic Features

The set of features is based on the one described in Brunato et al. (2020) and counts more than 150 features, distributed across distinct levels of linguistic annotation and computed according to the Universal Dependencies (UD) annotation framework. These features have been shown to be effective in a variety of different scenarios, all related to modeling the ‘form’ of a text, rather than the content: e.g., from the assessment of sentence complexity by humans (Brunato et al., 2018) to the identification of the native language of a speaker from his/her productions in a second language (L2) (Cimino et al., 2018). Specifically, they can be grouped into the following main phenomena:

Raw Text Features: Document length computed as the total number of tokens and of sentences (*#Tokens*, *#Sentences* in Table 2); average sentence length and token length, calculated in tokens and in characters, respectively (*Sent length*, *Word length*).

Lexical Richness: Distribution of words and lemmas belonging to the Basic Italian Vocabulary (De Mauro, 2000) (*BIV_Tok*, *BIV_Types*) and to the internal repertoires (i.e. fundamental, high usage and high availability, *BIV_Fund*; *BIV_High-US*; *BIV_High-AV*); Type/Token Ratio, a feature of lexical variety computed as the ratio between the number of lexical types and the number of tokens in the first 100 and 200 words of text (*TTR Lemma*); Lexical density.

Morpho-Syntactic Information: Distribution of all grammatical categories, with respect to the Universal part-of-speech tagset (*UPOS_** and the

language specific tagset (*XPOS_**); Distribution of verbs according to tense, mood and person, both for main and auxiliary verbs (*aux_**; *V_**)).

Verbal Predicate Structure: Average distribution of verbal roots and of verbal heads for sentences (*VerbHead*); features related to the arity of verbs (i.e. average number of dependents for verbal head, distribution of verbs by arity).

Global and Local Parsed Tree: Average depth of the syntactic tree (*MaxDepth*); average depth of embedded complement chains headed by a preposition; average length of dependency links and of the maximum link (*Links Len*; *Max Link Length*); relative order of the subject and object with respect to the verb;

Syntactic relations: Distribution of typed UD dependency relations (*dep_**);

Use of Subordination: Distribution of main and subordinate clauses (*Main clause*, *Subord clause*), average length of subordinate chains, distribution of subordinate chains by length.

4 Data Analysis

For each considered feature we calculated the average value and the standard deviation in the two classes. We then assessed whether the variation between mean values is significant using the Wilcoxon rank sum test. We found that 57% (i.e. 126 out of the 219) of features are differently distributed in a significant way between successful and unsuccessful stories. In Table 2 we report an extract of the most interesting ones.

As it can be seen, successful stories are on average longer in terms of number of tokens and sentences (1, 2), although these sentences are generally shorter (3), suggesting that readers appreciate more a plain writing style. However, when lexical factors are considered, the preference is given to texts exhibiting less frequent words, as suggested by the slightly lower distribution of words belonging to the Basic Italian Vocabulary (5,6) and especially to the Fundamental one (7). Inflectional morphology also appears as a domain of variation between the two classes. Successful fanfictions employ quite more often verbs in the second person (15), a feature typical of narrative writing related to direct speech. On the contrary, we observe a higher distribution of third person verb, specifically auxiliaries, both singular (14) and plural (13), in less successful texts, which can hint at a preference for reported speech.

Label	Example (<i>Italian</i>)	Example (<i>English</i>)
Successful	La città di Edimburgo era sommersa da una cascata d'acqua. Pioveva. Pioveva da giorni e giorni, senza sosta. Il cielo era illuminato di lampi e scosso da tuoni. Le strade erano vuote. Per la prima volta da giorni, allo scoccare della mezzanotte, la pioggia cessò di colpo. Il silenzio piombò sui quartieri che sembrarono improvvisamente più bui. E in quel silenzio penetrante, l'unico rumore che si riusciva a distinguere era un tac-tac-tac leggero e discontinuo. Proveniva da una finestra. La finestra di una lussuosa casa in centro, l'unica luce accesa a quell'ora. Joanne era davanti al computer, fonte di quel tremolio e scriveva. Batteva le dita sulla tastiera per alcuni istanti, poi si fermava, rileggeva, cancellava e riscriveva. Andava avanti così da giorni. I suoi occhi erano stanchi, ma la sua mente lavorava frenetica. Mancava poco ² .	The city of Edimburgh was flooded by a cascade of water. It was raining. It had been raining for days and days, relentlessly. The sky was lit by lightning and shaken by thunder. The streets were empty. For the first time in days, at the stroke of midnight, the rain stopped abruptly. Silence fell upon the districts that suddenly seemed darker. And in that piercing silence, the only noise that could be recognized was a faint and irregular tac-tac-tac. It was coming from a window. The window of a luxurious house in the city centre, the only light still on at that time. Joanne was in front of the computer, source of that trembling and was writing. She tapped her fingers on the keyboard for a few moments, then stopped, reread, deleted and rewrote. She had been going on like this for days. Her eyes were tired, but her mind was working frantically. Almost there.
Unsuccessful	Il cielo era tetro cosperso di nuvole che sembravano volere annunciare un acquazzone, il vento ulula forte facendo sbattere le finestre violentemente, come se volesse gridare, liberarsi da una rabbia repressa. La donna dai lunghi capelli rosso scuro continuava a fissare la devastazione attraverso il vetro che ora si era appannato dal suo stesso respiro. Aveva lo sguardo malinconico non più illuminato da quella dolce espressione che il riso le donava. Una mano le si poggiò sulla spalla e girò pian piano il volto verso la persona amata che con un ritmo lento cominciò ad accarezzarle le gote che assunsero un colorito roseo alla sua pelle pallida. Chiuse gli occhi come per assaporare quel dolce tocco che ora si era spostato nei suoi capelli. "Non guardare più oltre il vetro" Mormorò la voce con una nota di preoccupazione, apparteneva a James, marito di Lily la donna dai lunghi capelli rossi ³ .	The sky was bleak strewn with clouds that seemed to want to announce a downpour, the wind howls loudly making the windows slam violently, as if it wanted to scream, to free itself from a suppressed anger. The woman with the long dark red hair kept staring the devastation through the glass that was now clouded by her own breath. Her melancholic gaze was no longer lit up by that sweet look that laughter gave her. A hand rested on her shoulder and slowly turned her face towards the loved one who started slowly caressing her cheeks which took on a rosy tone on her pale skin. She closed her eyes, as if to savor that sweet touch that had now moved into her hair. "Don't look beyond the glass anymore" Whispered the voice with a note of concern, it belonged to James, husband of Lily the woman with long red hair.

Table 1: An extract of a 'successful' story (the most reviewed one) and of an 'unsuccessful' one.

Focusing on the distribution of morpho-syntactic categories, there is a significant difference in the usage of the most common punctuation marks, commas (25) and full stops (26), which are quite more frequent in highly-reviewed fan-fictions. These features relate themselves to the previously observed difference in terms of document length, as texts with more sentences necessarily use punctuation marks to divide them. Ad-

ditionally we can see that balanced marks (24), i.e. parenthesis and quotation marks, occur more in successful texts, strengthening our previous claim about a more frequent presence of direct speech in this class. At syntactic level, dependency relations are slightly shorter in successful texts, both considering the average value of all dependencies (29) and the value of the maximum dependency link (30). In readability assessment studies, longer syntactic dependencies are typically found in complex texts, and the same holds for deeper syntactic trees. Both these features have lower values in highly-reviewed stories, suggest-

²The full story can be found at <https://efpfanfic.net/viewstory.php?sid=607026&i=1>

³The full story can be found at <https://efpfanfic.net/viewstory.php?sid=27412&i=1>

Feature	Unsucc		Success	
	Avg	(StDev)	Avg	(StDev)
Raw Text Features				
1. # Tokens	1401	(1940)	2120	(2718)
2. # Sentences	78.4	(116.7)	125.1	(153.6)
3. Sent length	20.18	(12.39)	17.38	(6.43)
4. Word length	4.50	(.250)	4.52	(.193)
Lexical Features				
5. % BIV_Tok	85.7	(5.1)	84.8	(4.7)
6. % BIV_Types	73.4	(7)	70.1	(7)
7. % BIV_Fund	61	(7.5)	57.1	(7.7)
8. * % BIV_High-AV	3.1	(1)	3.1	(1)
9. % BIV_High-US	8.5	(2.4)	9.1	(2.5)
10. Lexical density	.498	(.033)	.503	(.031)
11. TTR Lemma 100	.560	(.118)	.560	(.112)
12. TTR Lemma 200	.433	(.114)	.436	(.110)
Morpho-Syntactic Features				
13. % Aux_3perPl	13.2	(9)	11.8	(7.6)
14. % Aux_3perSin	54.4	(17.1)	53.2	(15.6)
15. % Aux_2perSin	6.3	(8.1)	7.9	(8.5)
16. % Aux_Imperf.	38.5	(24.8)	31.3	(23.6)
17. % Aux_Pres.	52.4	(26)	60	(23.6)
18. % V_Gerund	5.7	(3.8)	6.3	(4)
19. % upos_VERB	12.5	(1.8)	12.3	(1.7)
20. % upos_NOUN	13.8	(2.3)	13.5	(2.1)
21. % upos_ADJ	4.7	(1.4)	4.6	(1.2)
22. % upos_PRON	8.59	(2.24)	8.51	(2)
23. % upos_ADP	10.8	(1.9)	10.4	(1.8)
24. % xpos_FB	1.7	(2)	2.3	(2.4)
25. % xpos_FF	6.5	(2.7)	7.1	(2.8)
26. % xpos_FS	5.5	(2.2)	6.1	(2.1)
27. % xpos_CC	3.1	(.9)	2.9	(.8)
28. % xpos_CS	1.7	(.7)	1.8	(.7)
Syntactic Features				
29. Links Len	2.78	(.438)	2.72	(.385)
30. *Max Link Len	1.19	(2.38)	.687	(1.33)
30. MaxDepth	3.96	(1.45)	3.58	(.857)
32. % Main clause	48.8	(9.9)	49.9	(9)
33. % Subord clause	51.2	(9.9)	50.1	(9)
34. % Verb Head	2.63	(1.72)	2.26	(.897)
35. % dep_nsubj	4.9	(1.1)	4.7	(1)
36. * % dep_obj	5.3	(1.1)	5.3	(1)
37. % dep_obl	5.5	(1.1)	5.2	(1)
38. % dep_punct	14.2	(3.8)	16	(4.1)
39. % dep_conj	4	(1.3)	3.7	(1.1)
40. % dep_det	10.9	(2)	10.5	(1.8)

Table 2: An extract of linguistic features varying significantly between *successful* and *unsuccessful* stories. All differences are significant at $p < 0.001$, except for features marked with an asterisk, which have $p < 0.05$.

ing that the style of successful writing is characterized by a simpler syntactic structure. Interestingly, these results, although preliminary, go in the opposite direction to those reported by Ganjigunte et al. (2013) for successful literary works in English, which were found to be less correlated with text readability scores. Finally, subordinate clauses (33) occur slightly more often than main

clauses (32) in unsuccessful texts, while there is a nearly even split between hypotaxis and parataxis in successful ones.

To deepen our analysis, we also computed the coefficient of variation σ^* for all features varying significantly between the two classes, where σ^* is the ratio between the standard deviation σ and the mean μ . This allowed us to evaluate the dispersion of values around the average in a standardized way, and thus to compare the stability of features pertaining to data measured on different scales. A feature that is much scattered in a class of texts and highly stable in the other has a greater chance of being a meaningful representative of the latter.

In Figure 1 we show the average variability in the two classes of the four groups of features distinguished according to the level of annotation they were extracted from. As a whole, we noticed that successful texts display less variability in nearly every considered feature: 117 of them (92%) are more stable in this class. In successful stories, features with greater stability compared to the other class are mainly raw text, e.g. number of sentences, number of tokens and syntactic ones, e.g. verbal heads per sentence and average depth of syntactic trees. Among the few features which are more stable in poorly received texts, we find instead verbal predicate features, such as the distributions of past tenses and of indicative moods, in addition to the frequency of usage of cardinal numbers. The set of lexical features is instead the most stable one for both classes.

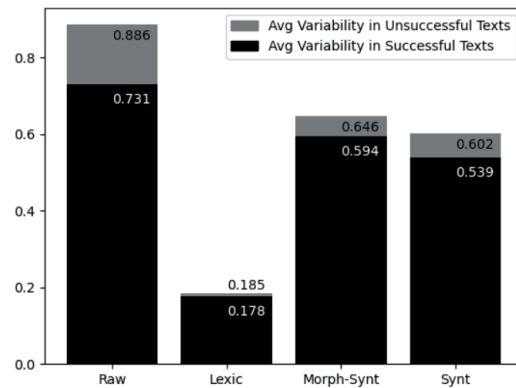


Figure 1: Average coefficient of variation in each class of features, both for successful and unsuccessful texts.

5 Conclusion

In this paper, we presented a NLP-based stylistic analysis on the emerging genre of fanfiction aimed at characterizing the writing style of a successful story. We collected a new large-scale corpus which – to the best of our knowledge – is the first one of this genre for Italian. We showed that successful stories, defined as those receiving a number of reviews higher than the average, are characterized by a variety of linguistic features at different levels of granularity and that these features are more uniformly distributed within them.

In the future, we would like to broaden the perspective to other genres in order to study whether there are linguistic predictors of successful writing which are constant across different genres, as well as across concepts somehow similar to success, such as virality and engagement.

References

- D. Brunato, A. Cimino, F. Dell’Orletta, G. Venturi and S. Montemagni. 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts. *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, 7145–7151.
- D. Brunato, L. De Mattei, F. Dell’Orletta, B. Iavarone and G. Venturi. 2018. Is this Sentence Difficult? Do you Agree? *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, 2018.
- A. Cimino, F. Dell’Orletta, D. Brunato and G. Venturi. 2018. Sentences and Documents in Native Language Identification. *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-IT)*, 1–6, Turin.
- W. Daelemans. 2013. Explanation in Computational Stylistics. Gelbukh A. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2013*, Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg.
- Tullio De Mauro. 2000. *Grande dizionario italiano dell’uso* (GRADIT). Torino, UTET.
- V. Ganjigunte Ashok, S. Feng and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1753–1764.
- S.C. Herring and J. Androutsopoulos. 2015. Computer-mediated discourse 2.0. *The handbook of discourse*, 2nd ed. Deborah Tannen, Heidi E. Hamilton, Deborah Schiffrin, eds. John Wiley Sons., 1753–1764.
- M. Koppel, S. Argamon and A. Rachel Shmuni. 2002. Automatically Categorizing Written Texts by Author Gender. *Lit. Linguistic Comput.*, 17, 4, 401–412.
- S. Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, 145–172.
- D. Nguyen, A.S. Doğruöz, C.P. Rosé, and F.M.G. de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, Vol. 42, No. 3, 537–593.
- John Paolillo. 2001. Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics*, 5, 180–213.
- C. Peersman, W. Daelemans, and L. Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, 37–44.
- M.G. Sindoni. 2011. ‘I Really Have No Idea What Non-Fandom People Do with Their Lives.’ A Multimodal and Corpus-Based Analysis of Fanfiction. *Lingue e Linguaggi*, (13), 2015, 277–300, doi.org/10.1285/i22390359v13p277.
- M. Smitha and D. Bamman. 2016. Beyond Canonical Texts: A Computational Analysis of Fanfiction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, Austin, Texas, USA, November 1–4, 2016.
- T. Solorio, M. Montes-y-Gómez, Suraj Maharjan, J. Ovalle and Fabio A. González. 2017. Multi-task Approach to Predict Likability of Books. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1217–1227.
- H. van Halteren. 2004. Linguistic profiling for author recognition and verification. *Proceedings of the Association for Computational Linguistics*, 200–207.

A Multimodal Dataset of Images and Text to Study Abusive Language

Stefano Menini
Fondazione Bruno Kessler
Trento, Italy
menini@fbk.eu

Alessio Palmero Aprosio
Fondazione Bruno Kessler
Trento, Italy
aprosio@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

English. In this paper, we present a novel dataset composed of images and comments in Italian, created with teenagers in classes using a simulated scenario to raise awareness on cyberbullying phenomena. Potentially offensive comments have been collected for more than 1,000 images and manually assigned to a semantic category. Our analysis shows that the presence of human subjects, as well as the gender of the people present in the pictures trigger different types of comment, and provides novel insight into the connection between images posted on social media and offensive messages. We also compare our corpus with a similar one obtained with WhatsApp, showing that comments to images show different characteristics compared to text-only interactions.¹

1 Introduction

In order to study abusive language online, the availability of datasets containing the linguistic phenomena of interest are of crucial importance. However, when it comes to specific target groups, for example teenagers, collecting such data may be problematic due to issues with consent and privacy restrictions. Furthermore, while text-only datasets for abusive language detection have been widely developed and used by the NLP community, limitations set by image-based social media platforms like Instagram make it difficult for researchers to experiment with multimodal data. We therefore present a novel corpus containing images and potentially offensive Italian comments and we analyse it from different perspectives, to investi-

gate whether the subject of the images plays a role in triggering a comment.

The data collection was carried out in several school classes, being part of a ‘living lab’ to raise awareness on cyberbullying and, more generally, on the use of social media by teenagers. The dataset is freely available on Github² and, since the comments were collected with the written consent of parents and teachers, they can be freely used for research purposes, without the ethical implications that would derive from using real data posted by teenage users. The images, instead, are released as a ResNet-18 neural network trained on ImageNet, similar to recent NLP works (Kruk et al., 2019), since they were taken from Instagram and cannot be shared as pictures.

2 Related Work

Several datasets have been created to study hate speech, abusive language and cyberbullying. Most of them include single textual comments or threads annotated as being hateful/offensive/abusive or not. For example Reynolds et al. (2011) propose a dataset of questions and answers from Formspring.me, a website with a high amount of cyberbullying content. It consists of 12,851 posts annotated for the presence of cyberbullying and severity. Another resource developed by Bayzick et al. (2011) consists of conversation transcripts (thread-style) extracted from MySpace.com, which are annotated for presence and typology of cyberbullying. For an overview on existing annotation schemes and datasets specific to cyberbullying see the survey presented in (Emmery et al., 2019). Similarly, a project called Hate Speech Datasets³ (Vidgen and Derczynski, 2020) collects a comprehensive list

¹”Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).”

²<https://github.com/dhfbk/creep-image-dataset>

³<https://github.com/leondz/hatespeechdata>

of datasets that are annotated with offensive language, online abuse, and so on.

Probably the most popular datasets shared within the NLP community have been extracted from Twitter because of its relatively easy-to-use APIs. Indeed, most of the shared tasks recently organised to build and evaluate hate speech detection systems use Twitter data (Basile et al., 2019; Struß et al., 2019; Bosco et al., 2018; Aragón et al., 2019).

The relationship between textual content and images and the role that they play together is a relatively understudied problem in relation to online hate speech. A notable exception is the dataset collected from Instagram by Hosseinmardi et al. (2015), which consists of 2,218 media sessions, each being annotated with information on cyber-aggressive behavior. In this dataset the annotation refers to a thread and not to single offensive messages. The corpus has been used also for classification tasks, for example in (Cheng et al., 2019) it was employed to detect cyberbullying through a hierarchical attention network that takes into account the hierarchical structure of social media sessions and the temporal dynamics of cyberbullying. Other Instagram datasets have been created but they cannot be shared due to the restrictions in the social network policy (Yang et al., 2019).

3 Annotation Tool and Process

The annotation was performed involving overall 95 students aged between 15 and 18. The activity was carried out in classes, during ‘living labs’ aimed at raising teenagers’ awareness on online harassment and cyberbullying. Students were given access to the CREENDER tool⁴, a web-based annotation system that displays pictures taken from a pre-defined batch of images, and allows users to add comments (Palmero Aprosio et al., 2021). In this case, the images were first extracted from Instagram by the authors of this paper and then manually checked to avoid nudity and explicit sexual content.

After a user logs in the system, a picture is displayed, and a prompt asks “*If you saw this picture on Instagram, would you make fun of the user who posted it?*”. If the user selects “No”, then the system picks another image randomly and the same question is asked. If the user clicks on “Yes”, a

second screen opens where the user is asked to specify the reason why the image would trigger such reaction by selecting one of the following categories: “*Body*”, “*Clothing*”, “*Pose*”, “*Facial expression*”, “*Location*”, “*Activity*” and “*Other*”. The user should also write the textual comment s/he would post below the picture. After that, the next picture is displayed, and so on.

The question posed by the system does not ask explicitly whether the user would *insult*, *harrass* or *offend* the person who posted the image, because in a preliminary test with students we observed that the answer would almost always be “No”. This showed that only in few cases a user would consciously harm another user, especially if the two know each other. Furthermore, comments with explicit hateful content are easy to find online and can be unambiguously annotated in most of the cases. We therefore decided to focus on a more nuanced form of offensive message, that is when a user makes fun of another one. We made this choice because we assumed that this kind of messages would be more ambiguous, containing ironic or sarcastic comments, and mixing humorous and abusive content without being necessarily explicit. This would make the collected data very interesting from a linguistic and computational point of view.

The data collection was embedded in a larger process that required two to three meetings with each class, one per week, involving every time two social scientists, two computational linguists and at least two teachers. During these meetings several activities were carried out with students, including simulating a WhatsApp conversation around a given plot as described in (Sprugnoli et al., 2018), commenting on existing social media posts, and annotating images as described in this paper. Since ethical issues were a main concern since the drafting of the study design, because all participants were underage students, all the activities had been co-designed with the schools involved and informed consent was gathered beforehand both from teachers and from parents.

The sessions with students were organised so that different school classes annotated the same set of images, in order to collect multiple annotations on the same pictures. However, since some users were quicker than others in giving a judgement on the pictures, we could not collect multiple annotations for all images included in the dataset (see

⁴<https://github.com/dhfbk/creender/tree/master>

Pictures with ↓ ... and having →	At least 1 comment	(Total comments)	No comments
At least 1 judgement	1,018	1,135	16,894
At least 2 judgements	901	1,018	9,876
At least 3 judgements	713	815	5,454
At least 4 judgements	495	563	3,060

Table 1: Number of pictures in the dataset with at least n judgments (‘yes’/‘no’) and number of comments.

Table 1 for details).

4 Annotated Corpus

Overall, 17,912 images have been judged at least once by the students. For 1,018 of them, at least an offensive comment has been written during the annotation sessions. Overall, the number of comments in the dataset is 1,135, which is higher than the number of pictures with a comment because the same image may be commented more than once by different students. An overview of the content of the dataset including images and comments is presented in Table 1. Note that the number of *judgements* refers to the ‘yes/no’ option selected by users in the first platform view, while the number of *comments* refers only to the images tagged with a ‘yes’, for which a student wrote also a comment. Overall, only one image has been tagged with four ‘yes’, and in most of the cases annotators selected only ‘no’. The number of images tagged with exactly three ‘yes’ is 13, those with two is 88. Since these images have been leveraged from Instagram with no particular criterion in mind, the distribution of ‘yes’ and ‘no’ may be considered realistic, with the majority of pictures not triggering any potentially negative reaction, and around 6% of them being associated with offensive comments.

In general, we observe that there is a low agreement on whether a picture triggers an offensive comment or not. This suggests that an offensive intent is more dependent on the attitude of a user posting a comment than on image-specific features. We also compute inter-annotator agreement – using Krippendorff’s alpha measure (Krippendorff, 1970) – on the trigger categories assigned to the comments, considering only the images that received at least two comments. Agreement is 0.19, which implies again that the reason to make fun of a user does not depend on a specific feature of the picture, but rather that multiple aspects of a posted image can be taken as an excuse for

	Female	Male	Both	None	Total
Body	27	20	3	4	54
Clothing	66	30	9	12	127
Pose	114	99	11	5	229
Facial Exp.	68	90	17	7	182
Location	16	17	7	57	97
Activity	12	14	7	36	69
Other	72	63	22	113	272
Total	377	318	76	252	1023

Table 2: Distribution of offence triggers per subject types

potentially offensive comments. In order to avoid the ambiguity introduced by the ‘Other’ label, we also compute IAA ignoring this class. This time the agreement value is 0.26, showing that on the one hand the ‘Other’ label covers uncertain cases, but also that the reason to comment a picture remains highly subjective.

Some typical comments collected during the simulation are *Coprìti (Cover yourself up)*, *Che schifo di foto (This picture sucks)* and *Inquietante (Disturbing)*. These comments have different features compared to hate speech messages extracted from Twitter: they tend to be short because they complement the image and they are rich in deictic expressions. In most of the cases, they are not self-contained from a semantic point of view.

5 Corpus Analysis

In order to analyse whether what is portrayed in a picture has an impact on the choice to write an offensive message, we manually assign each image with at least 1 comment to one of the following categories: male-only subject(s), female-only subject(s), mixed group, no human subject. How the different categories are distributed, taking into account also the triggers (i.e. self-declared reasons to write a comment) is displayed in Table 2.

The last column of the table shows that the categories are rather imbalanced, with a minority of comments associated with the ‘Body’ label and several comments concerning the ‘Pose’. How-

ever, for most of the comments the ‘Other’ label was used. When collecting feedback from students after the annotation sessions, several annotators suggested that it should be possible to assign multiple labels instead of just one, and reported that they used the ‘Other’ label for those cases. On the other hand, they did not express the need to include additional categories in the annotation.

As regards the picture subjects, the analysis shows that the main differences between pictures with male and female subjects concern the ‘Facial expression’ and ‘Clothing’ categories: the first is more frequently associated with male subjects, while the second seems to be more related to female subjects. When there are multiple subjects with different genders, instead, no particular differences are observed. As expected, when no person is portrayed in the picture, ‘Location’, ‘Activity’ and ‘Other’ are prevalent. In some cases, ‘Pose’ or ‘Expression’ are selected, because of the presence of animals or drawings in the images.

We manually assign the subject category also to a set of 3,200 pictures randomly taken from the images that were tagged with ‘No’. Then we compare the two category distributions, that are reported in Table 3. By applying the χ^2 test ($N = 4, 218$), we observe a statistically significant difference between the two distributions of categorical variables ($p < .001$). In particular, pictures with no human subject are less likely to get an offensive comment, while those with a female subject are the most commented ones. Also male subjects, however, trigger offensive comments very frequently, while they are only present in 19% of the images which were not commented by users.

	% Yes	% No
Females	36.85	32.14
Males	31.09	19.00
Mixed	7.43	9.33
Nobody	24.63	39.53

Table 3: Subject types for pictures annotated with ‘Yes’ (i.e. triggering a comment) and ‘No’

6 Dataset comparison

In order to better understand the peculiarities of the textual comments in our corpus, we compare them with the messages in another existing corpus created with a similar approach, i.e. simulated scenarios, and with the same goal, i.e. study how

teenagers communicate online. More specifically, the second corpus was created following the approach described in (Sprugnoli et al., 2018) using WhatsApp chats in classes to simulate cyberbullying interactions among teenagers. The target age group is the same as for our multimodal corpus, but in the second corpus the interactions are solely based on text. We select from the WhatsApp corpus the 3,004 comments manually tagged as offensive so to make them comparable with the 1,135 comments in our multimodal corpus, which were all written with the goal to make fun of someone. Both corpora are processed with the TINT suite (Aprosio and Moretti, 2018), through which a number of linguistic features were extracted.

As regards type/token ratio, it is 0.62 in the WhatsApp corpus and 0.82 in our data, suggesting that images may foster a richer, more creative use of the language, even if offensive. This difference may also be affected by the fact that WhatsApp chats followed a pre-defined plot, therefore limiting the topics to be mentioned in the interactions. Also lexical density is different, being 0.56 on WhatsApp and 0.65 in our corpus. This confirms that the language used in image comments is more complex and more similar to written standard language, while WhatsApp chats share some features of spoken interactions, where content is generally sparser (Stubbs, 1986). We also analyse the impact of nominal utterances over the corpus, counting how many turns do not contain any verb based on the PoS tagger output. While in the WhatsApp corpus these utterances are around 29%, in our multimodal corpus they are 35%. According to previous studies (Comandini et al., 2018) this kind of construction is used to express emphasis, and is particularly frequent on social networks and in spoken language. Our results are in line with previous findings on social media language, but show also that in our multimodal corpus the presence of images may boost communicative economy, making verbs less necessary than in other text-based media.

Concerning message length, both corpora contain rather short messages, with 5.9 tokens per sentence in the WhatsApp data and 5.3 tokens in the multimodal corpus on average. The standard deviation is rather high in both cases (4.80 and 4.99 respectively) because of the high length variability of the messages, ranging from one word (e.g. *Copriti*, *Certo*) to max. 50 tokens per message in our

multimodal corpus and 67 in the WhatsApp one.

Question marks are abundant in the WhatsApp dataset (0.51 per sentence on average, vs. 0.19 in the other corpus) because it contains interactions including questions and answers. Exclamation marks, instead, are much more frequent in the multimodal corpus (0.14 per sentence vs. 0.03), which contains more emphatic comments.

7 Conclusions

In this work we present a multimodal dataset in the abusive language domain created by teenage participants who, during online annotation sessions, judged whether images may trigger an offensive comment, left a possible comment and also assigned to it a trigger category.

The analysis of the collected data gives interesting insights into how Instagram-like platforms work. First of all, images containing persons are more likely to trigger potentially offensive comments than those without a human subject. Both female and male subjects are offended but the reasons may differ: the former are targeted more because of the pose and of the clothing, while the latter for the pose and the facial expression. In general, the reasons why a comment is triggered seems to be subjective, depending on the user leaving the comment rather on some actual characteristics of the person portrayed in the picture.

We conducted our data collection using Instagram pictures randomly taken from this platform, because it is the social network that is most used by teenagers, including those involved in our annotation sessions. However, this makes the release of the full dataset impossible, because of a very restrictive policy concerning images. We therefore adopt a strategy already used within the NLP research community (Kruk et al., 2019), releasing the images as a layer of a ResNet-18 neural network trained on ImageNet. The comments, instead, are freely available without restrictions due to the consent signed by all parents and by the anonymity granted to participants. This represents a very interesting dataset from a research point of view, since it includes comments written by underage students that are usually difficult to obtain because of privacy reasons.

In the future, we plan to extend our study to compare the judgements given by single users to those given by groups of peers. In a preliminary study, we observed that, when students are given

the possibility to discuss with a small group of peers whether they would like to write an offensive comment, they tend to be more aggressive and are more likely to select ‘yes’. While the comments collected so far with groups of annotators are not enough to allow a fair comparison between the two settings (single vs. group), we plan to extend them in the future and pursue also this interesting research line. Finally, we plan to train a classifier able to detect offensive messages by merging visual and textual features with the goal to integrate it in a monitoring tool like the one introduced in (Menini et al., 2019). This would enable a more holistic, context-aware understanding of offensive communication online.

Acknowledgments

Part of this work has been funded by the KID_ACTIONS REC-AG project (n. 101005518) on “Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologies”. We would like to thank the students involved in the corpus creation process and their teachers who supported the experimentation.

References

- Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Mario Ezra Aragón, Miguel Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF)*, Bilbao, Spain.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SEMEVAL 2019*, Minneapolis, Minnesota, USA, June.
- Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software. In *3rd Annual ACM Web Science Conference (WebSci 2011)*, pages 1–2.

- Cristina Bosco, Felice Dell’Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of EVALITA 2018*.
- Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 235–243. SIAM.
- Gloria Comandini, Manuela Speranza, and Bernardo Magnini. 2018. Effective Communication without Verbs? Sure! Identification of Nominal Utterances in Italian Social Media Texts. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. 2019. Current Limitations in Cyberbullying Detection: on Evaluation Criteria, Reproducibility, and Data Scarcity.
- Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishr. 2015. Prediction of cyberbullying incidents on the Instagram social network. *arXiv preprint arXiv:1508.06257*.
- Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.
- Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4621–4631.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110.
- Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2021. The CREENDER Tool for Creating Multimodal Datasets of Images and Comments. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*.
- Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE.
- Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.
- Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.
- Michael Stubbs. 1986. Lexical density: A technique and some findings. In Michael Coulthard, editor, *Talking about Text. Discourse Analysis*, chapter Lexical Density: A Technique and Some Findings, page 27–42. English Language Research.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.
- Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy, August. Association for Computational Linguistics.

A Resource for Detecting Misspellings and Denoising Medical Text Data

Enrico Mensa* **Gian Manuel Marino[†]** **Davide Colla***
Università di Torino, Università di Torino, Università di Torino,
Dipartimento di Informatica Dipartimento di Informatica Dipartimento di Informatica

Matteo Delsanto*
Università di Torino,
Dipartimento di Informatica

Daniele P. Radicioni*
Università di Torino,
Dipartimento di Informatica

*{firstname.surname}@unito.it; [†]marino.jnf@gmail.com

Abstract

English. In this paper we propose a method for collecting a dictionary to deal with noisy medical text documents. The quality of such Italian Emergency Room Reports is so poor that in most cases these can be hardly automatically elaborated; this also holds for other languages (e.g., English), with the notable difference that no Italian dictionary has been proposed to deal with this jargon. In this work we introduce and evaluate a resource designed to fill this gap.¹

Italiano. *In questo lavoro illustriamo un metodo per la costruzione di un dizionario dedicato all'elaborazione di documenti medici, la porzione delle cartelle cliniche annotata nei reparti di pronto soccorso. Questo tipo di documenti è così rumoroso che in genere le cartelle cliniche difficilmente possono essere direttamente elaborate in maniera automatica. Pur essendo il problema di ripulire questo tipo di documenti un problema rilevante e diffuso, non esisteva un dizionario completo per trattare questo linguaggio settoriale. In questo lavoro proponiamo e valutiamo una risorsa finalizzata a condurre questo tipo di elaborazione sulle cartelle cliniche.*

1 Introduction

Noise in textual data is a very common phenomenon afflicting text documents, especially

when dealing with informal texts such as chats, SMS and e-mails. This kind of text inherently contains spelling errors, special characters, non-standard word forms, grammar mistakes, and so on (Liu et al., 2012). In this work we focus on a type of text which can also be very noisy: *emergency room reports*. In the broader frame of a project aimed at detecting injuries stemming from violence acts in narrative texts contained in emergency room reports, we recently developed the VIDES, so dubbed after ‘Violence Detection System’ (Mensa et al., 2020). This system is concerned with categorizing textual descriptions as containing violence-related injuries (V) vs. non-violence-related injuries (NV), which is a relevant task to the ends of devising alerting mechanisms to track and prevent violence episodes. VIDES combines a neural architecture which performs the categorization step (thus discriminating V and NV records) and a Framenet-based approach, whereby semantic roles are represented through a synthetic description employing a set of word embeddings.² More specifically, a model of violent event has been devised: records that are recognized as containing violence-related injuries are further processed by an explanation module, which is charged to individuate the main elements corroborating that categorization (V) by identifying the involved agent, the type of injury, the involved body district *etc.*. Explaining the categorization ultimately involves filling the semantic components of the violence frame. All such ele-

¹Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²Related approaches have been designed as Semantic Role Labeling tasks (Gildea and Jurafsky, 2002; Zapirain et al., 2013), but also frame-based approaches have been proposed, paired to deep syntactic analysis, to extract salient information through a template-filling approach (Lesmo et al., 2009; Gianfelice et al., 2013).

ments contribute to recognizing a violent event as the source of the injuries complained by ER patients.

During the development of VIDES we realized that in order to run sophisticated algorithms for the detection and extraction of such violent traits we needed to cope with the noise contained in the input medical records. Some efforts have been invested to deal with different sorts of linguistic phenomena menacing the comprehension of texts; however, most existing works are focused on the English language, and rely on dictionaries that cannot be directly employed on Italian text documents.

In this preliminary work we start to tackle the issue of noisy words in medical records for Italian texts, by specifically focusing on misspellings. Our contribution is twofold: we first manually explore the dataset by analyzing a small sample of records in order to determine whether the main traits and issues present in other languages are also shared by Italian reports; secondly, we collect, merge and evaluate a set of Italian dictionaries, which constitute a brick fundamental to build any domain specific spell-checking algorithm (López-Hernández et al., 2019).

2 Related Work

Literature shows a limited but significant interest on the issue of detecting and correcting noisy medical text documents; nonetheless, some commonalities underlying this sort of text can be drawn.

Medical texts are often very noisy; among the most common mistakes we mention mistyping, lack or improper use of punctuation, grammatical errors and domain-specific abbreviations and Latin medical terminology (Siklósi et al., 2013). This is mainly due to the nature of the records themselves, and to the fact that the medical personnel compiling the entries is often under pressure and in a hurry.

Most of the spelling correction approaches have been carried out for English, with the exception of research in Swedish (Dziadek et al., 2017) and Hungarian (Siklósi et al., 2013), while no work has been found dealing with the Italian language. Regarding the methodologies, most works focus on non-word errors, while disregarding grammatical and real word mistakes. Non-word mistakes occur when a misspelling error produces a word that does not exist, such as ‘patienz’ instead of

‘patient’, while real word mistakes occur when a word is mistakenly replaced with another – existing – one, like the substitution of ‘abuse’ with ‘amuse’. The adopted algorithms are diverse, with the prevalence of approaches relying on embeddings (Kilicoglu et al., 2015; Workman et al., 2019) or regular expressions and rule-based systems (Patrick et al., 2010; Sayle et al., 2012; Lai et al., 2015). However, basically all contributions adopt a preliminary *dictionary look-up* step (López-Hernández et al., 2019). To this purpose, besides the general dictionaries provided in toolkits such as Aspell and Google Spell Checker,³ authors often rely on (medical) domain-specific dictionaries, such as The Unified Medical Language System (UMLS) (Aoki et al., 2004), the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT, 2020) and The SPECIALIST Lexicon (Browne et al., 2000). It is thus evident that the development of analogous resources for the Italian language is a crucial step for the design of tools and systems aimed at dealing with the spell-checking of Italian medical text documents.

Besides the treatment of misspellings, there are also works specifically focused on abbreviations. For instance, in (Wu et al., 2011) the authors present a corpus-based method to create a lexical resource of English clinical abbreviations via several machine learning algorithms. The resource has been used to automatically detect and expand abbreviations, and obtained interesting experimental results. More recently, another approach proposed in (Kreuzthaler et al., 2016) focuses on abbreviations ending with a period character; the proposed technique puts together statistical and dictionary-based strategies to detect abbreviations in German clinical narratives.

In the present work we are not proposing a specific technique for dealing with abbreviations, we are rather concerned with misspellings. However, the approaches already proposed for other languages will be considered in future work to also treat Italian abbreviations in our dataset.

3 Data Analysis

We analyze real data coming from a set of emergency room reports collected in Italian hospitals by the Italian National Institute of Health in the

³<https://git.savannah.gnu.org/git/aspell.git> and <https://languagetool.org>, respectively.

Table 1: Figures describing the complete dataset and the sample selected for manual annotation.

	Complete	Sample
Number of entries	136,144	592
Number of tokens	2,329,840	14,137
Number of unique tokens	49,116	1,842
Avg tokens per record	17.11	23.88

frame of the SINIACA project (Pitidis et al., 2014). The SINIACA project, so dubbed after ‘Sistema Informativo Nazionale sugli Incidenti in Ambiente di Civile Abitazione’ (National Information System on Accidents in Civil Housing Environment), is the Italian branch of the European Injury Database (EU-IDB) (Lyons et al., 2015), an EU-wide surveillance system concerned with accidents, collecting data from hospital emergency departments according to the EU recommendation no. C 164/2007/01, aimed at injury prevention and safety promotion.

Dataset. The whole dataset amounts to 136,144 non-empty entries, 592 of which were randomly selected for the manual analysis. Table 1 reports some figures describing the dataset. Double spaces and punctuation redundancy have been fixed through regular expressions, while tokens have been extracted by splitting the sentences based on spaces. Also, tokens containing numbers are presently discarded.

Analysis result. We performed a manual analysis on the subset of the original dataset: the 592 randomly selected entries herein were manually examined, and for each entry we looked for noisy words. Three main types of words were annotated: *i) misspellings*: a wrongly typed word, e.g., *fratura* instead of *frattura* – fracture; *ii) abbreviations*: a shortened form of a word or phrase, e.g., *dx* instead of *destra* – right; *iii) acronyms*: a word formed from the initial letters of other words, e.g., *ps* instead of *pronto soccorso* – emergency room. Interestingly enough, both abbreviations and acronyms can be at least partly considered as domain dependent: for example, in different settings, *ps* may denote *post scriptum* (something added at a later time, likely a letter, after the signature), but also ‘Polizia di Stato’ (Police) or ‘previdenza sociale’ (social security). Dealing with such phenomena thus involves access-

Table 2: Noise distribution on the annotated dataset; between parenthesis we report the percentage over the total number of tokens, while the last column indicates the average per record.

	With repetitions	Unique	Average
Noisy tokens	1,336 (9.4%)	424 (3%)	2.25
Misspellings	433 (3%)	304 (2%)	0.73
Abbreviations	670 (4.7%)	76 (0.5%)	1.13
Acronyms	233 (1.6%)	45 (0.3%)	0.39

ing a context dependent knowledge base that allows selecting the utterance appropriate for the context at hand. We are presently concerned with misspellings, acronyms and abbreviations as *noise*, but only the first category can be actually considered as an error. More specifically, while misspellings are actual errors, abbreviations and acronyms belong to a domain-specific language, and these are way too specific to be recognized as legitimate words through a general-purpose dictionary. As seen in literature, misspellings and abbreviations/acronyms must be treated with different techniques, and in this work we mainly focus on tackling the first category, while also obtaining interesting insights regarding the second one.

Table 2 illustrates the results of the annotation process. We discovered that the dataset contains a lot of noise, amounting to almost the 10% of the tokens, on average 2 noisy tokens per record. By looking separately at the different typologies of noise we observe that misspellings are more scattered and diverse, while the usage of abbreviations and acronyms seems to be more coherent: we have 670 instances of abbreviations but only 76 unique abbreviations, while 304 out of the 433 instances of misspellings are unique. This phenomenon is also depicted in Figure 1, where we provide the log-log plot of the frequency of each misspelling, abbreviation and acronym ordered by rank. We observe that the distribution of abbreviations and acronyms has a different magnitude, but is very similar in shape; on the other side, the misspellings are clearly more scattered with a very long tail of items appearing only once.

4 Dictionaries Creation and Evaluation

The manual analysis uncovered characteristics and features that are in line with those found in literature for English datasets (López-Hernández et al., 2019). However, to allow the development

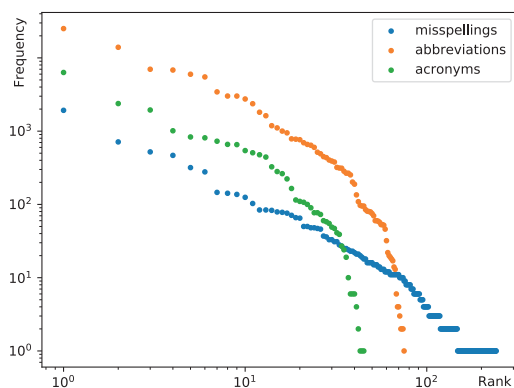


Figure 1: Log-log plot showing the frequency of misspellings, abbreviations and acronyms over the annotated dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of spell-checkers for Italian medical texts, another key component is still missing: most approaches aimed at error detection rely on dictionaries to determine if a token is a legitimate word or not. In fact, the simplest implementation of misspellings detection is as follows: if we have at our disposal the set W containing all of the terms of a given language, joined to all terms pertaining the specific domain at hand, any word $w \notin W$ can be likely considered as a misspell. To the best of our knowledge, no such dictionary exists that is able to cope with Italian medical text documents, so we built a resource to answer to this need.

4.1 Source Dictionaries

The automatic development of a dictionary is not a trivial task. We want to reach the highest possible coverage for both general terms and specific medical terminology, but at the same time we cannot rely too much on unverified sources (e.g., crowd-sourced data) with the risk of introducing misspellings and errors into the dictionary. We selected different sources and arranged them into four main classes:

- **MED**: a collection of medical terms built by putting together five medical online dictionaries (torrinomedica.it, 2020; abcsalute.it, 2020; codifa.it, 2020; my-personaltrainer.it, 2020a; my-personaltrainer.it, 2020b), containing medical specific terms and medications names;
- **ITA**: a collection of Italian terms built by

Table 3: Figures of the 5000 annotated tokens used to evaluate the dictionaries.

Class	Type	Amount (% on total)
Positive	Correct words	3,886 (77.7%)
	Abbreviations	184 (3.7%)
	Acronyms	126 (2.5%)
Negative	Misspells	804 (16.1%)

merging three well-known Italian online dictionaries (Hoepli, 2020; Sabatini-Colletti, 2020; De Mauro, 2020);

- **WMED**: a collection of terms from Wikipedia pages pertaining the medical domain. The list of Wikipedia medical pages has been obtained by querying the SPARQL endpoint of Wikidata (Vrandečić and Krötzsch, 2014), while the pages have been taken from the 20 August 2020 Wikipedia dump;
- **WMOV**: since medical records also contain a brief narrative text of the events that led to the (either violent or accidental) injuries, we added terms associated to eventive and narrative genres by collecting Wikipedia pages pertaining to movies, television series and literary work that are expected to contain narrative terminology.

The set of terms extracted from Wikipedia can potentially contain misspellings and errors, and so we also set a frequency minimum which allows for the pruning of the tokens herein. We annotate this parameter with a subscript next to the set name, e.g., $WMOV_1$ indicates that the threshold was set to 1 for the terms frequency.

4.2 Evaluation

Building the dataset. In order to assess the quality of the collected dictionaries we started from the 49,116 unique tokens in the dataset, removed the stop words⁴ and randomly selected 5,000 of them to be manually annotated. The annotation was carried out by four of the authors of this paper. The selection algorithm was designed so to increase the probability of a token to be selected in accordance to its frequency in the dataset. These 5,000 tokens were then annotated

⁴We used the set of stop words made available by Spacy (<https://spacy.io/>) for the Italian language.

Table 4: Results of the evaluation of the considered dictionaries. The first column reports the size of each dictionary, the second to fourth columns provide coverage and correctness along with their harmonic mean, while the last three columns illustrate the coverage of our dictionaries on tokens that were annotated as correct words, abbreviations and acronyms.

	Terms	Coverage	Correctness	F1-Score	Correct Words	Abbreviations	Acronyms
ITA	124,494	.542	.980	.700	.573	.179	.206
MED, ITA	155,650	.621	.975	.759	.652	.228	.261
MED, ITA, WMED ₀	287,279	.897	.907	.902	.918	.521	.785
MED, ITA, WMED ₀ , WMOV ₀	511,827	.926	.863	.894	.941	.641	.873
MED, ITA, WMED ₁ , WMOV ₁	343,264	.906	.898	.902	.925	.586	.793
MED, ITA, WMED ₁ , WMOV ₅	266,633	.892	.922	.907	.912	.554	.761
(LEM) MED, ITA, WMED ₁ , WMOV ₅	227,895	.903	.896	.900	.926	.559	.674

with one of the following four classes: correct words (regardless of their domain specificity), abbreviations, acronyms and misspellings. The first three classes represent terms that should be found in our resource, while the last category contains words that should not be present in the dictionary. Table 3 reports the statistics featuring the dataset annotated for evaluation purposes.

Evaluating the dictionary. In Table 4 we report the results of the dictionaries evaluation. Each dictionary has been built by taking into consideration one or more of the previously presented sources. Multiple sources have been simply merged into a unique set of terms, without repetitions. We assess the quality of each dictionary via two measures, coverage and correctness. The *coverage* is the percentage of words that were found in the dictionary (either correct words, abbreviations or acronyms), while the *correctness* is the percentage of misspellings that were not present in the dictionary. We considered different combinations of the sources, the tuning of the frequency-based filtering parameter, and an additional lemmatization step.

We observe that both the ITA and the MED sets are fundamentally correct, even though they also include words that in the common usage are frequently misspelled, such as *passaggiere* in place of the correct form *passaggero*. On the other side, its .62 coverage is unsatisfactory (please refer to the second row of Table 4, MED, ITA); it also witnesses that medical jargon is only partially grasped by dictionaries in the MED set. As expected, the introduction of terms from Wikipedia improves the coverage, but with detrimental effect on the correctness. This also holds for the WMOV set, which is rich but also pretty noisy. By fine tuning the frequency thresholds of both WMED and WMOV we were able to prune most of the noise and to pre-

serve the coverage at the same time, finally obtaining a good dictionary with the combination MED, ITA, WMED₁, WMOV₅.

This setting was also tested by applying a lemmatization step on both Wikipedia terms and our dataset tokens. Interestingly, the lemmatization introduces more mistakes than it solves: this is due the the fact that unpredictably the lemmatizer converts misspellings into legitimate words that do not necessarily correspond to their correct spelling. This fact shows also that lemmatization, which is acknowledged as a task almost completely solved from a scientific point of view, still poses relevant issues for the medical jargon and for domain-specific languages more in general.

A lot of abbreviations are not yet covered in the dictionary. Once again, these abbreviations are dataset-specific (and perhaps also follow local uses rather than widely accepted practices), and thus these are very hard to find even on specialized public medical resources. For instance, *incid* (*incidente* – accident) appears very frequently and its easily understandable by humans but its not a common or medical abbreviation. The same phenomenon can also be observed on acronyms, that are less sparse and more adherent to widely accepted practices and standards.

5 Conclusions and Future Work

In this work we tackled the issue of detecting textual noise in Italian room emergency reports, focusing specifically on misspellings. Firstly we examined the reports and found out that the sorts of issues reported in literature for other languages can also be found in Italian text documents. Secondly, we developed and evaluated an Italian dictionary suited for the task of noise detection. In future work we plan to expand the dictionary by

including the terms from the Italian ICD-9 and ICD-10 (International Classification of Diseases), that may be useful to interpret acronyms and resolve abbreviations. Moreover, we plan to employ this dictionary in a fully fledged spell-checking system. Finally, the usage of semantic —sense indexed— representations such as, e.g., (Mensa et al., 2018) and (Colla et al., 2020a; Colla et al., 2020b) will be explored, in order to deal with real word mistakes, and more in general contextual information (Basile et al., 2019) will be considered as a main cue in order to uncover and correct this sort of errors. For example, by leveraging the terminology surrounding noisy tokens we plan to distinguish the more scattered misspellings from the other terms that are not present in our dictionary.

Acknowledgments

The first author was supported by a grant provided by Università degli Studi di Torino. This research is also supported by Fondazione CRT, RF 2019.2263.

References

- [abcsalute.it2020] abcsalute.it. 2020. Abcsalute.it - Dizionario Medico. <http://www.abcsalute.it/dizionario-medico>.
- [Aoki et al.2004] Kiyoko F Aoki, Atsuko Yamaguchi, Nobuhisa Ueda, Tatsuya Akutsu, Hiroshi Mamitsuka, Susumu Goto, and Minoru Kanehisa. 2004. Kcam (kegg carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic acids research*, 32(suppl_2):W267–W272.
- [Basile et al.2019] Valerio Basile, Tommaso Caselli, and Daniele P. Radicioni. 2019. Meaning in Context: Ontologically and linguistically motivated representations of objects and events. *Applied Ontology*, 14:335–341.
- [Browne et al.2000] Allen C Browne, Alexa T McCray, and Suresh Srinivasan. 2000. The specialist lexicon. *National Library of Medicine Technical Reports*, pages 18–21.
- [codifa.it2020] codifa.it. 2020. codifa.it - Dizionario dei Farmaci. <https://www.codifa.it/farmaci>.
- [Colla et al.2020a] Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020a. Lesslex: Linking multilingual embeddings to sense representations of lexical items. *Computational Linguistics*, 46(2):289–333.
- [Colla et al.2020b] Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020b. Novel metrics for computing semantic similarity with sense embeddings. *Knowledge-Based Systems*, 206:106346.
- [De Mauro2020] De Mauro. 2020. Dizionario Italiano Nuovo De Mauro. <https://dizionario.internazionale.it/>.
- [Dziadek et al.2017] Juliusz Dziadek, Aron Henriksson, and Martin Duneld. 2017. Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235:241.
- [Gianfelice et al.2013] Davide Gianfelice, Leonardo Lesmo, Monica Palmirani, Daniele Perlo, and Daniele P. Radicioni. 2013. Modificatory Provisions Detection: a Hybrid NLP Approach. In Bart Verheij, editor, *Proceedings of ICAIL 2013: XIV International Conference on Artificial Intelligence and Law*, pages 43–52. ACM.
- [Gildea and Jurafsky2002] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.
- [Hoepli2020] Hoepli. 2020. Dizionario Italiano Hoepli. <https://dizionari.repubblica.it/italiano.html>.
- [Kilicoglu et al.2015] Halil Kilicoglu, Marcelo Fisman, Kirk Roberts, and Dina Demner-Fushman. 2015. An ensemble method for spelling correction in consumer health questions. In *AMIA Annual Symposium Proceedings*, volume 2015, page 727. American Medical Informatics Association.
- [Kreuzthaler et al.2016] Markus Kreuzthaler, Michel Oleyunik, Alexander Avian, and Stefan Schulz. 2016. Unsupervised abbreviation detection in clinical narratives. In *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*, pages 91–98.
- [Lai et al.2015] Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics*, 55:188–195.
- [Lesmo et al.2009] Leonardo Lesmo, Alessandro Mazzei, and Daniele P. Radicioni. 2009. Extracting Semantic Annotations from Legal Texts. In *Proceedings of the International Conference on Hypertext, HT09*, pages 167–172, Turin, Italy, July. ACM.
- [Liu et al.2012] Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1035–1044.

- [López-Hernández et al.2019] Jérica López-Hernández, Ángela Almela, and Rafael Valencia-García. 2019. Automatic spelling detection and correction in the medical domain: A systematic literature review. In *International Conference on Technologies and Innovation*, pages 95–108. Springer.
- [Lyons et al.2015] Ronan Lyons, Rupert Kisse, and Wim Rogmans. 2015. Eu-injury database introduction to the functioning of the injury database (idb). <https://bit.ly/37FAKaB>.
- [Mensa et al.2018] Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2018. Cover: a linguistic resource combining common sense and lexicographic information. *Language Resources and Evaluation*, 52(4):921–948.
- [Mensa et al.2020] Enrico Mensa, Davide Colla, Marco Dalmaso, Marco Giustini, Carlo Mamo, Alessio Pitidis, and Daniele P. Radicioni. 2020. Violence detection explanation via semantic roles embeddings. *BMC Medical Informatics and Decision Making*, 20(1):263–275, Oct.
- [my-personaltrainer.it2020a] my-personaltrainer.it. 2020a. Lista delle Malattie di My Personal Trainer. https://www.my-personaltrainer.it/malattie_a_z.php.
- [my-personaltrainer.it2020b] my-personaltrainer.it. 2020b. Lista di Sintomi di My Personal Trainer. https://www.my-personaltrainer.it/sintomi_a_z.php.
- [Patrick et al.2010] Jon Patrick, Mojtaba Sabbagh, Suvir Jain, and Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *Proceedings of 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2010)*, pages 1–8.
- [Pitidis et al.2014] Alessio Pitidis, Gianni Fondi, Marco Giustini, Eloïse Longo, Giuseppe Balducci, Gruppo di lavoro SINIACA-IDB, and Dipartimento di Ambiente e Connessa Prevenzione Primaria, ISS. 2014. Il Sistema SINIACA-IDB per la sorveglianza degli incidenti. *Notiziario dell'Istituto Superiore di Sanità*, 27(2):11–16.
- [Sabatini-Colletti2020] Sabatini-Colletti. 2020. Dizionario Italiano Sabatini Colletti. https://dizionari.corriere.it/dizionario_italiano/.
- [Sayle et al.2012] Roger Sayle, Paul Hongxing Xie, and Sorel Muresan. 2012. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *Journal of chemical information and modeling*, 52(1):51–62.
- [Siklósi et al.2013] Borbála Siklósi, Attila Novák, and Gábor Prószyk. 2013. Context-aware correction of spelling errors in hungarian medical documents. In *International Conference on Statistical Language and Speech Processing*, pages 248–259. Springer.
- [SNOMED-CT2020] SNOMED-CT. 2020. International Health Terminology Standards Development Organisation. <http://www.ihtsdo.org/snomed-ct/>.
- [torrinomedica.it2020] torrinomedica.it. 2020. torrinomedica.it - Dizionario dei Farmaci. <https://www.torrinomedica.it/schede-farmaci>.
- [Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.
- [Workman et al.2019] T Elizabeth Workman, Yijun Shao, Guy Divita, and Qing Zeng-Treitler. 2019. An efficient prototype method to identify and correct misspellings in clinical text. *BMC research notes*, 12(1):1–5.
- [Wu et al.2011] Yonghui Wu, S Trent Rosenbloom, Joshua C Denny, Randolph A Miller, Subramani Mani, Dario A Giuse, and Hua Xu. 2011. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1541. American Medical Informatics Association.
- [Zapirain et al.2013] Benat Zapirain, Eneko Agirre, Lluís Marquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Is Neural Language Model Perplexity Related to Readability?

Alessio Miaschi^{*,*}, Chiara Alzetta^{*,*}, Dominique Brunato^{*},
Felice Dell’Orletta^{*}, Giulia Venturi^{*}

^{*}Department of Computer Science, University of Pisa

^{*}Istituto di Linguistica Computazionale “Antonio Zampolli”, ItaliaNLP Lab, Pisa

^{*}DIBRIS, University of Genoa

alessio.miaschi@phd.unipi.it, chiara.alzetta@edu.unige.it,
{name.surname}@ilc.cnr.it

Abstract

This paper explores the relationship between Neural Language Model (NLM) perplexity and sentence readability. Starting from the evidence that NLMs implicitly acquire sophisticated linguistic knowledge from a huge amount of training data, our goal is to investigate whether perplexity is affected by linguistic features used to automatically assess sentence readability and if there is a correlation between the two metrics. Our findings suggest that this correlation is actually quite weak and the two metrics are affected by different linguistic phenomena.¹

1 Introduction and Motivation

Standard Neural Language Models (NLMs) are trained to predict the next token given a context of previous tokens. The metric commonly used for assessing the performance of a language model is perplexity, which corresponds to the inverse geometric mean of the joint probability of words w_1, \dots, w_n in a held-out test corpus C . While being primarily an intrinsic metric of NLM quality, perplexity has been used in a variety of scenarios, such as to classify between formal and colloquial tweets (González, 2015), to detect the boundaries between varieties belonging to the same language family (Gamallo et al., 2017) or to identify speech samples produced by subjects with cognitive and/or language diseases e.g. dementia (Cohen and Pakhomov, 2020) or Specific Language Impairment (Gabani et al., 2009). From the perspective of computational studies aimed at modeling human language processing, perplexity scores have also been shown to effectively match various

human behavioural measures, such as gaze duration during reading (Demberg and Keller, 2008; Goodkind and Bicknell, 2018).

In this paper we focus on a less investigated perspective addressing the connection between perplexity and readability. Since by definition perplexity gives a good approximation of how well a model recognises an unseen piece of text as a plausible one, our intuition is that lower model perplexity should be assigned to easy-to-read sentences, while difficult-to-read ones should obtain higher perplexity. On the other hand, state-of-the-art NLMs trained on huge data have shown to implicitly learn a sophisticated knowledge of language phenomena, also with respect to complex syntactic properties of sentences (Tenney et al., 2019; Jawahar et al., 2019; Miaschi et al., 2020). This could suggest that variations in terms of linguistic complexity, especially when related to subtle morpho-syntactic and syntactic features of sentence rather than lexical ones, could not impact on model perplexity to a great extent. This assumption seems to be confirmed by the (still unpublished) results by Martinc et al. (2019) which, to our knowledge, is the only one explicitly leveraging unsupervised neural language model predictions in the context of readability assessment. According to this study, a NLM is even less perplexed by articles addressed at adults than by documents conceived for a younger readership. From a relatively different perspective focused on the ability of automatic comprehension systems to solve cloze tests, Benzahra and Yvon (2019) showed that NLMs performance is not affected by the level of text complexity.

In order to test the validity of all these hypotheses, we rely on the perplexity score given by a state-of-the-art NLM for the Italian language to several datasets representative of different textual genres containing both easy- and complex-to-read sentences: ideally, such datasets should

¹Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

emphasise the correlation between perplexity and readability (if present) since the corpora are explicitly designed to contain both simple and difficult examples.

Contributions We inspect whether and to which extent it is possible to find a relationship between a readability score and the perplexity of a NLM. To this aim we investigate (i) if the perplexity of a NLM and the readability score of a set of sentences show a significant correlation and (ii) whether the two metrics are equally affected by the same set of linguistic phenomena that occur in the sentence.

2 Experimental Design

According to our research questions, we devised a set of experiments to study whether NLMs perplexity reflects the level of readability of a sentence and which are the linguistic phenomena mostly involved in each metric. For this purpose, we firstly investigated whether sentence-level perplexity scores computed with one of the most prominent NLM model correlate with the scores assigned to the same sentences by a supervised readability assessment tool. Secondly, we investigated which are the linguistic features of the considered sentences that correlate in a statistically significant way with the perplexity and readability score respectively. In order to verify whether correlations hold across different typology of texts, we tested our approach on five Italian datasets.

2.1 Models

READ-IT. Automatic readability (henceforth *ARA*) was assessed using READ-IT (Dell’Orletta et al., 2011) the first readability assessment tool for Italian which combines traditional raw text features with lexical, morpho-syntactic and syntactic information extracted from automatically parsed documents. In READ-IT, analysis of readability is modelled as a binary classification task, based on Support Vector Machines using LIBSVM (Chang and Lin, 2001). Training corpora are representative of two classes of texts, i.e. difficult– vs. easy–to-read ones, both containing newspaper articles. The set of features exploited for predicting readability has been proved to capture different aspects of sentence complexity. Thus, the assigned readability score ranges between 0 (easy-to-read) and 1 (difficult-to-read) referring to the percentage probability for unseen documents or sentences to be-

long to the class of difficult-to-read documents. For the purposes of our work, we carried out readability assessment at sentence level, making the analysis reliable for the comparison with sentence-based perplexity of a NLM.

GePpeTto. Sentence-level perplexity scores were computed relying on GePpeTto (De Mattei et al., 2020). GePpeTto is a generative language model trained on the Italian language and built using the GPT-2 architecture (Radford et al., 2019). The model was trained on a dump of Italian Wikipedia (2.8GB) and on the itWac corpus (Baroni et al., 2009), which amounts to 11GB of web texts. The perplexity (PPL) of the model was computed as follows:

$$PPL = e^{\left(\frac{NLL}{N}\right)}$$

where NLL and N correspond respectively to the negative log-likelihood and to the length of each sentence $w_{1:n} = [w_1, \dots, w_n]$ in the datasets.

2.2 Corpora

In order to test the reliability of our initial hypothesis, we chose four corpora containing different typologies of texts, i.e. web pages, educational materials, narrative texts, newspaper and scientific articles. Each corpus includes a balanced amount of difficult- and easy-to-read sentence. In addition, we also considered in the analysis the Italian Universal Dependency treebank. This is meant to verify whether the connection between sentence-level readability and perplexity also holds in a well-acknowledged benchmark corpus. For each of them, we excluded from our analysis short sentences, i.e. having less than 5 tokens.

PACCSS-IT² (Brunato et al., 2016): we took into account 125,977 sentences belonging to PACCSS-IT, a corpus of complex-simple aligned sentences extracted from the ItWaC corpus. The resource was build using an automatic approach for acquiring large corpora of paired sentences able to intercept structural transformations (such as deletion, reordering, etc.). For example, the two following sentences represent a pair in the corpus, where a reordering operation occurs at phrase level (i.e. the subordinate clause proceeds vs. follows the main clause):

- Complex: *Ringraziandola per la sua cortese attenzione, resto in attesa di risposta.* [Lit:

²<http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/>

Thanking you for your kind attention, I look forward to your answer.]

- Simple: *Resto in attesa di una risposta e ringrazio vivamente per l'attenzione.* [Lit: I look forward to your answer and I thank you greatly for your attention.]

Terence and Teacher³ (Brunato et al., 2015): two corpora of original and manually simplified texts aligned at sentence level. *Terence* contains short Italian novels for children and their manually simplified version carried out by linguists and psycholinguists targeting children with text comprehension difficulties. *Teacher* is a corpus of pairs of documents belonging to different genres (e.g. literature, handbooks) used in educational settings manually simplified by teachers. We exploited 1,644 sentences belonging to these corpora.

Multi-Genre Multi-Type Italian corpus: a collection of Italian texts representative of three traditional textual genres: Journalism, Scientific prose and Narrative. Each genre has been internally subdivided into two sub-corpora representative of an easy- vs difficult-to-read variety, which was defined according to the intended target audience for a given genre. The journalistic prose corpus includes articles automatically downloaded from the online versions of two general-purpose newspapers⁴, while the “easy” sub-corpus contains articles from two easy-to-read newspapers⁵ addressed to adults with low literacy skills or mild intellectual disabilities. The scientific prose collection consists of scholarly publications on linguistics and computational linguistics and Wikipedia pages downloaded from the portal “Linguistics”, representative of the complex and easy variety respectively. For the narrative genre, we included long novels written by novelists of the last century and contemporary writers in the corpora of complex variety, while for the easy variety we collected short novels for children. The complete corpus contains 56,685 sentences.

Italian Universal Dependency Treebank: it includes different sections of the Italian Universal Dependency Treebank (IUDT), version 2.5 (Zeman et al., 2019). In particular, we considered two groups: a first one containing the whole Italian

Stanford Dependency Treebank (ISDT)⁶ (Bosco et al., 2013), the Italian version of the multilingual Turin University Parallel Treebank (Sanguinetti and Bosco, 2015) and the Venice Italian Treebank (Delmonte et al., 2007) (24,998 sentences), all containing a mix of textual genres; and a second one including two collections of texts representative of social media language, i.e. generic tweets and tweets labelled for irony (PoSTWITA⁷ and TWITTITIRO⁸) (Sanguinetti et al., 2018; Cignarella et al., 2019) (3,660 sentences in total).

3 Sentence Perplexity and Readability

Our analysis starts from a comparison between the average perplexity and readability scores obtained for each sentence of the five considered datasets. As shown in Table 1, readability values (column *ARA*) are quite homogeneous across the datasets, with low standard deviation values. On the contrary, the range of perplexity scores is wider (column *PPL*), going from an average score of 3,905.83 of PACCSS-IT to 436.75 of the IUDT miscellaneous portion (Italian UD). These differences seem to provide a first evidence that perplexity and readability are not correlate to each other.

This intuition has been proved computing the Spearman’s rank correlation coefficient between the perplexity and readability scores for each dataset. Results are reported in Table 2, column *PPL-ARA*. As it can be seen, all correlation rates are significant, except for the result obtained on the Terence and Teacher corpus, possibly due to the fact that the size of the corpus is too small to allow a significant comparison. Contrary to our expectations, no correlation was detected between the two metrics for all corpora, suggesting that perplexity and readability are independent from each other.

To further investigate the reasons behind these scores and to deepen the analysis about the relationship between the two metrics, we investigated whether they capture the same (or similar) linguistic properties of the sentences. To this aim, we tested the presence and strength of the correlation between each of the two metrics and a set of 176 linguistic features, which have been shown to capture properties of sentence complex-

³<http://www.italianlp.it/resources/terence-and-teacher/>

⁴www.repubblica.it and <http://www.ilgiornale.it/>

⁵www.dueparole.it and <http://www.informazionefacile.it/>

⁶<https://github.com/UniversalDependencies/UD.Italian-ISDT>

⁷<https://github.com/UniversalDependencies/UD.Italian-PoSTWITA>

⁸<https://universaldependencies.org/treebanks/it-twittiro>

Dataset	PPL	ARA
<i>PACCSS-IT</i>	3,905.83 (\pm 21,306.07)	0.55 (\pm 0.24)
<i>Terence-Teacher</i>	790.85 (\pm 5,002.62)	0.46 (\pm 0.27)
<i>Multi-Genre Multi-Type</i>	570.85 (\pm 4,820.12)	0.58 (\pm 0.31)
<i>Italian-UD</i>	436.75 (\pm 3,633.64)	0.61 (\pm 0.30)
<i>Twitter-UD</i>	986.28 (\pm 2,479.64)	0.59 (\pm 0.30)

Table 1: Perplexity (PPL) and Readability (ARA) mean and standard deviation values for the 5 datasets.

Dataset	PPL-ARA	Feats
<i>PACCSS-IT</i>	-0.031*	0.169*
<i>Terence-Teacher</i>	0.014	0.149
<i>Multi-Genre Multi-Type</i>	0.026*	0.184*
<i>Italian-UD</i>	-0.054*	0.332*
<i>Twitter-UD</i>	-0.038*	-0.037

Table 2: Spearman’s correlation coefficients between sentence-level perplexity and readability scores (PPL-ARA) and between rankings of linguistic features (Feats). Statistically significant correlations ($p < 0.05$) are marked with *.

ity (Brunato et al., 2018). In particular, this analysis is based on the set of features described in Brunato et al. (2020), which are acquired from raw, morpho-syntactic and syntactic levels of annotation. They range from basic information on the average sentence and word length, to lexical information about the internal composition of the vocabulary of the text (e.g. the distribution of lemmas belonging to the *Basic Italian Vocabulary* (De Mauro, 2000)). They also include morpho-syntactic information (e.g. POS distribution and of inflectional properties of verbs) and more complex aspects of sentence structure derived from syntactic annotation and modeling global and local properties of parsed tree structure, e.g. the relative order of subjects and objects with respect to the verb, the use of subordination. In order to extract these features, the considered corpora were morpho-syntactically annotated and dependency parsed by the UDPipe pipeline (Straka et al., 2016), with the exception of the IUDT corpus.

Column *Feats* of Table 2 illustrates the results of this analysis: we report the Spearman’s correlation coefficients between the two rankings of linguistic features, each ordered by strength of correlation between feature value and perplexity score and readability score respectively. Once again we observe rather weak correlation values, with the only exception of Italian-UD which is the only

one reporting a medium correlation (.332). Overall, these results corroborate our previous findings that the two metrics are not particularly related with each other, and they further suggest that the linguistic phenomena affecting the perplexity of NLM and the readability level of a sentence are very different. Consider for example the two following sentences:

- (1) *Il furto è avvenuto giovedì notte.*
The theft has taken place Thursday night.
- (2) *Il comitato di bioetica: no all’eutanasia.*
The bioethics committee: no to euthanasia.

While (1) is very easy-to-read, with a readability score of 0.25, but it has a quite high perplexity score, i.e. 40,737.81, (2) is quite difficult-to-read (ARA=1) but is has a very low perplexity score (PPL=11.24).

4 In-Depth Linguistic Investigation

To better explore the motivation behind these results, we performed an in-depth investigation aimed at understating the relationship between our set of linguistic features and the two metrics taken into consideration. Since we noticed that for all datasets a higher number of features correlates with ARA than with PPL, we selected those that are significantly correlated with the two metrics. The number of shared features varies for each dataset, depending on their size. For example, for the two smallest ones, i.e. Terence and Teacher and the UD Twitter Treebank, we could only consider 34.65% (61) and 44.88% (79) of the whole set of features respectively, while for the larger corpora the sub-set is wider: 81.81% (144) in PACCSS-IT, 78.97% (139) for Multi-Genre Multi-Type and 84.65% (149) for the IUD Treebank.

Table 3 shows the top ten features for each dataset, i.e. those that obtained the strongest correlation with both PPL and ARA. As expected, correlations are generally stronger between linguistic features and readability scores, although they

are lower than expected. This could be due to the fact that, even if the READ-IT classifier is trained with a similar set of features, the non-linear feature space makes it difficult to identify clear correlations with individual features. Similarly, our set of features seem to play only a marginal role on perplexity. However, this is not the case of the PACCSS-IT corpus, for which the set of considered linguistic features have an higher correlation with PPL. This can be possibly related to the partial overlap between the GePpeTto training data and the PACCSS-IT sentences, since the latter is drawn from the ItWac corpus which is included in the GePpeTto’s training.

Inspecting these results, we can also observe that correlations between features and PPL seem to be more affected by genre-specific characteristics. This is particularly clear if we consider the Italian UD Twitter treebank, for which among the top ten most correlated features we find some of them characterising social media language, e.g. symbols (*upos-xpos_dist_SYM*) or the vocative relation, which marks a dialogue participant addressed in a text along with the specification, specifically used for Twitter @-mentions (*dep_dist_vocative:mention*).

5 Conclusion

The paper presented a study aimed at investigating the relationship between two metrics computed at sentence-level, i.e. perplexity of a state-of-the-art NLM for the Italian language and readability score automatically assigned to a sentence by a supervised classifier. We carried out our analysis considering several datasets differing at the level of textual genre and language variety. Specifically, we observed that comparing the rankings obtained using the two metrics we cannot find any significant correlation, either between the scores of the two metrics or with respect to the set of linguistic features that mostly impact their values. Further investigation within this line of research will explore whether we can draw the same observations when a different NLM is exploited to compute sentence perplexity.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically pro-

PACCSS-IT			
PPL		ARA	
Feats	Corr	Feats	Corr
aux_num_pers_dist_Sing+3	0,53	xpos_dist_FF	0,34
dep_dist_cop	0,51	dep_dist_punct	0,32
avg_max_depth	0,50	upos_dist_PUNCT	0,32
upos_dist_ADP	0,50	ttr_form	0,29
xpos_dist_E	0,50	aux_mood_dist_Cnd	0,25
dep_dist_case	0,49	upos_dist_DET	0,25
n_tokens	0,48	dep_dist_det	0,25
dep_dist_root	0,48	ttr_lemma	0,22
xpos_dist_FS	0,48	upos_dist_NOUN	0,21
Terence and Teacher			
PPL		ARA	
Feats	Corr	Feats	Corr
xpos_dist_B	0,25	dep_dist_det	-0,39
verbs_num_pers_dist_Sing+3	0,23	upos_dist_DET	-0,38
lexical_density	0,22	upos_dist_NOUN	-0,37
dep_dist_advmod	0,21	xpos_dist_S	-0,37
upos_dist_ADV	0,21	xpos_dist_RD	-0,29
verbs_num_pers_dist_Plur+3	-0,16	upos_dist_ADV	0,27
xpos_dist_V	0,16	dep_dist_advmod	0,25
avg_token_per_clause	-0,16	xpos_dist_FF	0,25
upos_dist_VERB	0,14	avg_sub_chain_len	0,24
Multi-Genre Multi-Type			
PPL		ARA	
Feats	Corr	Feats	Corr
n_tokens	-0,19	principal_prop_dist	-0,42
dep_dist_root	0,19	ttr_form	-0,34
dep_dist_advmod	0,19	xpos_dist_FF	0,34
upos_dist_ADV	0,18	dep_dist_det	-0,33
n_prepositional_chains	-0,18	upos_dist_DET	-0,33
xpos_dist_B	0,18	upos_dist_PUNCT	0,33
upos_dist_ADP	-0,17	dep_dist_punct	0,33
xpos_dist_E	-0,17	xpos_dist_FB	0,31
ttr_lemma	0,16	sub_prop_dist	0,27
Italian UD Treebank			
PPL		ARA	
Feats	Corr	Feats	Corr
n_tokens	-0,27	principal_prop_dist	-0,53
dep_dist_root	0,27	sub_proposition_dist	0,40
n_prepositional_chains	-0,26	n_tokens	0,39
avg_max_depth	-0,24	dep_dist_root	-0,39
upos_dist_ADP	-0,24	ttr_form	-0,37
ttr_lemma	0,23	avg_max_depth	0,36
max_links_len	-0,23	avg_links_len	0,35
avg_max_links_len	-0,23	max_links_len	0,34
xpos_dist_E	-0,22	avg_max_links_len	0,34
Italian UD Twitter Treebank			
PPL		ARA	
Feats	Corr	Feats	Corr
upos_dist_SYM	0,38	upos_dist_PUNCT	0,30
avg_max_depth	-0,28	dep_dist_punct	0,30
xpos_dist_SYM	0,28	dep_dist_det	-0,29
in_dict	-0,24	upos_dist_DET	-0,29
dep_dist_vocative:mention	0,23	verbal_root_perc	-0,27
in_dict_types	-0,22	xpos_dist_RD	-0,27
ttr_lemma	0,21	avg_token_per_clause	-0,27
in_FO	-0,21	subj_pre	-0,27
verbal_head_per_sent	-0,19	obj_post	-0,24

Table 3: Top 10 features along with their correlation scores between perplexity and readability.

cessed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Marc Benzahra and François Yvon. 2019. Measuring text readability with machine comprehension: a pilot study. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 412–422, Florence, Italy, August. Association for Computational Linguistics.

C. Bosco, S. Montemagni, and M. Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL*

- Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August.
- Dominique Brunato, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas, November. Association for Computational Linguistics.
- Dominique Brunato, Lorenzo De Mattei, Felice Dell’Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France, May. European Language Resources Association.
- Chih-Chung Chang and Chih-Jen Lin. 2001. LIB-SVM: a library for support vector machines.
- Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*.
- Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer’s type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, Online, July. Association for Computational Linguistics.
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.
- Tullio De Mauro. 2000. *Il dizionario della lingua italiana*, volume 1. Paravia.
- Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.
- Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT - Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*.
- V. Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.
- Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa Bedore, and Elizabeth Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–55, Boulder, Colorado, June. Association for Computational Linguistics.
- Pablo Gamallo, Jose Ramon Pichel, and Iñaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *VarDial2017 workshop at EACL 2017. Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–114, Valencia, Spain, April 3, 2017. c©2017 Association for Computational Linguistics (<http://web.science.mq.edu.au/smalmasi/wardial4/index.html>).
- M. González. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *TweetMT@SEPLN*.
- Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In Asad B. Sayeed, Cassandra Jacobs, Tal Linzen, and Marten Van Schijndel, editors, *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018, Salt Lake City, Utah, USA, January 7, 2018*, pages 10–18. Association for Computational Linguistics.
- Ganesh Jawahar, Benoît Sagot, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, Éric Fleury, et al. 2019. What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.
- Matej Martinc, Senja Pollak, and Marko Robnik-Sikonja. 2019. Supervised and unsupervised neural approaches to text readability. *Computing Research Repository*, *arXiv:1503.06733. Version 2*.
- Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

- Manuela Sanguinetti and Cristina Bosco. 2015. Part-TUT: The turin university parallel treebank. In Roberto Basili et al., editor, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, page 51–69. Springer.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018)*.
- M. Straka, J. Hajic, and J. Strakova. 2016. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2019. Universal dependencies 2.5. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*.

Italian Transformers Under the Linguistic Lens

Alessio Miaschi^{♣, *}, Gabriele Sarti^{♣, ♦, *}, Dominique Brunato^{*},
Felice Dell’Orletta^{*}, Giulia Venturi^{*}

[♣]Department of Computer Science, University of Pisa

[♦]Department of Mathematics and Geosciences, University of Trieste

[♦]International School for Advanced Studies (SISSA), Trieste

^{*}Istituto di Linguistica Computazionale “Antonio Zampolli”, ItaliaNLP Lab, Pisa

alessio.miaschi@phd.unipi.it, gsarti@sissa.it,

{name.surname}@ilc.cnr.it

Abstract

In this paper we present an in-depth investigation of the linguistic knowledge encoded by the transformer models currently available for the Italian language. In particular, we investigate whether and how using different architectures of probing models affects the performance of Italian transformers in encoding a wide spectrum of linguistic features. Moreover, we explore how this implicit knowledge varies according to different textual genres.

1 Introduction and Background

In the last few years, the study of Neural Language Models (NLMs) and their representations has become a key research area in the NLP community. Several methods have been devised to obtain meaningful explanations regarding the linguistic information encoded in NLMs (Belinkov and Glass, 2019). The most common approach is based on the development of *probes*, i.e. supervised models trained to predict a variety of language properties using the contextual word/sentence embeddings of a pre-trained model (Conneau et al., 2018; Zhang and Bowman, 2018; Miaschi and Dell’Orletta, 2020). This approach demonstrated that NLMs representations encode linguistic knowledge in a hierarchical manner (Belinkov et al., 2017; Blevins et al., 2018; Tenney et al., 2019b), and can even support the extraction of dependency parse trees (Hewitt and Manning, 2019). Jawahar et al. (2019) investigated the representations learned by BERT (Devlin et al., 2019), one of the most prominent NLM, across its layers, showing that lower ones are usually better for capturing surface features, while embeddings

from higher layers are better for syntactic and semantic properties. Using a suite of probing tasks, Tenney et al. (2019a) deeply explore this behavior showing that the linguistic knowledge encoded by BERT through its 12/24 layers follows the traditional NLP pipeline.

While the vast majority of this research focused on English contextual representations, relatively little work has been done to understand the inner workings of non-English models. The study by de Vries et al. (2020) represents an exception in this context: authors apply the probing task approach to compare the linguistic competence encoded by a Dutch BERT-based model and multilingual BERT (mBERT), showing that earlier layers of mBERT are consistently more informative than earlier layers of the monolingual model. The survey by Nozza et al. (2020) also provides a comparative study of mBERT and language-specific BERT models but focused on the performance that each model obtains after training on several specific downstream tasks.

In this paper, we adopt a task-agnostic perspective to carry out an in-depth investigation of the linguistic knowledge implicitly encoded by 6 Italian monolingual models and multilingual BERT. We define a broad set of probing tasks, each corresponding to a specific property of sentence structure. We then compare the average performance reached by each model in predicting the feature value, evaluating the results obtained by models using their layer-wise sentence-level representations. A further comparative perspective, which to our knowledge is still rather under-investigated, concerns the study of how the architecture of the probing model itself influences probing scores. To address this point, for each model, we perform the same suite of probing tasks using both a linear SVR and a multilayer perceptron (MLP), and compare whether and how each probing task’s resolution is affected by the two architectures.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Since all experiments were carried out on different sections of Italian Universal Dependency Treebank (Nivre et al., 2016), we were also able to investigate how linguistic knowledge of NLMs varies according to different textual genres.

Contributions To the best of our knowledge, this is the first study aimed at comparing the linguistic knowledge encoded in the representations of multiple non-English pre-trained transformer models. In particular: (i) we compare the probing performances of 6 Italian NLMs spanning three models over multiple linguistic feature categories; (ii) we investigate whether and how using different architectures of probing models affects the performance of transformers in encoding specific features; and (iii) we show how the implicit knowledge learned by these models differs across textual genres.

2 Approach

To inspect the inner knowledge of language encoded by Italian Transformers, we relied on a suite of 82 probing tasks, each of which corresponds to predicting the value of a corresponding feature modeling a specific property of the sentence. We designed two sets of experiments. The first one consists in comparing the linguistic knowledge encoded by the Italian Transformers and evaluating the best probing model for inferring such knowledge from the NLMs. We compared the results obtained with two simple probing models, a linear SVR and a multilayer perceptron (MLP), which take as input layer-wise sentence-level representations extracted from the Italian models. These representations are produced for each sentence of different sections of the Italian Universal Dependency Treebank (IUDT), version 2.5 (Zeman et al., 2019), and used to predict the actual value of each probing feature. In the second set of experiments, we evaluated how the Italian models’ linguistic knowledge differs across textual genres and varieties, considering different IUDT sections.

2.1 Models and Data

We relied on 7 pre-trained Italian Transformers models. Models statistics are reported in Table 1.

¹<https://github.com/dbmdz/berts>

²Polignano et al. (2019)

³<https://github.com/idb-ita/GilBERTo>

⁴<https://github.com/musixmatchresearch/umberto>

⁵De Mattei et al. (2020)

Name	Training data
BERT Architecture	
Multilingual-BERT	Wikipedia
BERT-base-italian ¹	Wikipedia + OPUS (13GB) (Tiedemann and Nygaard, 2004)
AlBERTo ²	TWITA (191GB) (Basile et al., 2018)
RoBERTa Architecture	
GilBERTo ³	OSCAR (71GB) (Suárez et al., 2019)
UmBERTo-Commoncrawl	OSCAR (69GB)
UmBERTo-Wikipedia ⁴	Wikipedia (7GB)
GPT-2 Architecture	
GePpeTto ⁵	Wikipedia + ItWAC (14GB) (Baroni et al., 2009)

Table 1: NLMs used in the experiments.

Short Name	Types of texts	# sent
ParTUT (Sanguinetti and Bosco, 2015)	Multi-genre	2,090
VIT (Delmonte et al., 2007)	Multi-genre	10,087
ISDT (Bosco et al., 2013)	Multi-genre	14,167
ISDT_tanl	Newswire	4,043
ISDT_tut	Legal/Newswire/Wiki	3,802
ISDT_quest	Interrogative sentences	2,162
ISDT_2parole	Simplified Italian news	1,421
ISDT_europarl	EU Parliament acts	497
PoSTWITA (Sanguinetti et al., 2018)	Tweets	6,713
TWITTIRÒ (Cignarella et al., 2019)	Ironie Tweets	1,424
Total		35,481

Table 2: Sections of the Italian Universal Dependency Treebank (IUDT).

Sentence level representations were computed performing a *Mean-pooling* operation over the word embeddings provided by the models.

NLM’s linguistic competences are probed against five IUDT sections including texts representative of different textual varieties and genres. As shown in the overview in Table 2, we also distinguish the whole ISDT into different sub-corpora according to the specific language variety they represent, e.g. transcription of spontaneous speech (*ISDT_europarl*), questions (*ISDT_quest*) or simplified language (*ISDT_2parole*).

2.2 Probing features

The set of probing tasks consists of predicting the value of a specific linguistic feature automatically extracted from each POS tagged and dependency parsed sentence of the IUDT datasets.

The set of features is based on the ones described in Brunato et al. (2020) and are acquired from raw, morpho-syntactic and syntactic levels of annotation and can be categorised in 9 groups corresponding to different linguistic phenomena. As shown in Table 3, these features model linguistic phenomena ranging from raw text one, to morpho-syntactic information and inflectional properties of verbs, to more complex aspects of sentence struc-

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distribution of UD and language-specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

Table 3: Probing Features used in the experiments.

ture capturing global and local properties of the whole parsed tree and of specific subtrees, such as the order of subjects and objects with respect to the verb, the distribution of UD syntactic relations, also including features referring to the use of subordination and to the structure of verbal predicates.

All these features have been shown to play a highly predictive role when leveraged by traditional learning models on a variety of classification problems, covering different aspects of stylistometric and complexity analysis. In addition, in their recent work, Miaschi et al. (2020) showed that these features can be effectively used to profile the knowledge encoded in the language representations of a pretrained NLM, specifically the English Bert, and how it changes across layers. Since these features are based on the UD formalism, which guarantees the comparative encoding of language phenomena between the two languages (Nivre, 2015), we focused on the same set to investigate the linguistic knowledge of Italian transformers.

3 Results

We first investigate which is the best architecture for probing the linguistic knowledge encoded by the Italian Transformers. Since many of our probing features are strongly related to sentence length, we compared the two probing models' results with the ones obtained by a baseline corresponding to a LinearSVR model trained using

Groups	LinearSVR	MLP	Baseline
RawText	0.84	0.80	0.50
Vocabulary	0.70	0.34	0.19
POS	0.69	0.68	0.03
VerbInflection	0.50	0.61	0.03
VerbPredicate	0.32	0.43	0.08
TreeStructure	0.61	0.64	0.40
Order	0.46	0.55	0.06
SyntacticDep	0.65	0.74	0.04
Subord	0.49	0.60	0.16
AllFeatures	0.60	0.64	0.10

Table 4: Average R^2 scores for all the NLMs obtained with the LinearSVR and the MLP probing models. Baseline scores are also reported.

only sentence length as input feature. Table 4 reports average R^2 results⁶ for all the 7 NLMs obtained with the LinearSVR and the MLP probing models, along with baseline scores. The MLP probe is a three-layer feedforward network with ReLU activations and was selected to investigate the presence of nonlinear relations in representations, which could hamper the probing performance of the LinearSVM probe, but would be highlighted by a sharp difference between MLP and LinearSVM performances. As a first remark, we notice that both probing models outperform the baseline. This proves that all NLMs encode a spectrum of phenomena that, although related to sentence length, require a more sophisticated linguistic knowledge to be accurately predicted. Best scores are obtained with the MLP model, which achieved higher R^2 scores especially for features grouping more complex syntactic phenomena (e.g. *TreeStructure*, *SyntacticDep*). Interestingly enough, the LinearSVR model outperforms the MLP by more than .30 R^2 points when predicting features related to vocabulary richness (*Vocabulary*).

In order to ensure that our probes are actually showing the linguistic generalization abilities of the NLMs rather than learning the linguistic tasks, we also tested the probing models using the *control task* approach devised in Hewitt and Liang (2019). We produced a control version of the IUDT corpus by randomly shuffling the linguistic features assigned to each sentence and performed the same probing tasks with the two probing classifiers for all NLMs representations. The correla-

⁶The Coefficient of determination (R^2) is a statistical measure of how close the data are to the fitted regression line and corresponds to the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

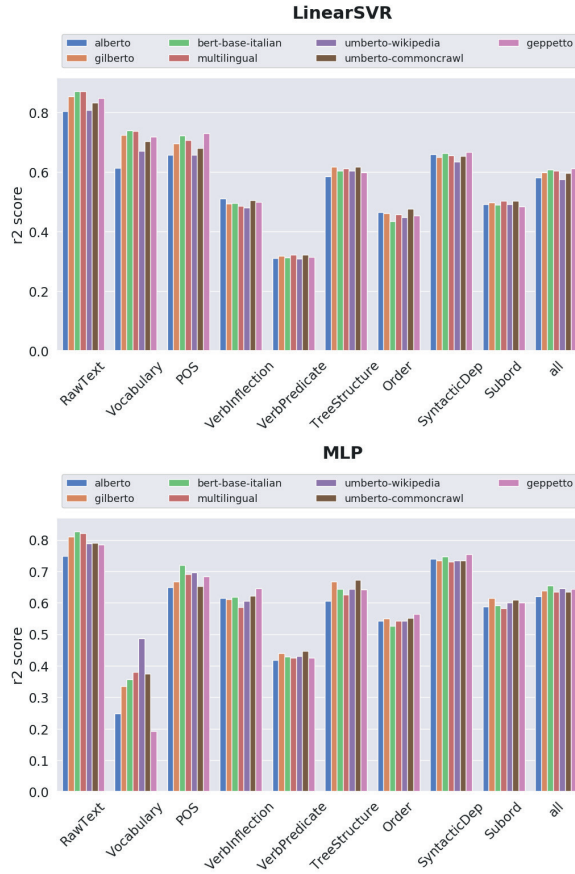


Figure 1: Average R^2 scores obtained by each NLM with the two probing models.

tion and R^2 scores between regressors’ predictions and shuffled scores were low (< 0.05) and comparable for both the SVR and the MLP. These results support the claim that NLMs representations encode information closely related to linguistic competence and that our probing models are not relying on spurious signals unrelated to our linguistic properties to solve the regression task.

To investigate how each transformer encodes the linguistic knowledge, we report in Figure 1 average R^2 scores obtained with the two probing models for all the 7 NLMs. As we can notice, the seven transformers achieve quite similar results when considering all features as a whole, although BERT-base-italian has the best overall performance (0.65 for *all* features). The same did not hold when we analyzed their performances in terms of R^2 scores for the different previously described groups of features. For instance, we can notice that, for both the probing models, features related to the distribution of syntactic relations (*SyntacticDep*) are better predicted by GePpeTto,

while GiLBERTO and UmBERTo-Commoncrawl are the best ones in the prediction of tree structure properties. Differences hold for what regards competencies related to vocabulary richness (*Vocabulary*): while UmBERTo-Wikipedia extensively outperforms all the other transformers using the MLP model, the best transformer is BERT-base-italian when these competences are probed with the LinearSVR model.

Similar trends can be observed in Figure 2, where we report how the linguistic knowledge encoded by the 7 NLMs evolves across layers according to the two probing models. Regardless of the architectures, for all transformers, raw text features (*RawText*) are mainly encoded in the first layers, while the knowledge about the order of subject/object (*Order*) and the use of subordination (*Subord*) increases consistently across layers and specifically in the first ones. Contrarily to what was observed by de Vries et al. (2020), mBERT’s linguistic knowledge is not encoded systematically earlier than in monolingual transformers. This perspective of analysis also reveals other differences among the considered transformers: e.g. even though GePpeTto has a lower average competence on verb inflection (see Figure 1), it achieves the highest scores in the middle layers. Focusing instead on differences between layerwise scores obtained by the two probing models, we can clearly notice that the encoding of linguistic knowledge shows a quite rough trend for what concerns the results obtained with the MLP. This is particularly the case of features belonging to the vocabulary, POS and tree structure groups.

Finally, we inspected whether the overall linguistic competence encoded in the contextual representations of each model changes according to the type of texts in the different IUDT sections we considered. As we could expect, the results reported in Figure 3 show that all transformers achieve lower performance when they have to predict the value of features extracted from treebanks representative of social media language (PoSTWITA and TWITTIRÒ). Quite surprisingly, it is also the case of AIBERTO which is trained on Twitter data. A possible explanation is that, although PoSTWITA and TWITTIRÒ contain sentences representative of Twitter language, these sentences are still quite close to the Italian standard language, in order to be compliant with the UD morpho-syntactic and syntactic annota-

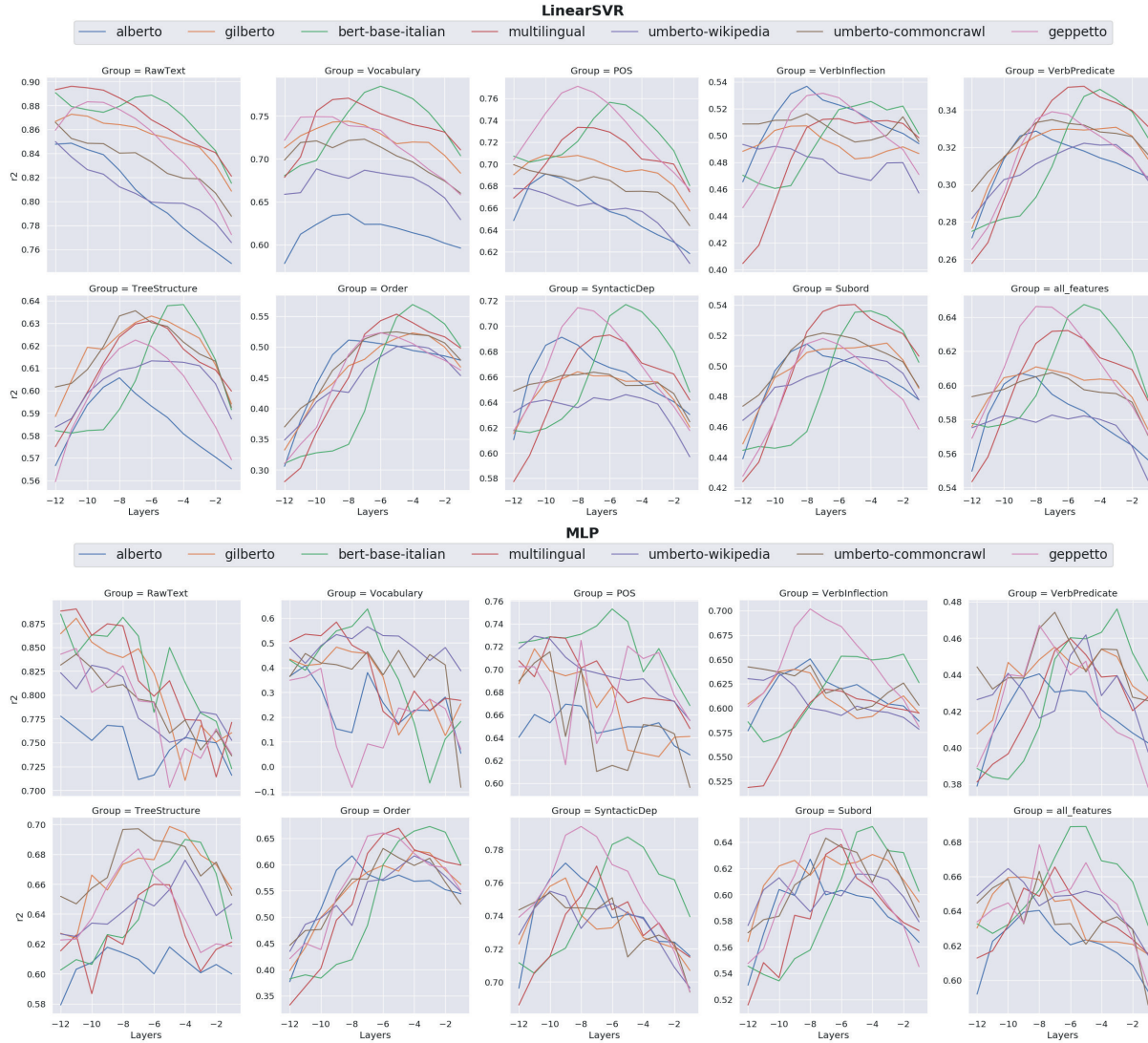


Figure 2: Average layerwise R^2 scores obtained with the LinearSVR (*top*) and the MLP (*bottom*) using the internal representations of the 7 NLMs.

tion schema. On the contrary, ALBERTo’s training set is derived from Twitter’s official streaming API that included all possible typologies of sentences. However, bert-base-italian is slightly less affected by the non-standard linguistic peculiarities of this genre. Similarly to what is observed for the whole Italian dataset (see Figure 1), this model also reaches the highest performance in almost all different IUDT sections, except for the one containing interrogative sentences (*isdt_quest*). Interestingly, this type of sentence is hardly mastered by all models. This is possible due to the fact that interrogative sentences are more likely to display a less canonical distribution of morphosyntactic and syntactic phenomena, hence being more difficult to encode effectively.

4 Conclusion

In this paper we presented an in-depth comparative investigation of the linguistic knowledge encoded in the Italian transformer models. Relying on a suite of more than 80 probing features and testing our approach with two different probing models, we showed that MLP is the best model for inferring the amount of information implicitly encoded in the NLMs representations. We also observed that BERT-base-italian achieved best scores in average, but the linguistic generalization abilities of the examined transformers vary according to specific groups of linguistic phenomena and across layers. Finally, we examined how the linguistic knowledge learned by the NLMs is affected by the distinct textual varieties available in Italian tree-

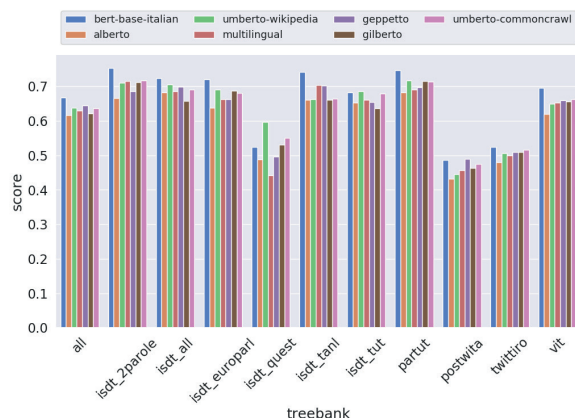


Figure 3: Average LinearSVM R^2 score considering all the UD Italian sentences (*all*) and according to the 10 treebanks previously described.

banks showing, for instance, that social media language represents a harder domain for all models.

We are currently investigating if the linguistic knowledge encoded by a NLM positively affects the resolution of downstream tasks, as already suggested by the recent work by Miaschi et al. (2020) for English. This connection, which is still rather investigated, can improve our understanding of how such models make their decisions.

References

- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6. CEUR-WS.
- Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.
- Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.
- Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7147–7153, Marseille, France, May. European Language Resources Association.
- Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An italian twitter treebank in universal dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $\$&!#*$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.
- Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What’s so special about bert’s layers? a closer look at the nlp pipeline in monolingual and multilingual models. *arXiv preprint arXiv:2004.06499*.
- Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT - Venice Italian Treebank: Syntactic and quantitative features. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong

- Kong, China, November. Association for Computational Linguistics.
- John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.
- Ganesh Jawahar, Benoît Sagot, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, Éric Fleury, et al. 2019. What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.
- Alessio Miaschi and Felice Dell’Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online, July. Association for Computational Linguistics.
- Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666.
- Joakim Nivre. 2015. Towards a universal grammar for natural language processing. *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–16.
- Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*.
- Manuela Sanguinetti and Cristina Bosco. 2015. Part-TUT: The turin university parallel treebank. In Roberto Basili et al., editor, *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, page 51–69. Springer.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in universal dependencies. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC 2018)*.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.
- Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: <http://logos.uio.no/opus>. Citeseer.
- Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2019. Universal dependencies 2.5. In *LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*.
- Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

BERTino: an Italian DistilBERT model

Matteo Muffo

Indigo.ai
Via Torino 61, Milano
matteo@indigo.ai

Enrico Bertino

Indigo.ai
Via Torino 61, Milano
e@indigo.ai

Abstract

English.¹ The recent introduction of Transformers language representation models allowed great improvements in many natural language processing (NLP) tasks. However, if on one hand the performances achieved by this kind of architectures are surprising, on the other their usability is limited by the high number of parameters which constitute their network, resulting in high computational and memory demands. In this work we present BERTino, a DistilBERT model which proposes to be the first lightweight alternative to the BERT architecture specific for the Italian language. We evaluated BERTino on the Italian ISDT, Italian ParTUT, Italian WikiNER and multiclass classification tasks, obtaining F1 scores comparable to those obtained by a $BERT_{BASE}$ with a remarkable improvement in training and inference speed.

Italiano. La recente introduzione dei Transformers come modelli di rappresentazione del linguaggio naturale ha permesso grandi avanzamenti sullo stato dell'arte in molte applicazioni di Natural Language Processing (NLP). Tuttavia, se da una parte i risultati raggiunti da queste architetture sono sorprendenti, dall'altra la loro fruibilità è limitata dall'elevato numero di parametri che costituiscono la loro architettura, con conseguenti elevate esigenze computazionali e di memoria. In questo lavoro presentiamo BERTino, un modello DistilBERT che è la prima

alternativa *leggera* all'architettura BERT specifica per la lingua italiana. Abbiamo valutato BERTino sui task ISDT italiano, ParTUT italiano, WikiNER italiano e classificazione multiclasse, ottenendo punteggi F1 paragonabili a quelli ottenuti da un modello $BERT_{BASE}$ con un notevole miglioramento nella velocità di addestramento e inferenza.

1 Introduction

In recent years the introduction of Transformers language models allowed great improvements in many natural language processing (NLP) tasks. Among Transformer language models, BERT (Devlin et al., 2018) affirmed itself as an high-performing and flexible alternative, being able to transfer knowledge from general tasks to downstream ones thanks to the pretraining-finetuning approach. The context-dependent text representations provided by this model demonstrated to be a richer source of information when compared to static textual embeddings such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2016) or Sent2Vec (Pagliardini et al., 2018). However, despite the substantial improvements brought by BERT in the NLP field, the high number of parameters that constitute its network makes its usage prohibitive in resource-limited devices, both at training and inference time, and with a non-negligible environmental impact. To address the aforementioned problem, recent research proposes several approaches to reduce the size of the BERT network, such as DistilBERT (Sanh et al., 2019), MobileBERT (Sun et al., 2020) or pruning (Gordon et al., 2020; McCarley et al., 2019).

The experiments conducted in Virtanen et al. (2019), de Vries et al. (2019) and Martin et al. (2020) demonstrate that monolingual BERT models outperform the same multilingual BERT ar-

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

chitecture (Devlin et al., 2018), justifying the effort for pre-training Transformer models required for specific languages. In this work we present **BERTino**, a DistilBERT model pre-trained on a large Italian corpus. This model proposes to be the first general-domain, lightweight alternative to BERT specific for the Italian language. We evaluate BERTino on two Part Of Speech tagging tasks, Italian ISDT (Bosco et al., 2000) and Italian PartUT (Sanguinetti and Bosco, 2015), on the Italian WikiNER (Nothman et al., 2012) Named Entity Recognition task and on a multi-class sentence classification. Comparing the scores obtained by BERTino, its teacher model and GilBERTo, the first obtains performances comparable to the other two architectures while sensibly decreasing the fine-tuning and evaluation time. In Section 2 we discuss the related works with a focus on DistilBERT, in Section 3 we describe the corpus and the pre-train followed by the results in Section 4.

2 Related work

In this section we will give a brief outline of the inner workings for Transformers, then we overview some lightweight alternatives to BERT.

The introduction of Transformer blocks (Vaswani et al., 2017) in language representation models is a keystone in recent NLP. The attention mechanism adopted by the Transformer encoder allows to provide contextualized representations of words, which proved to be a richer source of information than static word embeddings. Attention mechanism processes all words in an input sentence simultaneously, allowing parallelization of computations. This is a non-negligible improvement with respect to models like ELMo (Peters et al., 2018), which aim to provide contextualized text representations using a bidirectional LSTM network, processing each word sequentially.

Among language models that adopt Transformer technology, BERT (Devlin et al., 2018) affirmed itself as a flexible and powerful alternative, being able to establish new state-of-the-art for 11 NLP tasks at the time of publication. In its base version, this model adopts an hidden size of 768 and is composed of 12 layers (Transformer blocks), each of these involving 12 attention heads, for a total of 110 millions of parameters. As outlined in Section 1, the high number of parameters constituting BERT’s network can result prohibitive for deployment in resource-

limited devices and the computational effort is not negligible. For this reason, great effort has been devoted by researchers in order to propose smaller but valid alternatives to the base version of BERT. Gordon et al. (2020) studies how weight pruning affects the performances of BERT, concluding that a low level of pruning (30-40% of weights) marginally affects the natural language understanding capabilities of the network.

McCarley et al. (2019) conducts a similar study on BERT weight pruning, but applied to the Question Answering downstream task specifically.

Sanh et al. (2019) propose DistilBERT, a smaller BERT architecture which is trained using the knowledge distillation technique (Hinton et al., 2015). Since the model that we propose relies on this training technique, we propose a brief description of knowledge distillation in section 2.1. DistilBERT leverages the inductive biases learned by larger models during pre-training using a triple loss combining language modeling, distillation and cosine-distance losses. DistilBERT architecture counts 40% less parameters but is able to retain 97% of natural language understanding performances with respect to the teacher model, while being 60% faster.

Sun et al. (2020) propose MobileBERT, a compressed BERT model which aims to reduce the hidden size instead of the depth of the network. As DistilBERT, MobileBERT uses knowledge distillation during pre-training but adopts a $BERT_{LARGE}$ model with inverted bottleneck as teacher.

2.1 Knowledge distillation

Knowledge distillation (Hinton et al., 2015) is a training technique that leverages the outputs of a big network (called *teacher*) to train a smaller network (the *student*). In general, in the context of supervised learning, a classifier is trained in such a way that the output probability distribution that it provides is as similar as possible to the one-hot vector representing the gold label, by minimizing the cross-entropy loss between the two. By receiving a one-hot vector as learning signal, a model evaluated on the training set will provide an output distribution with a near-one value in correspondence of the right class, and all near-zero values for other classes. Some of the near-zero probabilities, however, are larger than the others and are the result of the generalization capabili-

ties of the model. The idea of knowledge distillation is to substitute the usual one-hot vector representing gold labels with the output distribution of the teacher model in the computation of the cross-entropy loss, in order to leverage the information contained in the near-zero values of the teacher’s output distribution. Formally, the knowledge distillation loss is computed as:

$$\mathcal{L}_{KD} = \sum_i t_i * \log(s_i) \quad (1)$$

with t_i being the output distribution of the teacher model relative to the i^{th} observation, and s_i being the output distribution of the student model relative to the i^{th} observation.

3 BERTino

As outlined in section 1, we propose in this work BERTino, a DistilBERT model pre-trained on a general-domain Italian corpus. As for BERT-like architectures, BERTino is task-agnostic and can be fine-tuned for every downstream task. In this section we will report details relative to the pre-training that we conducted.

3.1 Corpus

The corpus that we used to pre-train BERTino is the union of PAISA (Lyding et al., 2014) and ItWaC (Baroni et al., 2009), two general-domain Italian corpora scraped from the web. While the former is made up of short sentences, the latter includes a considerable amount of long sentences. Since our model can receive input sequences of at most 512 tokens, as for BERT architectures, we decided to apply a pre-processing scheme to the ItWaC corpus. We split the sentences with more than 400 words into sub-sentences, using fixed points to create chunks that keep the semantic sense of a sentence. In this way, most of the long sentences contained in ItWaC are split into sub-sentences containing less than 512 tokens. A certain number of the final sentences still contain more than 512 tokens and they will be useful for training the parameters relative to the last entries of the network.

The PAISA corpus counts 7.5 million sentences and 223.5 million words. The ItWaC corpus counts 6.5 million sentences and 1.6 billion words after preprocessing. Our final corpus counts 14 million sentences and 1.9 billion words for a total of 12GB of text.

3.2 Pre-training

Teacher model The teacher model that we selected to perform knowledge distillation during the pre-training of BERTino is *dbmdz/bert-base-italian-xxl-uncased*, made by *Bavarian State Library*². We chose this model because it is the Italian $BERT_{BASE}$ model trained on the biggest corpus (81 GB of text), up to our knowledge. Following Sanh et al. (2019), we initialized the weights of our student model by taking one layer out of two from the teacher model.

Loss function We report the loss function used to pre-train BERTino:

$$\mathcal{L} = 0.45\mathcal{L}_{KD} + 0.45\mathcal{L}_{MLM} + 0.1\mathcal{L}_{COS} \quad (2)$$

with \mathcal{L}_{KD} being the knowledge distillation loss as described in equation 1, \mathcal{L}_{MLM} being the masked language modeling loss and \mathcal{L}_{COS} being the cosine embedding loss. Sanh et al. (2019) describe the cosine embedding loss useful to “align the directions of the student and teacher hidden states vectors”. When choosing the weights of the three loss functions, we wanted our model to learn from the teacher and by itself in an equal way, so we set the same weights for both \mathcal{L}_{KD} and \mathcal{L}_{MLM} . Moreover, we considered the alignment of student and teacher hidden states vectors marginal for our objective, setting \mathcal{L}_{COS} as 10% of the total loss.

Architecture The architecture of BERTino is the same as in DistilBERT. Our model adopts an hidden size of 768 and is composed of 6 layers (Transformer blocks), each of which involving 12 attention heads. In this way BERTino’s network results to have half the layers present in the $BERT_{BASE}$ architecture.

Training details To pre-train BERTino we used a batch size of 6 and an initial learning rate of 5×10^{-4} , adopting Adam (Kingma and Ba, 2014) as optimization algorithm. We chose 6 as batch size due to the limited computational resources available. Results described in section 4 demonstrate that the small batch size that we adopted is sufficient to obtain a valid pre-trained model. We trained our model on 4 Tesla K80 GPUs for 3 epochs, requiring 45 days of computation in total. For some aspects of the training, we relied on the Huggingface Transformers repository (Wolf et al., 2019).

²<https://github.com/dbmdz/berts>

4 Results

We tested the performances of BERTino on benchmark datasets: the Italian ISDT (Bosco et al., 2000) and Italian ParTUT (Sanguinetti and Bosco, 2015) Part Of Speech tagging tasks, and the Italian WikiNER (Nothman et al., 2012) Named Entity Recognition task. To complete the evaluation of the model, we also tested it on a multi-class sentence classification task. In particular, we focused on intent detection, a task specific to the context of Dialogue Systems, creating a novel Italian dataset which is freely available at our repository³. The dataset that we propose collects 2786 real-world questions (2228 for training and 558 for testing) submitted to a digital conversational agent. The total number of classes in the dataset is 139.

For the first two tasks mentioned, we fine-tuned our model on the training set for 4 epochs with a batch size of 32 and a learning rate of 5×10^{-5} , for the NER task we performed 5-fold splitting of the dataset and fine-tuned BERTino for 2 epochs per fold with a batch size of 32 and a learning rate of 5×10^{-5} , while for the multi-class classification task we fine-tuned our model for 14 epochs on the training set with a batch size of 32 and a learning rate of 5×10^{-5} . To compare the results obtained, we fine-tuned the teacher model and a GilBERTo model⁴ on the same tasks with the same hyperparameters. Tables 1, 2, 3 and 4 collect the F1 scores gathered in these experiments together with fine-tuning and evaluation time. All the scores reported represent the average computed over three different runs. Results show that the teacher model slightly outperforms BERTino, with an increase of the F1 score of 0,29%, 5,15%, 1,37% and 1,88% over the tasks analysed. However BERTino results to be a sensibly faster network with respect to the teacher model and GilBERTo, taking almost half of the time to perform both fine-tuning and evaluation. We can conclude from the last observation that BERTino is able to retain most of the natural language understanding capabilities of the teacher model, even with a much smaller architecture.

5 Conclusions

In this work we presented BERTino, a DistilBERT model which aims to be the first lightweight alternative to BERT specific for the Italian language. Our model has been trained on a general-domain

corpus and can then be finetuned with good performances on a wide range of tasks like its larger counterparts. BERTino showed comparable performances with respect to both the teacher model and GilBERTo in the Italian ISDT, Italian ParTUT, Italian WikiNER and multi-class sentence classification tasks while taking almost half of the time to fine-tune, demonstrating to be a valid lightweight alternative to *BERT_{BASE}* models for the Italian language.

³<https://github.com/indigo-ai/BERTino>

⁴Available at <https://github.com/idb-ita/GilBERTo>

Italian ISDT			
Model	F1 score	Fine-tuning time	Evaluation time
BERTino	0,9800	9'10"	3"
Teacher model	0,9829	16'32"	6"
GilBERTo	0,9804	18'11"	5"

Table 1: F1 scores obtained by BERTino and the teacher model in the Italian ISDT task.

Italian ParTUT			
Model	F1 score	Fine-tuning time	Evaluation time
BERTino	0,9193	1'19"	1"
Teacher model	0,9708	2'19"	1"
GilBERTo	0,9621	2'21"	1"

Table 2: F1 scores obtained by BERTino and the teacher model in the Italian ParTUT task.

Italian WikiNER			
Model	F1 score	Fine-tuning time	Evaluation time
BERTino	0,9039	38'3"	3'2"
Teacher model	0,9176	67'2"	5'21"
GilBERTo	0,9136	66'33"	5'9"

Table 3: F1 scores obtained by BERTino and the teacher model in the Italian WikiNER task. The results reported are the average of the scores obtained in each of the 5 folds.

Multi-class sentence classification			
Model	F1 score	Fine-tuning time	Evaluation time
BERTino	0,7766	5'4"	6"
Teacher model	0,7954	9'48"	10"
GilBERTo	0,7381	10'0"	10"

Table 4: F1 scores obtained by BERTino and the teacher model in the multi-class sentence classification task.

References

- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.
- Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, Athens, Greece, May. European Language Resources Association (ELRA).
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing bert: Studying the effects of weight pruning on transfer learning.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.
- Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell’Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, April. Association for Computational Linguistics.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- J. S. McCarley, Rishav Chakravarti, and Avirup Sil. 2019. Structured pruning of a bert-based question answering model.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
- Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.
- Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.
- Manuela Sanguinetti and Cristina Bosco, 2015. *Part-TUT: The Turin University Parallel Treebank*, pages 51–69. Springer International Publishing, Cham.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

ItaGLAM: A corpus of Cultural Communication on Twitter during the Pandemic

Gennaro Nolano, Carola Carlino, Maria Pia di Buono, Johanna Monti

UniOr NLP Research Group
"L'Orientale" University of Naples
Italy

{gnolano, ccarlino, mpdibuono, jmonti}@unior.it

Abstract

This paper describes the compilation and annotation of **ItaGLAM**, a corpus of tweets written by Italian Galleries, Libraries, Archives and Museums (GLAMs) during the lockdown period in Italy due to the COVID-19 pandemic. ItaGLAM has been annotated with a set of labels which may be useful to identify different types of communication. Furthermore, the collected data have been used to train a set of classifiers.

The results are analyzed to evaluate the information flow between GLAM and users and to analyze cultural communication on the Web.

1 Introduction

Over the last years, Social Networks have become one of the most popular platforms for sharing experiences and opinions through the use of simple strings of text (Zhao and Rosson, 2009). Indeed, this way of communicating has become an essential interaction tool, not only among private users, but also among companies to engage with their audience and to promote their brands (Alturas and Oliveira, 2016).

The use of social networks has also been adopted by museums, that, over time, have changed their way of communicating with their audience². In particular, in regards to the GLAM sector, a new trend has been observed in recent years: the use of

the web as a way to create and foster an online community (Langa, 2014; Allen-Greil and MacArthur, 2010).

While during the first decade of this century museum professionals considered the exhibition of collections on social networks (Laws, 2015) as 'excessive', nowadays the use of these platforms has become the norm. As Amanatidis et al. (2020) pointed out in their study about the use of social networks (and in particular Instagram) by museums in the Greek culture scene: 'social media has become a key factor in the way that cultural organizations communicate with their public in supporting the marketing of performing art organizations'.

Such centrality makes the Social network a potentially effective means that allows GLAMs to reach a wide and heterogeneous audience and to adapt to it. Therefore, we believe that the analysis of the cultural communication implies an analysis of how cultural corporations interact with the audience through social networks.

After considering the most used social networks (namely Facebook, Instagram, Twitter) in the cultural sector, we have decided to focus our research on the use of Twitter, which has already been proven to be a solid basis to analyze institutional communication, as Preoțiuc-Pietro et al. (2015) have highlighted.

Therefore, the main aim of our research is to investigate how Italian GLAMs have extraordinarily (Giraud, 2020) interacted with their audience during the lockdown in Italy due to the COVID-19 pandemic (NEMO - Network of European Museum Organisations, 2020), i.e. in the period from the 8th of March to the 5th of May 2020 (as per *DPCM March 11 2020*).

Over this time many cultural initiatives have been launched with the aim of strengthening the dialogue with the audience and make sure that, despite the impossibility of any kind of physical access, the connection between GLAMs and their

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.osservatori.net/it/ricerche/comunicati/-stampa/laumento/-del/-livello/-di/-interesse/-per/-le/-attivita/-online/-dei/-musei/-incentivato/-dal/-covid/-19/-e/-gli/-investimenti/-per/-migliorare/-i/-servizi/-offerti>

visitors would not be interrupted³.

It has been observed that during the aforementioned period, while GLAMs institutions have drastically increased their use of Facebook, Instagram and Twitter, the latter one was the only Social Network for which an increase in the interaction audience-institution has been registered (Politecnico di Milano, 2020).

The study of communicative intents by GLAMs through Social Networks in the Italian language is still novel and, as such, best practices and tools to use still need to be tested and honed. In particular, there is still the need for an annotated corpus and a classifier that can be used on large amounts of data.

Despite the time frame taken into account is relatively short (covering 58 days in total), we think that investigating how Italian GLAMs used the web when it was the only form of communication at their disposal, represents a good training ground to test our practices and to train and evaluate different kinds of classifiers useful also in future works.

The paper is organized as follows: Section 2 describes the related works in the analysis of communication on Twitter by cultural institutions. Section 3 introduces the methodology used in this analysis: namely, it describes the creation of the corpus, the creation and use of the annotation set, and the training and evaluation of different classifiers. Finally, in Section 4 we explain the results of the research.

2 Related Work

The large amount of data available on Twitter makes this platform ideal for several studies. As such, during the years tweets have been used in several research projects regarding disaster response (Zahra et al., 2019), content classification (Dann, 2010; Stvilia and Gibradze, 2014) and, in particular, sentiment analysis (O'Connor et al., 2010; Gamallo and Garcia, 2014; Talbot et al., 2015). Despite these efforts, only a few studies have focused on the classification of communicative intents of organizations and institutions on Twitter, like Lovejoy and Saxton (2012) and Foucault and Courtin (2016), who focused on French tweets written during the MuseumWeek event.

Similar kinds of study can be found in researches dealing with Italian tweets, with several contributions dealing with sentiment analysis (Basile and

Nissim, 2013; Cimino et al., 2014) and automatic misogyny identification (Anzovino et al., 2018). To the best of our knowledge, no work has been done so far on communicative intent classification for Italian tweets.

3 Methodology

The task of tweet classification has turned out to be rather challenging for various reasons, many inherent to the platform itself. First and foremost, tweets are very short texts (with the maximum length of 280 characters), and with an average token count of 16.80 in our corpus.

Secondly, it is not unusual to find tweets composed only of hashtags, or URLs. While URLs by themselves are rarely if ever useful in a classification task, hashtags could represent a source of information only if they are used according to their original communicative intent or to the initiative to which they are related.

In the following subsections we describe how:

- the corpus was created;
- the annotation set has been chosen and then applied;
- the classifiers have been trained and tested.

3.1 Dataset

Because of the COVID-19 outbreak, the Italian Government (as many others around the world) imposed a lockdown policy, which lasted from the 8th March to the 5th May 2020 (58 days in total as per *DPCM March 11 2020*).

During this period of time, museums and art galleries adopted several strategies to continue engaging with their audience in order to maintain the communication alive, and to grant access to digital cultural heritage media. As already mentioned in Section 1, they increased the scope of their communication on the main social platforms, i.e. Facebook, Twitter and Instagram.

In this context, the focus of our analysis is the use of Twitter. The communication on Twitter is characterised by the use of certain *hashtags*, which have been used by GLAMs to propose several types of initiatives to their audience. Initially, the set of hashtags we used was made up of 33 hashtags promoted and used by Italian GLAMs and Italy's Ministry of Cultural Heritage and Activities (Italian: *Ministero per i Beni e le Attività Culturali e per il Turismo - MiBACT*), and

³<https://icom.museum/en/news/how-to-reach-and-engage-your-public-remotely/>

selected on the basis of their popularity according to the Twitter *trend topics* (TT)⁴.

Among these hashtags, #museitaliani (and its graphic variation #museiitaliani) is the only one already existing before the pandemic, and subsequently adapted by museums for the initiatives proposed during the pandemic; while others, such as #artyouready and #emptymuseum have been created *ad hoc* during the lockdown period to describe specific initiatives.

By using these hashtags as a queue in the public Twitter API⁵ we have created a corpus with a total of 23,716 tweets.

To better focus on the tweets and their intents concerning cultural communication, we have decided to filter out of the corpus any hashtag with less than 1,000 occurrences. We have thus obtained a queue of six hashtags (#artyouready, #emptymuseum, #museitaliani, #museichiusimuseiaperti, #laculturanonsiferma, #laculturaincasa) and a corpus of 15,988 tweets.

This corpus has been filtered once again so that only unique tweets (i.e. no retweets) written in Italian have been kept. By using a list of GLAMs manually extracted from the corpus, we have then extracted out of the remaining 8,038 tweets those written by a GLAM institution, thus ending up with our final corpus of 3,429 tweets published by 213 Italian cultural institutions. Table 1 shows the occurrences of the hashtags in the final corpus.

Hashtag	# Occ.
#artyouready	367
#emptymuseum	373
#museitaliani	906
#museichiusimuseiaperti	1560
#laculturanonsiferma	668
#laculturaincasa	283
Total	4,157

Table 1: Number of occurrences for each hashtag.

3.2 Annotation Process

In order to define the intents of GLAMs towards the users, the corpus has been annotated with four communication categories first presented by Courtin et al. (2015), and then used by Foucault and Courtin (2016), and Juanals and Minel (2018)

⁴This process is described in details in Carlino et al. (2020)

⁵<https://developer.twitter.com/en>

to annotate the information flow on a social network during a cultural event.

The annotation has been done at tweet level, using a set of labels composed as follows:

- **Sharing Experience - SE:** tweets that share an experience, an opinion or one's feeling

Example: *Eccoci qui oggi a ricordare e a raccontare come i musei chiusi non siano chiusi e i musei vuoti non siano vuoti. Forza!*

(Here we are today, reminding and telling how closed museums are not actually closed and empty museums are not actually empty. Come on!);

- **Promoting Participation - PP:** tweets that require some kind of activity from the users, either in real life or on-line

Example: *Art you ready? Domani partecipa anche tu al contest di @ museitaliani condividendo con noi le tue foto dei musei privi di persone. Cerca fra i ricordi, seleziona la foto, e condividi con # artyouready # MuseumFromHome # iorestoacasa. Ti aspettiamo!*

(Art you ready? Take part in tomorrow's @ museitaliani contest by sharing with us your photos of empty museums. Search through your memories, choose the photo and share it with # artyouready # MuseumFromHome # iorestoacasa. We are waiting for you!);

- **Interacting with the Community - ItC:** tweets through which Institutions create and foster their communities by directly interacting with the users

Example: *Siete stati davvero tanti ad accogliere l'invito a partecipare al flashmob # artyouready e tutti avete postato foto meravigliose! Ecco i tre scatti selezionati tra i più belli*

(So many of you accepted to take part in the # artyouready flashmob, and you all posted great photos! Here are the three shots selected among the most beautiful ones);

- **Promoting-Informing - PI:** tweets that promote or inform other users about activities, exhibitions, or about any sort of information on the museum.

Example: *Il castello di Fénis si trova in Valle d'Aosta circondato da una doppia cinta di mura merlate è caratterizzato da torri quadrate e cilindriche con feritoie e caditoie. (Fénis Castle is located in Aosta Valley, with its double crenellated surrounding walls, it is*

characterized by square and cylindrical towers with loopholes and storm drains).

A fifth category **N/A** has been included in order to classify tweets that do not fit in any of the aforementioned categories, like the ones composed of only hashtags.

Following this set of categories and our guidelines, the tweets have been annotated using the open source platform INCEpTION⁶, and a first round of annotation has been carried on 400 tweets, double annotated by a domain expert and a non-expert in order to calculate the Inter-Annotator Agreement (IAA).

The use of a non-expert was necessary so that the annotation would not have been influenced by any external knowledge (for example the original meaning behind the various hashtags).

The resulting Fleiss' Kappa has revealed to be moderately good at 0.629, which is considered sufficient for the task at hand. As it can be seen from the confusion matrix in Figure 1, the agreement is very strong on **PI** and **ItC**, moderately strong on **SE**, and very weak on **PP**.

Furthermore, 89 tweets have been deemed unus-

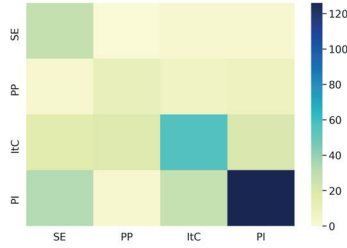


Figure 1: Confusion matrix for the agreement on every label.

able as they have been tagged with the label **N/A**, therefore, they have been removed from the corpus.

Table 2 presents the number of occurrences for each label for the remaining 3,340 tweets. These results show an issue regarding the label **PP**, that is severely underrepresented in the corpus. The effects of this underrepresentation on our classifiers will be explained in detail in Section 4, and the analysis of possible solutions will be the focus of future work.

3.3 Intent classification

In order to train the classifiers, the corpus has been preprocessed so that all tweets are lowercase, and all punctuation marks, URLs, numbers and

⁶<https://inception-project.github.io/>

Label	# Occ.	% corpus
SE	843	25.24%
PP	165	4.94%
ItC	1110	33.23%
PI	1222	36.59%

Table 2: Number of occurrences and percentage in the corpus for each label.

stopwords⁷ removed. The cleaning process has been done via the NLTK package for Python⁸, which has also been used for tokenization.

The experiments have been conducted on six classifiers: five more traditional classifiers trained on a TF-IDF vectorized text (created using the machine learning library for Python Scikit-learn⁹), and a Feed Forward Neural Network¹⁰ created with Keras¹¹ and trained on a 100-dimensions GloVe¹² embedded text.

The set of classifiers is thus the following: a Naive Bayes (**NB**, also used as baseline); a Support Vector Classifier (**SVC**); a the K-Nearest Neighbors classifier (**KNN**); a Decision Tree (**DT**); a Multilayer Perceptron (**MLP**) and a Neural Network classifier (**NN**).

The dataset was split using the *train_test_split* tool found in the sklearn library for Python, which splits the data into random train and test subsets given a test set size. With test size set at 0.3, the training set is composed of 2,338 tweets, and the testing set is composed of the remaining 1,002 tweets.

In order to evaluate the classification task, the values of precision, recall and F1 have been all weighted by the number of samples of each label. The final results are shown in Table 3.

Classifier	P	R	F1
NB	0.69	0.66	0.64
SVC	0.70	0.68	0.67
KNN	0.70	0.39	0.35
DT	0.56	0.55	0.55
MLP	0.66	0.66	0.66
NN	0.64	0.63	0.63

Table 3: System results.

⁷The list of stopwords used is the default one for Italian found in the NLTK package. Furthermore, the term 'Twitter' has been added to this list after the first experiments.

⁸<https://www.nltk.org/>

⁹<https://scikit-learn.org/stable/>

¹⁰Parameters: 4 layers, dropout=0.7, Adam Optimizer

¹¹<https://keras.io/>

¹²<https://nlp.stanford.edu/projects/glove/>

4 Evaluation and Result Analysis

The results show that the methodology adopted in this work can be useful in better understanding how cultural institutions communicate on the Web. The tools used in this specific task are adequate in annotating and automatically classifying the way cultural institutions communicate on the Twitter platform.

That being said, the results shown in Section 3 demonstrate that our experiments can still be improved.

Firstly, the increase in the size of the dataset would surely enhance the performances of the classifiers. In particular, this should be done focusing on the label PP, that, as it can be observed in Table 4, is the less frequent among the four.

Furthermore, while the precision for the label PP is usually higher than the average (note how it reaches 1.00 in our baseline), its recall is very low, even for our **SVM classifier**, which shows the best results overall. The intuition here is that, while it is usually easy for the classifiers to understand which tweet has the PP label, they are also very “picky”, and cannot really learn all the features needed in order to classify this label against the others.

Classifier		SE	PP	ItC	PI
NB	P	0.66	1.00	0.58	0.61
	R	0.37	0.02	0.74	0.76
SVC	P	0.66	0.88	0.72	0.60
	R	0.48	0.52	0.65	0.81
KNC	P	0.36	0.69	0.74	0.68
	R	0.79	0.32	0.43	0.39
DT	P	0.47	0.43	0.51	0.50
	R	0.51	0.50	0.53	0.48
MLP	P	0.59	0.71	0.67	0.68
	R	0.62	0.54	0.68	0.67
NN	P	0.61	0.77	0.60	0.71
	R	0.64	0.47	0.71	0.61

Table 4: Precision (P) and Recall (R) for each label.

Other possible solutions to this issue can be the use of techniques such as resampling and cost-based methods.

Secondly, by focusing on the textual features of the tweets, we can further investigate where improvements can be made.

In particular, looking at the top 5 tf-idf scores for each label (Table 5), we notice that the selected hashtags may occur in all types of tweets with a low difference among their scores. Such a low deviation does not contribute enough

Token	SE	PP	ItC	PI
#museichiusimuseiaperti	1.55	2.47	2.12	1.99
#iorestoacasa	1.66	1.92	1.51	1.71
#museitaliani	2.43	2.35	2.05	2.33
#laculturanonisiferma	2.88	2.47	2.83	2.35
#emptymuseum	3.02	1.97	3.29	3.4
#artyouready	2.9	1.71	3.04	3.23
#laculturaincasa	4.1	-	3.47	2.96
flashmob	-	2.1	-	-
mibact	3.34	2.4	2.7	3.09
oggi	3.03	-	-	2.78
youtube	-	-	3.16	-
cultura	4.18	-	3.12	4.18

Table 5: Top 5 word by their tf-idf score on each label.

to the classification process, as shown by #museichiusimuseiaperti values which are seemingly strong enough as a feature to differentiate PP against the others, but does not do a good job differentiating the other labels against each other.

Those data could give us some insight on how museums communicate through the Twitter platform. Indeed, usually, GLAMs tend to use the same hashtags regardless of their communicative intents (even when the hashtag used was initially linked to certain initiatives), which was already expected with some general hashtags, like #iorestoacasa.

The effects of possible removal or reweighting of these hashtags needs to be further explored.

5 Conclusion and Future Work

In this work, we have described our project for classifying communicative intents in tweets written by Italian GLAMs during the COVID-19 lockdown. Through the experiments and the following analysis we have shown how this task can be challenging.

As future work we will focus on: increasing the size of the corpus, integrating statistical techniques to help dealing with imbalanced labels, and finally improving the selection and reweighting of the features (in particular concerning the hashtags). Another topic which needs further investigation concerns the use of different kinds of textual embeddings, which might improve the result.

Once honed, the methodology and the tools we have used in this research could become an important asset in better understanding and analyzing cultural communication on the Web.

Acknowledgments

This work has been partially supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018 and by PON Ricerca e Innovazione 2014-2020 “Dottorati innovativi con caratterizzazione industriale”.

Authorship Attribution is as follows: Gennaro Nolano is author of Section 3.3 and 4, Carola Carlino is author of Section 3, 3.1 and 5, Maria Pia di Buono is author of Section 2 and 3.2, and Johanna Monti is author of Section 1.

References

- Dana Allen-Greil and Matthew MacArthur. 2010. Small towns and big cities: How museums foster community on-line. *Museums and the Web 2010: Proceedings*. Toronto: Archives & Museum Informatics. Published March, 31(2010):84–96.
- Bráulio Alturas and Liliana Oliveira. 2016. Consumers using social media: impact on companies’ reputation. In *Proceedings of the Academy of Marketing Conference 2016: Radical Marketing*. Academy of Marketing.
- Dimitrios Amanatidis, Ifigeneia Mylona, Spyridon Marmalis, and Irene Eirini Kamenidou. 2020. Social media for cultural communication: A critical investigation of museums’ instagram practices. *Journal of Tourism, Heritage & Services Marketing*, 6(2):38–44.
- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.
- Carola Carlino, Gennaro Nolano, Maria Pia di Buono, and Johanna Monti. 2020. Laculturanonisferma-report su uso e la diffusione degli hashtag delle istituzioni culturali italiane durante il periodo di lockdown. *arXiv preprint arXiv:2005.10527*.
- Andrea Cimino, Stefano Cresci, Felice Dell’Orletta, and Maurizio Tesconi. 2014. Linguistically-motivated and lexicon features for sentiment analysis of italian tweets. In *The 4th Conference for Evaluation of NLP and Speech Tools for Italian (EVALITA)*.
- Antoine Courtin, Brigitte Juanals, Jean-Luc Minel, and Mathilde de saint leger. 2015. A tool-based methodology to analyze social network interactions in cultural fields: The use case “museumweek”. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*.
- Stephen Dann. 2010. Twitter content classification. In *First Monday*, volume 15.
- Nicolas Foucault and Antoine Courtin. 2016. Automatic classification of tweets for analyzing communication behavior of museums. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3006–3013, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Pablo Gamallo and Marcos Garcia. 2014. Citius: A naive-Bayes strategy for sentiment analysis on English tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 171–175, Dublin, Ireland. Association for Computational Linguistics.
- Claudia Giraud. 2020. Il Ministero della Cultura presenta un report sulla reputazione online dei musei durante il Covid, 07. Available online at <https://www.artribune.com/professionie-professionisti/2020/07/il-ministero-della-cultura-presenta-un-report-sulla-reputazione-online-dei-musei-durante-il-covid/>, last accessed 10/09/2020.
- Brigitte Juanals and Jean-Luc Minel. 2018. Analysing cultural events on twitter. In Irene Garrigós and Manuel Wimmer, editors, *Current Trends in Web Engineering*, pages 213–222, Cham. Springer International Publishing.
- Lesley A Langa. 2014. Does Twitter help museums engage with visitors? In *iConference 2014 Proceedings*.
- Ana Sánchez Laws. 2015. *Museum websites and social media: issues of participation, sustainability, trust and diversity*, volume 8. Berghahn Books.
- K. Lovejoy and Gregory D. Saxton. 2012. Information, community, and action: How nonprofit organizations use social media. In *Economic & Social Impacts of Innovation eJournal*.
- NEMO - Network of European Museum Organisations. 2020. Survey on the impact of the COVID-19 situation on museums in Europe: Final Report, 12. Available online at https://www.ne-mo.org/fileadmin/Dateien/public/NEMO_documents/NEMO_COVID19_Report_12.05.2020.pdf, last accessed 10/09/2020.
- Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to

public opinion time series. In *ICWSM*, volume 11, pages 1–2.

Politecnico di Milano. 2020. La reputazione online di musei, parchi archeologici, istituti e luoghi della cultura italiani - Report relativo al mese di maggio 2020., 5. Available online at <http://musei.beniculturali.it/notizie/notifiche/report-maggio-2020-sulla-reputazione-online-dei-musei>.

Daniel Preotjiuc-Pietro, Vasileios Lamps, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Besiki Stvilia and Leila Gibradze. 2014. What do academic libraries tweet about, and what makes a library tweet useful? In *Library & Information Science Research*.

Ruth Talbot, Chloe Acheampong, and Richard H Wicentowski. 2015. Swash: A naive bayes classifier for tweet sentiment identification. In *SemEval@NAACL-HLT*.

Kiran Zahra, Muhammad Imran, and Frank Ostermann. 2019. Automatic identification of eyewitness messages on twitter during disasters. In *Information Processing and Management*, volume 57.

Dejin Zhao and Mary Beth Rosson. 2009. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252.

Creativity Embedding: a vector to characterise and classify plausible triples in deep learning NLP models

Isabeau Oliveri
Politecnico di Torino
isabeau.oliveri@
polito.it

Luca Ardito
Politecnico di Torino
luca.ardito@
polito.it

Giuseppe Rizzo
LINKS Foundation
giuseppe.rizzo@
linksfoundation.com

Maurizio Morisio
Politecnico di Torino
maurizio.morisio@
polito.it

Abstract

English. In this paper we define the creativity embedding of a text based on four self-assessment creativity metrics, namely *diversity*, *novelty*, *serendipity* and *magnitude*, knowledge graphs, and neural networks. We use as basic unit the notion of triple (*head*, *relation*, *tail*). We investigate if additional information about creativity improves natural language processing tasks. In this work, we focus on triple plausibility task, exploiting BERT model and a WordNet11 dataset sample. Contrary to our hypothesis, we do not detect increase in the performance.

Keywords - Creativity Embedding; Creativity Metric; NLP; Creativity Evaluation; Triple; Knowledge Graph; BERT.

1 Introduction

Current conversational agents have emerged as powerful instruments for assisting humans. Oftentimes, their cores are represented by natural language processing (NLP) models and algorithms. However, these models are far from being exhaustive representation of reality and language dynamics, trained on biased data through deep learning algorithms, where the flow among various layers without could result in information loss (Wang et al., 2015). As a consequence, NLP techniques still find it challenging to manage conversation that they have never encountered before, reacting not efficiently to novel scenarios.

One way to mitigate these issues is the integration of structured information, which knowledge graphs are one of the best-known sys-

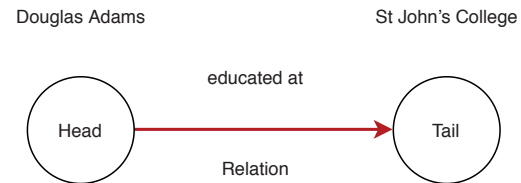


Figure 1: The triple (Douglas Adams, educated at, St John's College), from Wikidata knowledge base (Vrandečić and Krötzsch, 2014), is an example of statement.

tems for representing them. The most prominent example is the Semantic Web (Berners-Lee et al., 2001), where the information is represented through linked statements, each one composed of *head*, *relation*, *tail*, forming a *triple* (Figure 1). This semantic embedding allows significant advantages such as reasoning over data and operating with heterogeneous data sources.

Integration of structured information is not the only method that literature provides us to improve NLP techniques. Previous researches pointed out that analysis of creativity features could improve self-assessment evaluation, with benefits for solutions generated and inputs understanding (Lamb et al., 2018; Karampiperis et al., 2014; Surdeanu et al., 2008). We specify that in this work creativity is intended as capability to create, understand and evaluate novel contents. The concepts of Creativity AI have been discussed in their interconnections with the Semantic Web (Ławrynowicz, 2020), generalizable to knowledge graphs. Kuznetsova et al. (Kuznetsova et al., 2013) define quantitative measures of creativity in lexical compositions, exploring different theories, such as divergent thinking, compositional structure and creative semantic subspace. The crucial point is that no every novel combinations are perceived creative and useful, distinguishing creativity perceived in unconventional, uncommon or

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

”expressive in an interesting, imaginative, or inspirational way”.

Despite it is made clear the interest of the scientific community in exploring this direction, little research is conducted over creativity in the NLP field. The results and the considerations made by Kuznetsova and Ławrynowicz, led us to investigate the possible correlations between improvements in NLP tasks and creativity, with a particular focus on self-assessment. In this paper we introduce a novel approach for supporting deep learning algorithms with a mathematical representation of creativity feature of a text. We named it creativity embedding and based it on metrics of self-evaluation creativity over graph knowledge base.

2 Approach

2.1 Self-assessment creativity metrics

When humans face a problem they never encountered before, they usually perform a self-assessment procedure respect their previous knowledge and context, generally voting for the best solution. Following the example reported in Figure 2, we can imagine that a person has to describe the colour of a grey desk. He does not remind the name of the colour at that time, and performs a creative process. He use a metaphor to describe the grey colour of the desk, referring to the stereotype colour of a ”mouse”. This metaphor is widely accepted, and the colour would be ideally understand by the interlocutor. If in place of ”mouse” the random term ”mask” is used, the meaning will not probably received if not particular context or knowledge is shared between the person and the interlocutor, resulting in a not effective creative process. To emulate this self-assessment procedure, we propose metrics inspired by the related-concept literature, such as recommender systems (Monti et al., 2019) and machine learning (Pimentel et al., 2014; Ruan et al., 2020). The knowledge is represented by a graph of items interconnected by their relation (triples).

We define four metrics, namely diversity (1), novelty (2), serendipity (3), and magnitude (4). In these metrics we make use of a similarity function. In fact, to define the similarity (or the diversity, from another angle) between two or more items, we need a method and a representation that allows us to define a distance

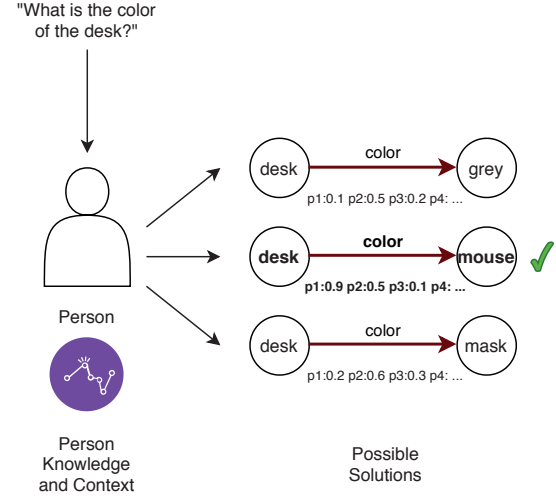


Figure 2: A person produces different solutions to answer a question. Therefore he performs a self-assessment procedure, taking into account several parameters p based on its knowledge and the context. Finally, he chooses the possible best solution. Parameters are expressed as numbers, for simplicity.

between them. In the literature, there is no fixed notion of similarity. However, a common strategy for texts is transforming words and sentences in vectors, taking in account and keeping their distributional properties and connections. Subsequently, mathematical distance functions are applied. The similarity function could defines a semantic similarity function between two items (words or sentences) under these conditions. For prompt understanding, we anticipate that in our experiment we use cosine similarity function and BERT vectors (embeddings) as words representation, as will be discussed in following sections. Nevertheless, thus defined metrics could be computed with different item vector representation and similarity function, as long as it is adopted a similarity function with output domain $[0,1]$, with high value for high similarity.

Diversity (1) represents the semantic diversity between the head h_T and tail t_T of the triple T . This information tells how these two elements are not semantically close. It could be considered as T internal semantic diversity.

$$div(T) = 1 - similarity(h_T, t_T) \quad (1)$$

Novelty (2) of a triple T is its average semantic diversity respect others triples in the context.

Context C is the sub-graph of triple obtained by traversing the paths of length p in the knowledge graph, starting from the triple h_T under examination, collecting n nearest triples. It could be considered as external semantic diversity of T respect to the context C retrieved.

$$nov(T) = \frac{1}{n} \sum_{i=1}^n 1 - similarity(T, C_i) \quad (2)$$

Serendipity (3) is here intended as the semantic novelty of the triple T , taking into account the s most novel triples considering the knowledge graph (refined context S). It could be considered as T novelty relevance.

$$ser(T) = \frac{1}{s} \sum_{i=1}^s 1 - similarity(T, S_i) \quad (3)$$

Magnitude (4) outlines the rarity of the triple, ranking rk each component of the triple by the number of its occurrences over the total number of items in the knowledge graph. The ranking function thus defined has an output domain $[0,1]$.

$$mag(T) = \frac{rk(h_T) + rk(rel_T) + rk(t_T)}{3} \quad (4)$$

2.2 Creativity Embedding

There were no annotated datasets on the creativity characteristics of interest. For this reason, a direct comparison with the ground truth was hampered. To overcome this obstacle, we indirectly measured the effectiveness of this approach by applying it to an external model and judging the results on the triple plausibility task (Yao et al., 2019; Wang et al., 2018; Wang et al., 2015; Padó et al., 2009). The triple plausibility task consists of classifying a dataset's triples in plausible or not plausible classes, comparing the result respect to the ground truth. We choose this task to perform an indirect evaluation of our proposal, rely on the correlation between plausibility and creativity (Lamb et al., 2018), as plausibility could represent a positive outcome of an effective creative process. The current trend in machine learning and natural language processing models pushes the use of mathematical representation of meaningful information utilising vectors, commonly known in this field as embeddings. For these reasons, we outline and train a neural network using the computed ground truth to predict creativity values, and define as creativity embedding the weight of last

hidden layer. This creativity embedding can be added and adapted in its dimension. Stated the above concepts, we define the subsequent research questions.

Research Question: *A creativity embedding extracted from the creativity neural network could improve triple plausibility classification in deep learning models?*

3 Model Architecture

3.1 BERT

We select Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) as a model for investigating the effects of creativity embedding, due to its flexibility and modularity, as well as being state of the art for various NLP tasks. The BERT model could be divided into three main parts: preprocessing of the input, stack of transformer layers, and other layers on top to perform a particular task - typically a classifier. A stack of *Transformers* forms the BERT core. A transformer exploits the attention mechanism to learn the contextual relationship between sentences and words input. The input is not considered in one direction, but figuratively in all ones at one time, defining the context of a word considering the entire surrounding words. The model is trained with a sort of play, where some words or entire sentences are masked, and the model has to predict them. We do not modify the core of the model; we are more interested in the preprocessing part, where we will inject the creativity embedding, as explained in the next section.

3.2 Creativity Neural Network and Creativity CLS Embedding

The outline of the architecture proposed for the task is shown in Figure 3. In the lower part, the triple flows through the BERT model. We used a modified tokenization technique of Knowledge Graph BERT (KG-BERT) (Yao et al., 2019), adapted for the structure of the triple. The triple is split in tokens respect the BERT vocabulary of known words. Special tokens are included in the sequence, classification (CLS) and separator (SEP) tokens. CLS corresponding embeddings are in charge of representing the sentence mathematically, and SEP tokens that separate different sentences. On the KG-BERT version for triple plausibility, SEP is used to separate head words from

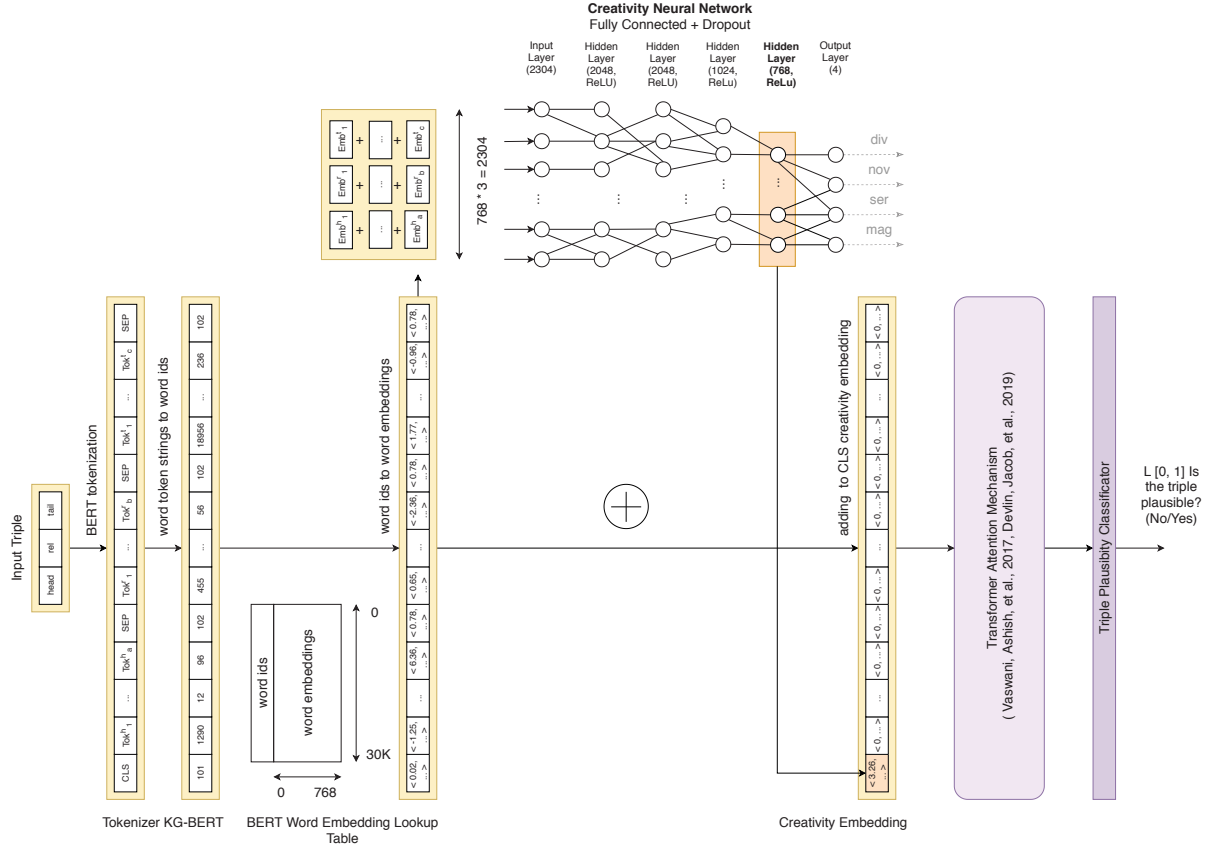


Figure 3: For each triple, Creativity Embedding computed by Creativity Neural Network is added to BERT CLS embedding, defining the Creativity CLS Embedding. A linear classifier on top perform the triple plausibility classification.

relation and tail words in three different sentences. The corresponding token identifiers and embeddings are retrieved through two lookup tables, provided by the BERT model. At the top of Figure 3, we show our creativity neural network. A compact and fixed-size version of the embeddings is obtained from BERT, summing the embeddings of each component of the triple. This compact version feeds the proposed neural network in charge of predicting creativity’s four values and producing creativity embedding. The neural network consists of an input layer ($768 * 3$ neurons), an output layer (4 neurons), 4 fully connected hidden layers with a dropout probability = 0.5. The activation function used is *ReLU*. This neural network structure is basic since its main task is to have a flexible last hidden layer adaptable to the technology that would leverage the creativity embedding. The CLS token is one of the most representative tokens to perform classification and other types of predictions. Came to us exploiting CLS token to adding creative embedding of the triple,

providing the model with a non-empty CLS, Creativity CLS Embedding. In this case, the penultimate layer has been described with several neurons equal to 768, the same size as the BERT embeddings. On the top of the architecture, a linear classifier is in charge of predictions of the plausibility task relying on Creativity CLS Embedding.

4 Experiment

In this experiment we random sample triples from WordNet11 (Miller, 1995) dataset (50000 train, 5000 validation, 3000 test, with positive and negative labels balanced).

Creativity Neural Network. As stated in the previous sections, we compute the four metrics on each triple dataset to create the ground truth. As a similarity function we use cosine similarity, that returns a value between 0 and 1, with high value for high similarity. We applied the cosine similarity function after transforming words and sentences in embeddings, provided by BERT

model. We encountered slowdowns only with novelty metric. The number of nodes is not predictable a priori in our setting, and the mathematical nature of the formula is sensitive to a high number of nodes. Peaks of memory allocation could occur, as well as long computation time. We limit the failure due to out of memory or timeout of the scheduled jobs applying the "divide et impera" paradigm and other adjustments. The length of the path p , seen as recursion deep, is fixed to 5. For each node interested by recursion, the number of maximum neighbor nodes n considered is fixed to 20. Once we obtain all the metrics values, we can train the Creativity Neural Network, as a regression problem. We use: as loss criterion mean squared error loss; as optimizer AdamW with learning rate = 0.001, betas = (0.9, 0.999), epsilon = $1e^{-08}$, weight decay = 0.01; as scheduler StepLR with parameters step size = 10 and gamma = 0.1; we train the model for 10 epochs, size batch of 512. To evaluate performance on test set we compute explained variance score = -0.4493, mean absolute error = 0.1733, mean squared error = 0.0388 and R2 score = -6.7694. Although small values of mean squared and absolute error, R2 tells us that the model do not approximate the distribution better than the "best-fit" line. This is probably due to low entropy of the inputted metrics values, that inspected, result in stationing around 0.5 value.

Triple Plausibility Task. The tokenized triple is inputted to the Creativity Neural Network, obtaining the creativity embeddings. This is added to the CLS embedding token, and the triple flows through the Transformers stack. Therefore, the BERT model is used to make predictions and address the triple plausibility task, putting a linear classifier on top of the Transformer stack. We use as loss function the binary cross-entropy loss function. The literature suggests few epochs and samples for the finetuning process. We finetune BERT for 2 epochs; after we freeze the weights of the model, training only the classifier layer for 3 epochs. We select BERT base uncased as baseline model; as optimizer AdamW with learning rate = $5e^{-05}$, as scheduler a linear scheduler with warm up proportion = 10%; for the classifier dropout probability = 0.5. We fix the maximum sequence length at 100 tokens, as all the triples after tokenization do not exceed this number of tokens.

5 Result and Conclusion

In this paper we investigate if defined creativity embedding improves triple plausibility task, exploiting BERT model. We do not detect an increase in the performance (Table 1), comparing ourselves to KG-BERT results. In this comparison we should point out that the sample used is one fifth of the complete WN11 dataset. This result is somewhat contrary to our expectations, as the creativity embeddings represent in some way a priori information. A possible explanation might be the learning methodology of the creativity embedding: we suppose that a significant loss of information in the process has occurred. Further research might explore other types of embeddings (Grohe, 2020), as graph2vec, and different integration of the proposed metrics. Future experimental investigations may try different parameter configurations. For example, the number of nodes considered intuitively could change the values of metrics as a novelty. Nevertheless, more in-depth data analysis on the used dataset, corresponding knowledge graph, and data correlations could provide additional insights. In future work, we will consider different combinations of metrics defined to train the creativity neural network. It is possible that there are metrics more or not relevant for the task. Selecting metrics strictly relevant will result in a lightening of the computational effort and will give us information about correlations between metrics and results. To conclude, we aim to bring the NLP community's attention to new research topics on creativity.

Acknowledgments

Computational resources provided by HPC@POLITO, which is a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino². We thank the reviewers from CLiC-it 2020 conference for the comments and advices.

References

- Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of

²<http://www.hpc.polito.it>

	Number of triples			Model Metrics			
	Train	Val	Test	Accuracy	Recall	Precision	F1
CE+BERT	50000	3000	5000	0.5093	0.8510	0.5102	0.6379
KG-BERT	225162	5218	21088	0.9334	0.9345	0.9324	0.9334

Table 1: Triple plausibility experiment results.

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Martin Grohe. 2020. Word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS’20*, page 1–16, New York, NY, USA. Association for Computing Machinery.
- P. Karampiperis, A. Koukourikos, and E. Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512.
- Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013. Understanding and quantifying creativity in lexical composition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1246–1258, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Carolyn Lamb, Daniel G. Brown, and Charles L. A. Clarke. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Comput. Surv.*, 51(2), February.
- Agnieszka Ławrynowicz. 2020. Creative ai: A new avenue for the semantic web? *Semantic Web*, pages 69–78.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Diego Monti, Enrico Palumbo, Giuseppe Rizzo, and Maurizio Morisio. 2019. Sequeval: An offline evaluation framework for sequence-based recommender systems. *Information*, 10(5):174.
- Ulrike Padó, Matthew W Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.
- Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing*, 99:215 – 249.
- Yu-Ping Ruan, Zhen-Hua Ling, Xiaodan Zhu, Quan Liu, and Jia-Chen Gu. 2020. Generating diverse conversation responses by creating and ranking multiple candidates. *Computer Speech Language*, 62:101071.
- Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online qa collections. In *Proceedings of ACL-08: HLT*, pages 719–727.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.
- Quan Wang, Bin Wang, and Li Guo. 2015. Knowledge base completion using embeddings and rules. *IJCAI’15*, page 1859–1865. AAAI Press.
- Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

The CREENDER Tool for Creating Multimodal Datasets of Images and Comments

Alessio Palmero Aproso
Fondazione Bruno Kessler
Trento, Italy
aprosio@fbk.eu

Stefano Menini
Fondazione Bruno Kessler
Trento, Italy
menini@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

English. While text-only datasets are widely produced and used for research purposes, limitations set by image-based social media platforms like Instagram make it difficult for researchers to experiment with multimodal data. We therefore developed CREENDER, an annotation tool to create multimodal datasets with images associated with semantic tags and comments, which we make freely available under Apache 2.0 license. The software has been extensively tested with school classes, allowing us to improve the tool and add useful features not planned in the first development phase.¹

Italiano. Mentre i dataset testuali sono ampiamente creati e usati per scopi di ricerca, le limitazioni imposte dai social media basati sulle immagini (come Instagram) rendono difficile per i ricercatori sperimentare con dati multimodali. Abbiamo quindi sviluppato CREENDER, un tool di annotazione per la creazione di dataset multimodali in cui immagini vengono associate a etichette semantiche e commenti, e che abbiamo reso disponibile gratuitamente con la licenza Apache 2.0. Il software è stato testato in un laboratorio con alcune classi scolastiche, permettendoci di ottimizzare alcune procedure e di aggiungere feature non previste nella prima release.

1 Introduction

In the last years, the NLP community has started to focus on the challenges of combining vision

and language technologies, proposing approaches towards multimodal data processing (Belz et al., 2016; Belz et al., 2017). This has led to an increasing need of multimodal datasets with high-quality information to be used for training and evaluating the developed systems. While several datasets have been created by downloading and often adding textual annotation to real online data (see for example the Flickr dataset²), this poses privacy and copyright issues, since downloading and using pictures posted online without the author's consent is often forbidden by social network privacy policies. Instagram terms of use, for example, explicitly forbid collecting information in an automated way without express permission from the platform.³

In order to address this issue, we present CREENDER, a novel annotation tool to create multimodal datasets of images and comments. With this tool it is possible to simulate a scenario where different users access the platform and are displayed different pictures, having the possibility to leave a comment and associate a semantic tag to the image. The same pictures can be shown to different users, allowing a comparison of their comments and online behaviour.

CREENDER can be used in contexts where simulated scenarios are the only solution to collect datasets of interest. One typical example, which we detail in Section 4, is the analysis of the online behaviour of teenagers and young adults, a task that poses relevant privacy issues since underage users are targeted. Giving the possibility to comment images in an Instagram-like setting without giving any personal information to register is indeed of paramount importance, and can be easily achieved with the tool presented in this paper.

¹"Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)."

²<https://yahooresearch.tumblr.com/post/89783581601/one-hundred-million-creative-commons-flickr-images>
³See, for example, <https://help.instagram.com/581066165581870>.

Given its flexibility, CREENDER can however be used for any task where images need to be tagged and/or commented, and multiple annotations of the same image should be preferably collected.

2 Related Work

Several tools have been developed to annotate images with different types of information. Most of them are designed to be run only on a desktop computer and are meant to select parts of the picture to assign a semantic tag or a description, so that the resulting corpora can be used to train or evaluate image recognition or captioning software. In this scenario, users often need to be trained to use the annotation tool, which requires some time that is usually not available in specific settings like schools (Russell et al., 2008). Other tools for image annotation or captioning are web-based, like CREENDER, but the software is not available for download and must be used as a service. This paradigm can lead to privacy issues, as the data are not stored locally or on an owned server (Chapman et al., 2012). This could be problematic when the pictures to be annotated are copyright-protected or when users involved in the data collection do not want/cannot create an account with personal information. Finally, some software is not distributed open source, and could suddenly become unavailable or not usable when not maintained any more (Halaschek-Wiener et al., 2005; Hughes et al., 2018).

Regarding the datasets, Mogadala et al. (2019) focus on prominent tasks that integrate language and vision by discussing their problem formulations, methods, existing datasets, and evaluation measures, comparing the results obtained with different state-of-the-art methods. Ethical and legal issues on the use of pictures and texts taken from social networks are also relevant, as discussed in (Lyons, 2020; Prabhu and Birhane, 2020; Fiesler and Proferes, 2018). Our tool has been developed to address specifically also this kind of issues, preserving the privacy of users and avoiding the collection of real data.

3 Annotation Tool

The CREENDER tool can be accessed both via browser and mobile phone, so that users can use it even if no computer connected to Internet is available. The web interface is multi-language, since English, French and Italian are already included,

while other language files can be added as needed. The interface language can be assigned at user level, meaning that the interface for users on the same instance can be configured in different languages.

Once the tool is installed on a server, a super user is created, who can access the administration interface where the projects are managed with the password chosen during installation (see Figure 2).

For each project, on the configuration side, a set of photos (or a set of external links to images on the web) needs to be given to the tool. Then, one can set the number of users and the number of annotations that are required for each photo. Finally, the system assigns the photos to the users and creates the login information for them. Social login is also supported (only Google for now), so that there is no need to spread users and password: the administrator chooses a five-digit code and gives it to every annotator, that can then log in using the code and his/her social account.

Given a picture, the system can be set to perform three actions in sequence or in isolation, as needed by the task: *i*) the picture can be skipped by the user, so that no annotation is stored and the next one is displayed; *ii*) the user can insert free text associated to the image. This can be used to write a caption, comment the picture, list the contained objects, etc. *iii*) one or more pre-defined categories can be assigned to the picture. Categories can range from specific ones related to the portrayed objects (e.g. male, female, animals, etc.) to more abstract ones, like for example the emotions provoked by looking at the picture.

In the configuration screen, the administrator can edit the prompted questions and the possible answers, so that the tool can be used for a variety of different tasks.

Using the administration web interface, it is also possible to monitor the task with information about the number of annotations that each user has performed. This enables to check whether some users experience difficulties in the annotation, or if some annotators are anomalously fast (for example by skipping too many images). Once the annotation session is closed, the administrator can download the resulting corpus containing the images and the associated information. The export is available in three formats: SQL database, CSV, and JSON.

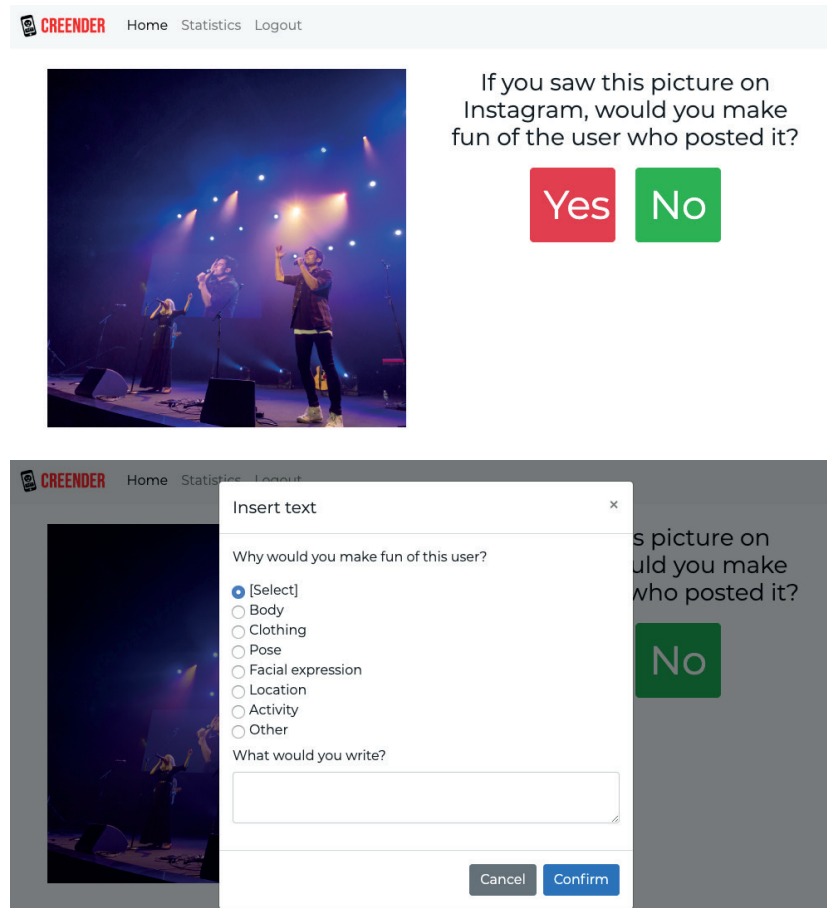


Figure 1: CREENDER interface configured for the collection of potentially offensive comments

4 Use Case: Creation of Offensive Posts

The CREENDER tool was used to collect abusive comments associated to images, simulating a setting like Instagram in which pictures and text together build an interaction which may become offensive. The data collection was carried out in several classes of Italian teenagers aged between 15 and 18, in the framework of a collaboration with schools aimed at increasing awareness on social media and cyberbullying phenomena (Menini et al., 2019). The data collection was embedded in a larger process that required two to three meetings with each class, one per week, involving every time two social scientists, two computational linguists and at least two teachers. During these meetings several activities were carried out with students, including simulating a WhatsApp conversation around a given plot as described in (Sprugnoli et al., 2018), commenting on existing social media posts, and annotating images as described in this paper.

Overall, 95 students were involved in the anno-

tation. The sessions were organised so that different school classes annotated the same set of images, in order to collect multiple annotations on the same pictures. The pictures were retrieved from online sources and then manually checked by the researchers involved in the study to remove pornographic content. In the preparatory phase, the filtered pictures were uploaded in the CREENDER image folder. Then, a login and password were created for each student to be involved in the data collection and printed on paper, so that they could be given to each student before an annotation session without the possibility to associate login information with the students' identity. CREENDER was configured to first take a random picture from the image folder, and display it to the user with a prompt asking "If you saw this picture on Instagram, would you make fun of the user who posted it?". If the user selects "No", then the system picks another image randomly and the same question is asked. If the user clicks on "Yes", a second screen opens where the user is asked to specify the reason why the image would trigger

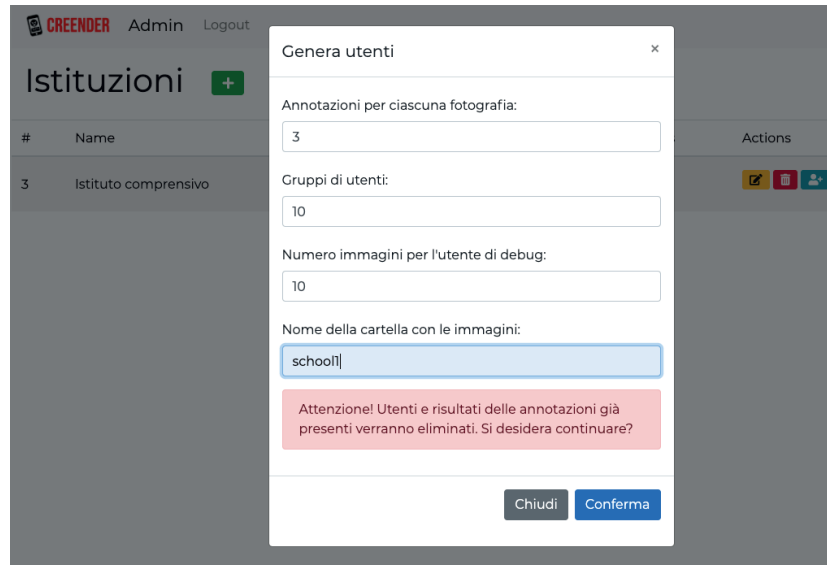


Figure 2: The administration interface to define the number of users and the images per user

such reaction by selecting one of the following categories: “*Body*”, “*Clothing*”, “*Pose*”, “*Facial expression*”, “*Location*”, “*Activity*” and “*Other*”. Two screenshots of the interface are displayed in Figure 1. The user should also write the textual comment s/he would post below the picture. After that, the next picture is displayed, and so on. A screenshot of the tool configured for this specific task is displayed in Figure 1.

At the end of the activities with schools, all collected data were exported. The final corpus includes almost 17,912 images, 1,018 of which have at least one associated comment, as well as a trigger category (e.g. *facial expression, pose*) and the category of the subject/s (*female, male, mixed* or *nobody*). The number of annotations for each picture may vary between 1 to 4. A more detailed description of the corpus is reported in (Menini et al., 2021).

The use of CREENDER allowed a seamless and very fast data collection, without the need to send images to each student, to exchange or merge files and to install specific applications. On the other hand, the data collection with students, who used the online platform in classes while researchers were physically present and could check the flow of the interaction, was useful to improve the tool. Some bug fixes and small improvements were indeed implemented after the first sessions. For example, a small delay (2 seconds) was added after the image is displayed to the user and before the *Yes/No* buttons appear, so that users are more

likely to look at the picture before deciding to skip it or not.

5 Release

The software is distributed as an open source package⁴ and is released under the Apache license (version 2.0). The API (backend) is written in php and relies on a MySQL database. The web interface (frontend) is developed using the HTML/CSS/JS paradigm using the modern Bootstrap and VueJS frameworks.

The interface is responsive, so that one can use it from any device that can open web pages (desktop computers, smartphones, tablets).

6 Conclusions

In this work we present a methodology and a tool, CREENDER, to create multimodal datasets. In this framework, participants in online annotation sessions can write comments to images, assign pre-defined categories or simply skipping an image. The tool is freely available with an interface in three languages, and allows setting up easily annotation sessions with multiple users.

CREENDER has been extensively tested during activities with schools around the topic of cyberbullying, involving 95 Italian high-school students. The tool is particularly suitable for this kind of settings, where privacy issues are of paramount importance and the involvement of un-

⁴<https://github.com/dhfbk/creender>

derage people requires that personal information is not shared.

In the future, we plan to continue the annotation of images related to cyberbullying, creating and comparing subsets of pictures related to different topics (e.g. religious symbols, political parties, football teams). From an implementation point of view, we will extend the analytics panel, adding for example scripts for computing inter-annotator agreement.

Acknowledgments

Part of this work has been funded by the KID ACTIONS REC-AG project (n. 101005518) on “Kick-off preventIng and responDing to children and AdolesCent cyberbullyIng through innovative mOnitoring and educatioNal technologieS”. In addition, the authors want to thank all the students and teachers who participated in the experimentation.

References

- Anya Belz, Erkut Erdem, Krystian Mikolajczyk, and Katerina Pastra, editors. 2016. *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany, August. Association for Computational Linguistics.
- Anya Belz, Erkut Erdem, Katerina Pastra, and Krystian Mikolajczyk, editors. 2017. *Proceedings of the Sixth Workshop on Vision and Language*, Valencia, Spain, April. Association for Computational Linguistics.
- Brian E Chapman, Mona Wong, Claudiu Farcas, and Patrick Reynolds. 2012. Annio: a web-based tool for annotating medical images with ontologies. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 147–147. IEEE.
- Casey Fiesler and Nicholas Proferes. 2018. “participant” perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.
- Christian Halaschek-Wiener, Jennifer Golbeck, Andrew Schain, Michael Grove, Bijan Parsia, and Jim Hendler. 2005. Photostuff-an image annotation tool for the semantic web. In *Proceedings of the 4th international semantic web conference*, pages 6–10. Citeseer.
- Alex J Hughes, Joseph D Mornin, Sujoy K Biswas, Lauren E Beck, David P Bauer, Arjun Raj, Simone Bianco, and Zev J Gartner. 2018. Quanti.us: a tool for rapid, flexible, crowd-based annotation of images. *Nature methods*, 15(8):587–590.
- Michael J Lyons. 2020. Excavating” excavating ai”: The elephant in the gallery. *arXiv preprint arXiv:2009.01215*.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110.
- Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. A multimodal dataset of images and text to study abusive language. In *7th Italian Conference on Computational Linguistics, CLiC-it 2020*.
- Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.
- Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision?
- Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173.
- Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.

(Stem and Word) Predictability in Italian verb paradigms: An Entropy-Based Study Exploiting the New Resource *LeFFI*

Matteo Pellegrini¹, Alessandra Teresa Cignarella^{2,3}

1. Liceo Statale “Augusto Monti” di Chieri, Italy

2. Dipartimento di Informatica, Università degli Studi di Torino, Italy

3. PRHLT Research Center, Universitat Politècnica de València, Spain

pellegrini.matteo@liceomonti.edu.it, cigna@di.unito.it

Abstract

English. In this paper we present *LeFFI*, an inflected lexicon of Italian listing all the available wordforms of 2,053 verbs. We then use this resource to perform an entropy-based analysis of the mutual predictability of wordforms within Italian verb paradigms, and compare our findings to the ones of previous work on stem predictability in Italian verb inflection.

1 Introduction

The pioneering work of Aronoff (1994) has inspired an influential line of research where predictability within inflectional paradigms is modelled by resorting to the notion of morphomic stems – i.e., stems that cannot be considered as bearing any meaning, as they appear in groups of cells that do not share a fixed morphosyntactic content. In this perspective, every lexeme is seen as equipped with a set of indexed stems, that only for regular lexemes are mutually predictable, while for irregular verbs they need to be independently stored. From each of these stems, a fixed set of wordforms can be obtained by adding the appropriate inflectional endings. An analysis relying on these assumptions was proposed by Maiden (1992) and subsequent work – see Maiden (2018) for a recent survey – to account for the patterns of stem allomorphy that are found in the verbal inflection of Romance languages in general. More detailed implementations of these ideas have then been provided for individual languages, among them Italian (Pirrelli and Battista, 2000; Montermini and Boyé, 2012; Montermini and Bonami, 2013). Another possibility that has been explored in more recent times is tackling the issue of inflectional predictability in terms of predictions of wordforms from one another, without

assuming a given segmentation in stems vs. endings, in a fully word-based, abstractive (Blevins, 2016) approach. Within this framework, Ackerman et al. (2009) propose to estimate the reliability of inflectional predictions by means of the information-theoretic notion of conditional entropy. Building on this work, Bonami and Boyé (2014) outline a procedure that allows to compute entropy values estimating the uncertainty in predicting one cell from another one directly from a lexicon of fully inflected wordforms in phonological transcription, using the type frequency of different inflectional patterns to estimate their probability of application. This method has been applied to French by Bonami and Boyé (2014), to Latin by Pellegrini (2020), and it has been used for typological comparison on a small sample of languages by Beniamine (2018), who also provides a freely available toolkit (Qumin) allowing to perform this computation automatically for any language.

A similar entropy-based analysis has not been proposed for Italian yet. To be able to use the Qumin toolkit to perform it, it is necessary to have an inflected lexicon listing all the wordforms of a representative number of lexemes in phonological transcription, like e.g. Flexique for French (Bonami et al., 2014) or LatInflexi for Latin (Pellegrini and Passarotti, 2018). Looking for such a resource for Italian, we can see that in most lexicons wordforms are given in orthographic transcription – see e.g. Morph-it! (Zanchetta and Baroni, 2005) and CoLFIS (Bertinetto et al., 2005). On the other hand, in PhonItalia (Goslin et al., 2014) there are phonological transcriptions, but not all the inflected wordforms of each lexeme are listed. To the best of our knowledge, the only resource providing phonological transcriptions of the full paradigm of lexemes is GLAFF-IT (Calderone et al., 2017), but due to the way in which it was created, it proves to be too noisy to be used for entropy computations as such.

In this paper, we describe the work that was done to obtain a smaller, but cleaner version of GLAFF-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

IT. We then use this resource to perform an entropy-based analysis of predictability in Italian verb inflection. After briefly describing the methodology, we present our results comparing them with the findings of previous stem-based analyses.

2 The Resource

In order to build *LeFFI* (*Lessico delle Forme Flesse dell'Italiano*), we have firstly consulted GLAFF-IT, a free machine-readable dictionary based on *Wikizionario*, the Italian language edition of Wiktionary. It is a morphophonological Italian lexicon which contains a total of 485,135 word-forms among verbs, nouns, adjectives and adverbs, in both orthographic and phonological IPA transcription. Since our interest for the present research lies only in verbs, in this step a total of 411,770 verbal forms in phonological transcription have been extracted from GLAFF-IT, together with the citation form (the infinitive) of the lexeme they belong to, thus resulting in a list of the complete paradigms of 7,552 verbs. To indicate the morphosyntactic properties expressed by each wordform, we use the notation of the Leipzig Glossing Rules (Comrie et al., 2008), both in our resource and in the examples shown in this paper.

Due to the large amount of manual work needed in order to obtain our resource, for the time being we have decided to focus only on a fraction of this list. So as not to lose quantitatively relevant data, our selection was based on the frequency of lexemes, as reported in the CoLFIS frequency lexicon. We have thus crossed the list of 7,552 verbs extracted from GLAFF-IT with the 5,193 verbal lexemes contained in CoLFIS, and kept only the ones with a frequency higher than 10. The resulting dataset, listing the 53 available, non-periphrastic cells of 2,053 verbs, is still large enough to allow for reasonably safe generalizations on Italian verb inflection.

After these automatic steps, several manual changes have been made in order to obtain the current version of our resource. Firstly, it should be noticed that many of the phonological transcriptions provided by GLAFF-IT are obtained automatically from the orthographic form. In some cases, however, it is not possible to infer a precise phonological transcription from orthography alone, because some graphemes can correspond to different phonemes. In such cases, the phonological transcriptions provided by GLAFF-IT are underspec-

ified: for instance, the symbol E is used for the grapheme ⟨e⟩, that can correspond to /e/ or /ɛ/, and similarly O for ⟨o⟩ (/o/ or /ɔ/), S for ⟨s⟩ (/s/ or /z/), Z for ⟨z⟩ (/ts/ or /dz/). While we have manually reconducted ⟨s⟩, ⟨z⟩ and a few other marginal ambiguous graphemes to the actual phonemes they correspond to, for ⟨e⟩ and ⟨o⟩ we have decided to keep the same neutralization as in GLAFF-IT. This choice is due to the fact that manually disambiguating all cases to reflect the actual pronunciation in the standard variety of Italian would have been very time consuming, but it is also justified by the fact that in many varieties (including the northern ones of the authors) these distinctions are not made.

Another systematic correction concerns the placement of stress, that for many wordforms have been obtained automatically in GLAFF-IT, and sometimes turns out not to be in the right place: for instance, in many third-plural forms, the stress is incorrectly placed on the penultimate (e.g. PRS.IND.3SG /divent'ano/ 'they become', /okkup'ano/ 'they occupy'), while in our resource we move it to the (pre)antepenultimate (e.g. /div'entano/, /'okkupano/). While in other cases it was possible to correct stress position in an automatic way, by moving the stress to the syllable where it is systematically placed (e.g. the antepenultimate in forms like PRET.IND.3SG /f'etfero/ 'they did'), in this case, since there are two alternatives, the changes had to be done semi-automatically, by automatically moving the stress to the antepenultimate, and then manually moving it to the preantepenultimate whenever needed.

In cases of cells containing more than one wordform, we keep only one of the cell-mates. Wherever it was possible, we have used Thornton (2008)'s description of overabundance in Italian verb inflection to select the less marginal variant (e.g., keeping /d'evo/ rather than /d'ebbo/ in the PRS.IND.1SG of *DOVERE* 'must').

Several other punctual corrections were manually made on the data of GLAFF-IT, yielding the current version of our resource, that is clean enough to be able to perform an entropy-based analysis shedding light on the patterns of interpredictability between wordforms in Italian verb paradigms.

3 The Method

The Qumin toolkit computes implicative entropy values estimating the uncertainty in predicting each paradigm cell assuming knowledge of one (or more

than one) wordform, following the procedure described in Beniamine (2018). Here, we illustrate the methodology using the data given in Table 1.

lexeme	conj.	GER	PRS.IND.2PL
AMARE ‘love’	1 st	/am'ando/	/am'ate/
VEDERE ‘see’	2 nd	/ved'endo/	/ved'ete/
SENTIRE ‘hear’	3 rd	/sent'endo/	/sent'ite/

Table 1: Italian verbs of different conjugations.

The first step of the procedure consists in classifying verbs according to the patterns of formal alternation between wordforms, and the phonological context in which such alternations are attested. As is shown in the second column of Table 2, 1st and 2nd conjugation verbs display the same pattern (1), while 3rd conjugation verbs use another pattern (2). The second step is another classification based on the patterns that can potentially be applied to GER to obtain PRS.IND.2PL. As can be seen in the third column of Table 2, verbs of the 2nd and 3rd conjugation are in the same class (B), because patterns 1 and 2 can potentially be applied to a GER ending in /endo/, while only pattern 1 can be applied to 1st conjugation verbs with GER in /ando/. Entropy is then computed for each of the classes of this second classification, weighing the probability of application of different patterns by means of their type frequency in the data, i.e., the number of verbs in which they are attested: here, data from *LeFFI* are given in the last column of Table 2.

lexeme	pattern/context (1SG ↔ 3SG)	applicable patterns	n. verbs
AMARE	1 (_ndo ↔ _te / V_#)	A (1)	1,505
VEDERE	1 (_ndo ↔ _te / V_#)	B (1,2)	320
SENTIRE	2 (_endo ↔ _ite / C_#)	B (1,2)	215

Table 2: Information used to compute the entropy of predicting PRS.IND.2PL from GER.

$$\begin{aligned}
& H(\text{PRS.IND.2PL}|\text{GER}) \\
&= - \left(\frac{1,505}{2,040} \times \left(\frac{1,505}{1,505} \times \log_2 \frac{1,505}{1,505} \right) + \right. \\
&\quad \left. + \frac{535}{2,040} \times \left(\frac{320}{535} \times \log_2 \frac{320}{535} + \frac{215}{535} \times \log_2 \frac{215}{535} \right) \right) = \\
&= - \left(\frac{1,505}{2,040} \times 0 + \frac{535}{2,040} \times -0.972 \right) = \\
&= 0.255
\end{aligned}
\tag{1}$$

As is shown in Equation 1, there is no uncertainty in class A: given a GER in /ando/, PRS.IND.2PL cannot but be in /ate/. On the other hand, given a GER in /endo/, PRS.IND.2PL can be in /ete/ (applying pattern 1) or in /ite/ (applying pattern 2). As a consequence, there is some uncertainty in this case. The entropy values of different classes are then summed and weighed – again on the basis of type frequency – in a single entropy value, that estimates the overall uncertainty in predicting PRS.IND.2PL from GER in Italian verbs.

4 Results

Giving the data of *LeFFI* as input to the Qumin toolkit, the output is an entropy-based distance matrix of all the cells of Italian verb paradigms. We do not show it here for reasons of space as it comprises 53 columns and rows, but we use its values to draw a mapping of the paradigm in zones of full interpretability, where two cells *A*, *B* are conflated in the same zone if they can be predicted from one another with no uncertainty, i.e. if $H(A|B) = H(B|A) = 0$. The outcome of this grouping is given in Table 3.

	1SG	2SG	3SG	1PL	2PL	3PL
FUT.IND	Z6	Z6	Z6	Z6	Z6	Z6
PRS.COND	Z6	Z6	Z6	Z6	Z6	Z6
PRS.SBJV.	Z2	Z2	Z2	Z4	Z4	Z2
PRS.IND	Z12	Z14	Z15	Z4	Z10	Z13
IPRF.IND	Z9	Z9	Z9	Z9	Z9	Z9
IPRF.SBJV	Z1	Z1	Z1	Z1	Z1	Z1
PRET.IND	Z5	Z1	Z5	Z1	Z1	Z5
IMP		Z3		Z4	Z10	

	M.SG	F.SG	M.PL	F.PL
PST.PTCP	Z7	Z7	Z7	Z7
PRS.PTCP	Z11		Z11	
GER	Z11			
INF	Z8			

Table 3: Zones of interpredictability in Italian verb paradigms: verbal forms.

Given this mapping, we can obtain what Stump and Finkel (2013) call a distillation of the paradigm by keeping only one cell for each zone, since all the other cells are trivially predictable. In Table 4, we show entropy values for a distillation of the Italian verbal paradigm.

It is interesting to compare the mapping of Table 3, based on the predictability of wordforms, to the one based on the predictability of stems that has been proposed in the works cited in Section 1, where cells are grouped together if they

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	Z11	Z12	Z13	Z14	Z15
Z1 (IPRF.SBJV.3SG)		.431	.364	.416	.045	.029	.029	.091	.009	.019	.009	.385	.347	.444	.357
Z2 (PRS.SBJV.3SG)	.405		.213	.474	.394	.423	.393	.436	.413	.405	.342	.213	.213	0	.213
Z3 (IMP.2SG)	.269	.006		.708	.226	.273	.235	.310	.273	.268	.172	.005	.005	0	.002
Z4 (PRS.IND.1PL)	1.239	1.238	1.419		.852	.743	.893	1.215	1.230	1.228	1.033	.965	1.322	.775	1.394
Z5 (PRET.IND.3SG)	.015	.443	.374	.451		.034	.056	.035	.044	.015	.044	.397	.359	.457	.370
Z6 (FUT.IND.3SG)	.495	.864	.856	.527	.231		.187	.526	.498	.503	.487	.460	.760	.466	.853
Z7 (PST.PTCP.F.SG)	.013	.435	.378	.426	.005	.026		.027	.006	.006	.006	.387	.352	.458	.376
Z8 (INF)	.032	.435	.366	.524	.045	.160	.033		.030	0	.026	.386	.358	.450	.377
Z9 (IPRF.IND.3SG)	.011	.429	.367	.417	.044	.458	.031	.086		.010	0	.380	.346	.442	.357
Z10 (PRS.IND.2PL)	.041	.435	.366	.428	.053	.494	.033	.083	.031		.026	.386	.358	.451	.367
Z11 (GER)	.265	.557	.503	.417	.128	.262	.141	.325	.254	.256		.474	.472	.582	.502
Z12 (PRS.IND.1SG)	.731	.830	.567	.703	.442	.366	.429	.712	.731	.727	.682		.830	.118	.572
Z13 (PRS.IND.3PL)	.248	0	0	.620	.229	.243	.228	.278	.248	.248	.176	0		0	0
Z14 (PRS.IND.2SG)	.991	.559	1.033	.544	.622	.502	.627	.976	.991	.986	.909	.329	.744		1.038
Z15 (PRS.IND.3SG)	.269	.003	0	.717	.233	.268	.239	.312	.270	.268	.173	.003	.003	0	

Table 4: Entropy values for a distillation of the Italian verb paradigm.

are based on the same stem. For this comparison, we refer to Montermini and Bonami (2013), where the most recent version of the stem-based mapping is provided. In their description, 8 stems are identified, while our word-based mapping is composed of 15 zones. In particular, Z1-9-10-11 of our mapping correspond to the zones including cells that are based on the same stem S1 in Montermini and Bonami (2013)’s analysis: this is why they are all colored with different shades of red in Table 3. Similarly, our Z2-12-13 (different shades of blue) include cells based on Montermini and Bonami (2013)’s S2 and our Z3-14-15 (different shades of green) include cells based on Montermini and Bonami (2013)’s S3. As for the other zones of our mapping, there is a one-to-one correspondence with the stems identified by Montermini and Bonami (2013).

The discrepancies between the two approaches are mostly due to two different reasons: (i) the presence of a few, highly irregular verbs¹ that are not accounted for by Montermini and Bonami (2013)’s analysis, but are included in our dataset, and, therefore, in our entropy-based analysis; (ii) more systematic opacities of some wordforms, that are poorly informative on the conjugation of lexemes.

As an example of case (i), PRS.IND.2PL and IPRF.IND.3SG can almost always be predicted from one another by replacing the final segments /te/ with /va/, or *vice versa*: e.g. AMARE (PRS.IND.2PL /am'ate/, IPRF.IND.3SG /am'ava/) and SENTIRE (PRS.IND.2PL /sent'ite/, IPRF.IND.3SG /sent'iva/).

¹Namely: ANDARE ‘to go’, AVERE ‘to have’, DARE ‘to give’, DIRE ‘to say’, ESSERE ‘to be’, FARE ‘to do’, SAPERE ‘to know’, and STARE ‘to stay’.

However, this generalization does not hold for a handful of highly irregular verbs, as is exemplified by DIRE ‘say’, with PRS.IND.2PL /d'ite/ but IPRF.IND.3SG /ditf'eva/. Of course, the picture is different depending on the presence of such irregular verbs in the data. If they are excluded, as in Montermini and Bonami (2013), the two cells can be considered as based on the same stem (S1) and, thus, as being fully interpredictable. If they are included, as happens in our data, the two cells have to be assigned to different zones, since there is some uncertainty in predicting the cells from one another. However, entropy is very low in such cases, thanks to the weighing based on type frequency (see the corresponding values in Table 4). It should be noticed that the lexemes that are not considered by Montermini and Bonami (2013) because of their irregularity are among the verbs with higher token frequency in Italian (all ranking among the first 13 positions in COLFIS). This makes their exclusion less worrisome, as the irregular formal patterns they display can plausibly be considered as being learned by rote. Nevertheless, our entropy-based picture can be considered as achieving a higher level of granularity in the description.

As an example of case (ii), PRS.IND.2SG and PRS.IND.3SG are in the same zone in Montermini and Bonami (2013), because they are both considered as obtained from S3: in particular, PRS.IND.3SG is identical to S3, while to obtain PRS.IND.2SG the final vowel of S3 has to be replaced by /a/. In both cases, knowing the shape of S3 is sufficient to infer the cell without any uncertainty. However, in our word-based perspective there is uncertainty when guessing PRS.IND.3SG from PRS.IND.2SG: the latter al-

ways ends in /i/ (e.g. AMARE /'ami/, VEDERE /v'edi/), neutralizing the distinction between verbs of different conjugations, and, thus, not allowing to discriminate between 1st conjugation verbs with S3 and PRS.IND.3SG in /a/ (e.g. AMARE /'ama/) and 2nd and 3rd conjugation verbs with S3 and PRS.IND.3SG in /e/ (e.g. VEDERE /v'ede/).

These examples show that our method allows to identify sources of uncertainty that are downplayed in the stem-based picture, either because of their quantitative marginality – case (i) – or because they are obscured by the use of an abstract stem, that however is not always inferrable by the shape of the single wordform used as predictor – case (ii).

However, it should be noticed that at least the possible availability of more exhaustive stem spaces accounting for all the formal variation of Italian verb inflection, without excluding highly irregular verbs – thus corresponding to our case (i) – was already acknowledged in the works cited above: see e.g. Pirrelli and Battista (2000, Footnote 16) and Montermini and Bonami (2013, Footnote 9). Indeed, there is of course a trade-off between the number of zones in which the paradigm is split on the one hand, and the coverage of the identified zones with respect to the whole lexicon on the other hand. In the stem-based mapping, the choice is not to make the number of zones too high, at the (minimal) cost of not accounting for a handful of irregular verbs. Conversely, in the word-based mapping that we adopt in the present paper, the higher number of zones is compensated by a complete coverage of the whole lexicon. Now, how many of the zones are actually identified and learned by speakers is an empirical matter that should be tackled by means of psycholinguistic experiments. However, what is important to keep in mind is that this gap between the two approaches can be filled, either by drawing the stem space in such a way that it covers also for irregular verbs, or by reducing the number of zones in the word-based analysis gradually collapsing zones of interpredictability for increasing values of implicative entropy. For instance, if the criterion for two cells to be assigned to the same zone is for them to be predictable from one another with an implicative entropy value lower than 0.01, rather than 0, then Z3,13,15 can be merged in a same zone. If the threshold is set at 0.02, also Z1 and Z9 can be conflated in the same zone, to which also Z7 can be added with threshold set at 0.03.

On the other hand, the discrepancy between the

two approaches generated by more systematic, but unidirectional opacities such as the one described above in (ii) could be avoided if in the entropy-based mapping we decided that having null entropy in one direction would be a sufficient criterion for two cells to be assigned to the same zone – i.e., two cells belong to the same zone if either $H(A|B)$ or $H(B|A) = 0$.

5 Conclusions

In this paper, we have presented the inflected lexicon of Italian verbs *LeFFI*. We have then exploited it to investigate predictability in Italian verb inflection, using implicative entropy to estimate the uncertainty in predicting wordforms from one another. The results have been used to obtain a mapping of the paradigm in zones of interpredictability, that we have compared to the mapping of stems proposed in previous work, showing that our word-based procedure is capable of capturing aspects that are downplayed, if not ignored in the stem-based approach.

Besides their theoretical interest, both the resource and the information-theoretic approach potentially have more practical applications, for instance in the field of psycholinguistics. The resource provides a very clean but sufficiently large dataset of forms that can be used as a source of input for fine-grained experiments. In such experiments, it would be possible to test if the different levels of predictability between cells identified by different values of implicative entropy find a correspondence in the process of acquisition of inflectional morphology by L1 and L2 speakers – i.e., if the pairs of cells between which there are higher implicative entropy values are indeed the ones on which learners are more uncertain. More generally, our entropy-based evaluation of uncertainty in inflectional predictions can be considered as a measure of (at least one aspect of) morphological complexity, that can be used also in other areas, for instance to assess text readability.

6 Availability of Data and Tools

The data and tools used in this study are freely available online, allowing for an easy replication of the presented results. *LeFFI* can be found in the following repository: <https://github.com/matteo-pellegrini/LeFFI>. The Qumin toolkit that was used to automatically perform entropy computations can be freely downloaded at: <https://github.com/XachaB/Qumin>.

References

- Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins, editors, *Analogy in grammar: Form and acquisition*, pages 54–82. Oxford University Press, Oxford.
- Mark Aronoff. 1994. *Morphology by itself: Stems and inflectional classes*. MIT press, Cambridge.
- Sacha Beniamine. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. Ph.D. thesis, Université Sorbonne Paris Cité-Université Paris Diderot.
- Pier Marco Bertinetto, Cristina Burani, Alessandro Laudanna, Lucia Marconi, Daniela Ratti, Claudia Rolando, and Anna Maria Thornton. 2005. CoLFIS (Corpus e Lessico di Frequenza dell’Italiano Scritto).
- James P. Blevins. 2016. *Word and paradigm morphology*. Oxford University Press, Oxford.
- Olivier Bonami and Gilles Boyé. 2014. De formes en thèmes. In Florence Villoing, Sarah Leroy, and Sophie David, editors, *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pages 17–45. Presses Universitaires de Paris-Ouest, Paris.
- Olivier Bonami, Gauthier Caron, and Clément Plancq. 2014. Construction d’un lexique flexionnel phonétisé libre du français. In *Congrès Mondial de Linguistique Française — CMLF 2014*, volume 8, pages 2583–2596. EDP Sciences.
- Basilio Calderone, Matteo Pascoli, Franck Sajous, and Nabil Hathout. 2017. Hybrid method for stress prediction applied to GLAFF-IT, a large-scale Italian lexicon. In *International Conference on Language, Data and Knowledge*, pages 26–41, Cham. Springer.
- Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig*.
- Jeremy Goslin, Claudia Galluzzi, and Cristina Romani. 2014. PhonItalia: a phonological lexicon for Italian. *Behavior research methods*, 46(3):872–886.
- Martin Maiden. 1992. Irregularity as a determinant of morphological change. *Journal of linguistics*, 28(2):285–312.
- Martin Maiden. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford University Press, Oxford.
- Fabio Montermini and Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio*, 12(2):171–190.
- Fabio Montermini and Gilles Boyé. 2012. Stem relations and inflection class assignment in Italian. *Word Structure*, 5(1):69–87.
- Matteo Pellegrini and Marco Passarotti. 2018. LatInFLexi: an inflected lexicon of Latin verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.
- Matteo Pellegrini. 2020. Using LatInFLexi for an Entropy-Based Assessment of Predictability in Latin Inflection. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 37–46.
- Vito Pirrelli and Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection: 2628. *Italian Journal of Linguistics*, 12(2):307–380.
- Gregory Stump and Raphael A. Finkel. 2013. *Morphological typology: From word to paradigm*. Cambridge University Press, Cambridge.
- Anna M. Thornton. 2008. A non-canonical phenomenon in Italian verb morphology: double forms realizing the same cell. *Poster presented at OxMorph1–Oxford*.
- Eros Zanchetta and Marco Baroni. 2005. Morph-it!: A free corpus-based morphological resource for the Italian language. *Proceedings of corpus linguistics*.

A deep learning model for the analysis of medical reports in ICD-10 clinical coding task

Marco Polignano University of Bari A. MORO Dept. Computer Science E.Orabona 4, Italy marco.polignano@uniba.it	Pierpaolo Basile University of Bari A. MORO Dept. Computer Science E.Orabona 4, Italy pierpaolo.basile@uniba.it	Marco de Gemmis University of Bari A. MORO Dept. Computer Science E.Orabona 4, Italy marco.degemmis@uniba.it
--	--	---

Pasquale Lops
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
pasquale.lops@uniba.it

Giovanni Semeraro
University of Bari A. MORO
Dept. Computer Science
E.Orabona 4, Italy
giovanni.semeraro@uniba.it

Abstract

English. The practice of assigning a uniquely identifiable and easily traceable code to pathology from medical diagnoses is an added value to the current modality of archiving health data collected to build the clinical history of each of us. Unfortunately, the enormous amount of possible pathologies and medical conditions has led to the realization of extremely wide international codifications that are difficult to consult even for a human being. This difficulty makes the practice of annotation of diagnoses with ICD-10 codes very cumbersome and rarely performed. In order to support this operation, a classification model was proposed, able to analyze medical diagnoses written in natural language and automatically assign one or more international reference codes. The model has been evaluated on a dataset released in the Spanish language for the eHealth challenge (CodiEsp) of the international conference CLEF 2020, but it could be extended to any language with latin characters. We proposed a model based on a two-step classification process based on BERT and BiLSTM. Although still far from an accuracy sufficient to do without a licensed physician opinion, the results obtained show the feasibility of the task and are a starting point for future studies in this direction.

Italian. La pratica di assegnare un codice univocamente identificabile e facilmente riconducibile ad una patologia a partire da diagnosi mediche e un valore aggiunto alla attuale modalità di archiviazione dei dati sanitari raccolti per costruire la storia clinica di ciascuno di noi. Purtroppo però, l'enorme numero di possibili patologie e condizioni mediche ha portato alla realizzazione di codifiche internazionali estremamente ampie e di difficile consultazione anche per un essere umano. Tale difficoltà rende la pratica di annotazione delle diagnosi con i codici ICD-10 molto complessa e raramente svolta. Col fine di supportare tale operazione si è proposto un modello di classificazione, in grado di analizzare le diagnosi mediche scritte in linguaggio naturale ed assegnarle automaticamente uno o più codici internazionali di riferimento. Il modello è stato valutato su un dataset rilasciato in lingua Spagnola per la challenge (CodiEsp) di eHealth della conferenza internazionale CLEF 2020 ma è di semplice estensione su qualsiasi lingua con caratteri latini. Abbiamo proposto un modello basato su due passi di classificazione e basati sull'utilizzo di BERT e delle BiLSTM. I risultati ottenuti, seppur ancora lontani da una accuratezza sufficiente per far a meno di un parere di un medico esperto, mostrano la fattibilità del task e si pongono come punto di partenza per futuri studi in tale direzione.

1 Introduction

In many of the existing systems for storing patient clinical data, the medical report of the doctor is stored in the form of textual data. Only in a few cases, one or more identification codes are assigned to each of the diagnosed conditions. The process of assigning a unique code to pathologies, symptoms, clinical situations, and drugs is commonly referred to as Clinical Coding. Among the most widely used coding systems, we can find the tenth version of the international medical glossary published by WHO (World Health Organization), commonly known as ICD-10. It contains codes for diseases, signs, and symptoms, abnormal results, complaints, social situations, and external causes of injury or illness. The United States uses its national variant of the ICD-10 called the ICD-10 Clinical Modification (ICD-10-CM). A procedure classification called the ICD-10 Procedure Coding System (ICD-10-PCS) has also been developed for the acquisition of hospitalization procedures. There are over 70,000 ICD-10-PCS procedure codes and over 69,000 ICD-10-CM diagnosis codes, compared to about 3,800 procedure codes and about 14,000 diagnosis codes in the previous ICD-9-CM. The use of an international classification to annotate medical diagnoses makes the health system interoperable between different countries. Among the many possibilities of using ICD codes, a doctor of any nationality could thus be able to read, analyze, and use the medical history of a patient even if of a different nationality. In addition, diagnostic patterns used by clinicians could be identified to improve automatic disease prediction strategies and provide automatic specialist support for decision making. These observations strongly support the need for automated systems to support clinicians to perform this task quickly and without human intervention. The contribution of our work is a novel model for ICD-10 codes annotation. We proposed a model based on Bi-LSTM and BERT to assign one or more ICD-10 codes to the medical diagnosis. Specifically, we have designed our approach as a two-step process. In the first one, we use a BERT-based classifier to select from the dataset only a subset of sentences that could be candidates for the annotation step. In particular, those phrases are them that could be annotated with one or more codes. The phrases left were generally speaking expressions. In the second step, we used a Bi-LSTM model to

analyze the candidate sentences and assign one or more codes to them. The results are encouraging and a good starting point for further investigations. The rest of the paper is structured as follows. We start analyzing related works, and we go through the description of the model and the dates. Finally, we analyze the results obtained, and we expose our consideration of the task in the conclusion section.

2 Related Work

The scientific community has long addressed the task of analyzing medical diagnoses in order to assign a unique code for each pathology. In particular, since 1973, the first corpus and state of the art analysis of the clinical coding task have been released. In 1999, Chapman stated that a system based on algorithms would be better than a human being to perform this task. If we think about the high number of codes in the ICD-10 glossary (about 70000), it immediately comes to mind that even a human being cannot be very accurate in the assignment, as long as he must know perfectly each of the countless codes. In 2006 Kukafka et al. affirm that the algorithmic path to be followed for the resolution of the task was through the NLP. It is, in fact the most suitable for the resolution of the task, currently. To get automatic systems able to afford the task efficiently, it is necessary to wait until 2017. Miftahutdinov (Miftahutdinov and Tutubalina, 2017) at CLEF eHealth 2017 uses an LSTM on a TF-IDF representation of the text to identify the most suitable ICD-10 code for the input sequence. This allows to obtain an F1 score equal to 0.85, considering a classification on 1,256 distinct classes. In 2018, Atutxa et al. (Atutxa et al., 2018), proposed a three-level sequence-to-sequence neural network-based approach. The first neural network tries to assign one set of ICD-10 codes to the whole document, then they are refined to assign one set of codes to the line, and finally one specific code. This strategy allowed the model to obtain an F1 score between 0.7086 and 0.9610, depending on the language of the dataset on which the system has been evaluated. At CLEF eHealth 2019, the best system was proposed by Sanger et al. (Sanger et al., 2019), obtaining an F1 score of 0.80. The proposed model utilized a multilingual BERT (Devlin et al., 2019) text encoding model, fine-tuned on additional training data of German clinical trials also annotated with ICD-10 codes. The model is

extended by a single output layer to produce probabilities for specific ICD-10 codes. Considering the successful models presented as state of the art, we decided to use a machine learning approach that combines CNNs, Bidirectional LSTMs, Attention Layers, and BERT.

3 Model and dataset

The CodiEsp evaluation track proposed by CLEF 2020 (Stanfill et al., 2010; Goeuriot et al., 2020; Miranda-Escalada et al., 2020) was structured as three sub-tracks about the analysis of clinical reports: CodiEsp-D requires automatic ICD10-CM [CIE10 Diagnóstico] code assignment; CodiEsp-P requires automatic ICD10-PCS [CIE10 Procedimiento] code assignment; CodiEsp-X requires to submit the reference to the predicted codes (both ICD10-CM and ICD10-PCS). The correctness of the provided reference is assessed in this sub-track, in addition to the code prediction. We decided to create a model based on deep learning able to deal with the first two subtasks, without performing an operation of reference span detection as required in the third task. An high level architecture of the model is described in Fig. 1. Specifically, we decided to use a BERT-based classifier to perform a pre-filtering operation in order to select a subset of sentences possibly referring to a clinical state. Later, the candidate sentences are submitted to a classifier based on BiLSTM, CNN, and self-attention to assign them one or more clinical codes. As pre-trained BERT model, we decided to use BETO (Caete et al., 2020), a Spanish pre-trained version of BERT. The authors trained BETO using 12 self-attention layers with 16 attention-heads each and 1,024 as hidden size. They used all the data from Wikipedia and all of the sources of the OPUS Project (Tiedemann, 2012), having the text in Spanish. We decided not to use the multilingual version of BERT because it has been shown that a version trained on the native language performs much better in many NLP tasks (Polignano et al., 2019b; Polignano et al., 2019c). The sentences classified as possible references of clinical codes are consequently passed to the second part of our model. First of all, the sentences are encoded into word embeddings. In this step, we decided to use a FastText embedding strategy (Bojanowski et al., 2017), which proved to be more effective than GLoVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013)

when many domain-specific words occur in the dataset (Polignano et al., 2019a). For our final configuration of the model, we chose the one released by José Cañete¹ made of 300 dimensions, trained on the Spanish Unannotated Corpora² containing more than 3 billion words. The block of the model that uses BiLSTM, CNN and self attention has been already proposed by the authors of this contribution for the emotions classification task. More details about it can be found in (Polignano et al., 2019a). As model parameters of the BiLSTM architecture we decided to set the value of hidden units to 64 and the internal dropout value to 0.3. We have also decided to vary the function of activation used by the net, setting it to the hyperbolic tangent function (tanh). A level of self-attention is added following the LSTM. Consequently, we applied the CNN layer on the result of the attention algorithm. In detail we apply a 1D Convolutional network with 64 filters and 5x5 kernel. We used ReLu as activation function, that unlike the hyperbolic tangent is faster to calculate. On the top of the CNN layer, we added a Max Pooling function for subsampling the values obtained, reducing the computational load and, the number of parameters of the model. The hidden model obtained until this step has been merged with the output of the previous Bi-LSTM. After that, we used a max-pooling layer for 'flattening' the results and reduce the model parameters. Finally, another dense layer with a soft-max activation function has been applied for estimating the probability distribution of each clinical code available in the dataset. Further details about the model can be found in (Polignano et al., 2019a; Polignano et al., 2020) and the source code of the model is publicly available on GitHub³.

4 Experimental evaluation

The CodiEsp corpora contains manually annotated clinical reports, written in Spanish, with corresponding clinical codes. The training set contains 500 clinical cases, while the development and the test set provide 250 clinical cases each. The CodiEsp corpus format is plain text with UTF8 encoding, where each clinical case is stored in a single file whose name is the clinical unique case identifier. The final collection of the

¹<https://github.com/dccuchile/spanish-word-embeddings>

²<http://crscardellino.github.io/SBWCE/>

³<https://github.com/marcopoli/CODIESP-10>

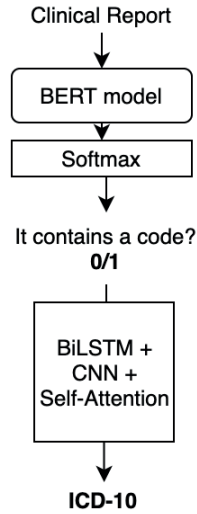


Figure 1: General design of the proposed model.

1,000 clinical cases of the corpus contains 16,504 sentences, with 16.5 sentences per clinical case on average. It contains 396,988 words, with 396.2 words per clinical report on average. The final architecture of the model, previously proposed was obtained after conducting several experiments on 20% of the training dataset released for the Codiesp-D subtask. For each task we used as classification labels only them with at least one example set in the training set. In particular we used 1788 codes for the Codiesp-D and 546 codes for the Codiesp-P task. For lacking of space, we are going to report in this contribution only the most relevant. As first experiment we trained different classification models with the purpose of directly classify the single medical reports with one ICD-10 code. In particular we developed the following models: *LSTM*, *BiLSTM*, *CNN*, *CNN + Self Attention*, *BERT*, *BiLSTM + CNN*, *BiLSTM + CNN + Self Attention*, *(Pre-filtering) BERT - (Classification) BiLSTM + CNN + Self Attention*.

Analyzing the results in Tab. 1, it is possible to notice that considering the F1 score as a metric, models based on deep learning approaches with LSTM and CNN strategy are able to obtain very similar results. It is evident that the differences between these methodologies are minimal and that generally, a combination of them improves the performance. Starting from an F1 measure of 0.09789 obtained using a biLSTM layer, a score of 0.10410 is reached when combining the biLSTM, CNN, and self-attention layers. The BERT-based

Table 1: Results obtained holding the FastText SUC word embedding, and varying the model.

	Macro-P	Macro-R	Macro-F1
LSTM	0.09357	0.09903	0.08845
BiLSTM	0.10354	0.10909	0.09789
BiLSTM + CNN	0.09995	0.11552	0.09831
BiLSTM + CNN + SelfAtt.	0.10629	0.11887	0.10410
CNN	0.09511	0.10095	0.09100
CNN + SelfAtt.	0.09279	0.09484	0.08706
BERT (BETO)	0.10381	0.10821	0.10294

classification model requires particular attention. It succeeds, in fact, to obtain a score of F1 comparable to that obtained by the model that combines the single techniques (BiLSTM+CNN+Self.Att). Thus, we considered good candidates for the final classification model, both BERT and the BiLSTM+CNN+Self.Att. models. The second step of experimentation was to understand if using a classifier that performs the task of pre-filtering diagnoses not containing an ICD-10 code could help the classification performance. Tab. 2. shows how we decided to test the combinations of the two models previously chosen as candidates with a pre-filtering approach followed by a classification approach. Observing the results in terms of F1 measure, it is possible to observe that the model using BERT in the first phase of selection and BiLSTM + CNN + Self.Att. in the phase of choice of the final codes, is the one that leads to better results. We were thus able to increase the F1 score by about 0.03 points compared to the previous step, reaching the value of 0.13632. Finally, we decided to use a threshold on the result of the last layer of the here proposed model (dense layer with a softmax function) in order to extract more than one label for each medical report. We performed different experiment in order to decide this threshold and finally we used a value of 0.10 able, from our evaluation, to maximize the F1 score of the model, reaching a score of 0.16011. The model we implemented, has been used for participating at both CodiEsp subtasks, i.e., CodiEsp-D and CodiEsp-P.

5 Conclusion

The ability to automatically annotate medical reports with international codes is an open and relevant research challenge for future technologies of global medical data sharing. In this work, we have

Table 2: Results obtained holding the FastText SUC word embedding, the models (BERT, BiLSTM + CNN + SelfAttention) and varying their combinations for pre-filtering and classification.

	Macro-P	Macro-R	Macro-F1
(Pre-filtering) BiLSTM + CNN + SelfAtt. – (Classification) BiLSTM + CNN + SelfAtt.	0.13241	0.10934	0.11534
(Pre-filtering) BiLSTM + CNN + SelfAtt. – (Classification) BERT (BETO multi-class)	0.09180	0.08871	0.10022
(Pre-filtering) BERT (BETO) – (Classification) BERT (BETO multi-class)	0.09704	0.10092	0.11734
(Pre-filtering) BERT (BETO) – (Classification) BiLSTM + CNN + SelfAtt.	0.13823	0.12053	0.13632

proposed a model based on the state of the art technologies to support the doctors during this time-consuming task. For example we can imagine the use of our model in a system that can suggest to the doctor a set of possible codes to choose from to note down the medical diagnosis. Such a suggestion could make this annotation process less complex for any human annotator. It is important to keep in mind that the ICD-10 code contains more than 70,000 codes and their number is constantly increasing. During the competition it was shown that the large number of possible labels was also the most difficult problem to deal with even for automatic systems. In this regard, in fact, many systems based on machine learning have found many more difficulties than hybrid systems able to select through linguistic rules a subset of possible assignable codes. This observation will be used by us in the future to improve the model proposed here. Currently, the results obtained are encouraging and confirm the possibility of tackling it with current technologies. Nevertheless, our system cannot obtain a score of reliability, such as acting independently, and the opinion of a human coder is still essential. The system proposed here has not won the challenge at CLEF 2020 but is still a good starting point for further studies that want to use those technologies to resolve the clinical coding task.

Acknowledgment

This work is funded by project "DECiSION" codice raggruppamento: BQS5153, under the Apulian INNONETWORK programme, Italy.

References

- Aitziber Atutxa, Arantza Casillas, Nerea Ezeiza, Víctor Fresno, Iakes Goenaga, Koldo Gojenola, Raquel Martínez, Maite Oronoz Anchordoqui, and Olatz Perez-de Viñaspre. 2018. Ixamed at clef ehealth 2018 task 1: Icd10 coding with a sequence-to-sequence approach. In *CLEF (Working Notes)*, page 1.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Jos Caete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Prez. 2020. Spanish pre-trained bert model and evaluation data. In *to appear in PML4DC at ICLR 2020*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Saez Gonzales, Marco Viviani, and Chenchen Xu. 2020. Overview of the CLEF eHealth evaluation lab 2020. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsirikaki, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurlie Nvol, and Linda Cappellato and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*, LNCS Volume number: 12260.
- Zulfat Miftahutdinov and Elena Tutubalina. 2017. Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. In *CLEF (Working Notes)*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estap, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings.

- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019a. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019b. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CLiC-it*.
- Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019c. ALBERTo: Modeling Italian Social Media Language with BERT. *Italian Journal of Computational Linguistics - IJCOL*, -2, n.2.
- Marco Polignano, Vincenzo Suriano, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2020. A study of machine learning models for clinical coding of medical reports at codiesp 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurlië Nvol, editors, *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. CEUR.org.
- Mario Sanger, Leon Weber, Madeleine Kittner, and Ulf Leser. 2019. Classifying german animal experiment summaries with multi-lingual bert at clef ehealth 2019 task 1. In *CLEF (Working Notes)*.
- Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersch. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.
- Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Exploring Attention in a Multimodal Corpus of Guided Tours

Andrea Amelio Ravelli[◊], Antonio Origlia[•], Felice Dell’Orletta[◊]

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)

ItaliaNLP Lab - www.italianlp.it

{andreaamelio.ravelli, felice.dellorletta}@ilc.cnr.it

[•] University of Naples “Federico II”

antonio.origlia@unina.it

Abstract

This paper explores the possibility to annotate engagement as an extra-linguistic information in a multimodal corpus of guided tours in cultural sites. Engagement has been annotated in terms of gain or loss of perceived attention from the audience, and this information has been aligned to the transcription of the speech from the guide. A preliminary analysis suggests that the level of engagement correlates with some specific linguistic features, opening up to possible future exploitation.

1 Introduction

Understanding a message expressed through the speech channel in face-to-face interactions involves more than the ability to decipher a string of characters and to assign a meaning to words and sentences. The linguistic information conveyed by lexicon is only the tip of the iceberg: intonation, gesture, facial expression, gaze, body movement play a key role in spoken communication. By summing the information in all these complementary modalities acquired through different channels (i.e. auditory and visual systems), the human brain is capable to analyse and decode a message not only on the basis of the words it contains. Moreover, the vision modality enables the speaker to evaluate the effectiveness of his/her message on the audience. In fact, face-to-face interactions offer the possibility to have an on-line feedback from the addressee even without an ongoing active dialogue. Simply by interpreting unconscious signals accessible from the vision modality, such as body postures and movements, facial expressions, eye-gazes, the speaker can understand if the addressee

is engaged with the discourse, and continuously fine-tune his/her communication strategy in order to keep the attention high in the audience.

Engagement can be explained as the process by which two or more actors establish, maintain and end their perceived connection during interactions they jointly undertake (Rich et al., 2010). It is composed of a series of verbal and non verbal behaviours, useful to understand the involvement between the actors, and specifically between the actors and the content of their communication scene, and it can be used to provide evidence of the waning of connectedness (Sidner et al., 2005).

In this work we describe a pilot annotation of audience engagement during guided tours in cultural sites, by evaluating the observable behaviours of the visitors in response to the speech from the guide. The main goal is to trace the level of attention of the visitors. Engagement is defined as a multidimensional meta-construct (Fredricks et al., 2004), and attention is considered a component of its the visible cues.¹ The paper is organised as follows: section 2 introduces the CHROME project and its multimodal corpus; section 3 describes the visual annotation; section 4 reports the results of the annotation in terms of agreement and some linguistic analysis on the available set of aligned transcriptions; section 5 concludes with some discussions on possible future works and exploitation for this kind of resource.

2 The CHROME Project

The Italian national project Cultural Heritage Resources Orienting Multimodal Experience (Origlia et al., 2018) aims at developing a data collection and annotation procedure to support the develop-

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Per definition, cognitive engagement refers to internal processes, whereas only the emotional and behavioral components are manifested in visible cues. Nevertheless, all engagement elements are highly interrelated and do not occur in isolation (Fredricks et al., 2004). Thus, attention plays a crucial role (Goldberg et al., 2019).

ment of new interactive technologies for cultural heritage. The project concentrates on the three Campanian Charterhouses: an integrated description of these from different point of views (textual, behavioural, geometrical, etc...) is being developed. In the framework of this project, a data collection campaign to document how professional guides present architectural heritage contents when on-site was defined.

2.1 The CHROME multimodal corpus

The collected data consist of audiovisual recordings involving three art historians with strong experience in accompanying groups of visitors. Given the limited number of informants considered in the CHROME project, only female experts were recruited to remove gender effects in multimodal and linguistic analysis.

Recorded data include two Full-HD video recordings: the first one is a fixed shot of the art historian, taken from a position immediately next to the attending group, while the second one is a fixed shot of the group of recruited visitors. A close-range digital microphone with background noise cancellation is used to record the guide's voice.

Each recruited expert accompanied four groups of four people in an hour long guided tour at the San Martino Charterhouse in Naples. Recruited members of the audience vary on a socio-demographic basis and each group is gender balanced. The visit is divided into six points of interest (POIs), selected as the most relevant parts of the Charterhouse from an architectural and artistic point of view:

- *Pronaos*: outside the doorstep of the church. The introductory part of the visit is recorded in this POI. Environmental elements mainly consist of architectural details;
- *Great cloister*: a large external place, near the monks' cemetery. Further details about the monks' life are given. Environmental elements consist of the natural setting of a large garden and of the cemetery elements (e.g. *memento mori*);
- *Parlor*: the first internal setting. Specific details about the Charthusians' rules are given here. Environmental elements mainly consist of frescoes;

- *Chapter hall*: next to the parlor. Specific details about the Charthusians' order are given here. Environmental elements mainly consist of frescoes;
- *Wooden choir*: inside the church, behind the altar. The history of the church decoration process is given here. Environmental elements consist of both architectural details (e.g. the choir and the harmonic chassis) and artistic elements (frescoes and statues);
- *Treasure hall*: deeper inside the complex. Details about the relationship between the monks and the different governing parties in Naples are given. Environmental elements mainly consist of architectural details.

The selected POIs allow us to capture the social behaviour visitors and gatekeepers exhibit to negotiate the approach to the visit and to document postural and gestural behaviour of an art historian presenting a complex environment.

Videos and audio recordings are synchronised *a posteriori* using a visual-acoustic marker. Linguistic and multimodal annotations, performed on the synchronised versions of the collected material, are merged and aligned using the ELAN software (Wittenburg et al., 2006). An ELAN project file is produced for each POI visit in order to allow cross-domain research and closed vocabularies for the label sets belonging to each annotation domain are used to ensure consistency. Specifically about linguistic annotations, the considered levels consisting of word, syllable and phone level transcriptions are obtained using WebMAUS (Kisler et al., 2017) and manually checked by human experts. Also, tonal units are manually marked by a human expert, as well as syntactic structures.

3 Engagement annotation

A subset of data from the CHROME Project has been used for this work. More specifically, we acquired data for one guide accompanying four different groups of visitors in the Charterhouse of St. Martin in Naples, consisting in 24 video couples (aligned videos of both the guide and the audience, one couple for each POI). Annotation has been performed by two annotators by means of PAGAN annotation web-based platform (Melhart et al., 2019), which enables the users to easily align and play two videos. Annotators have been

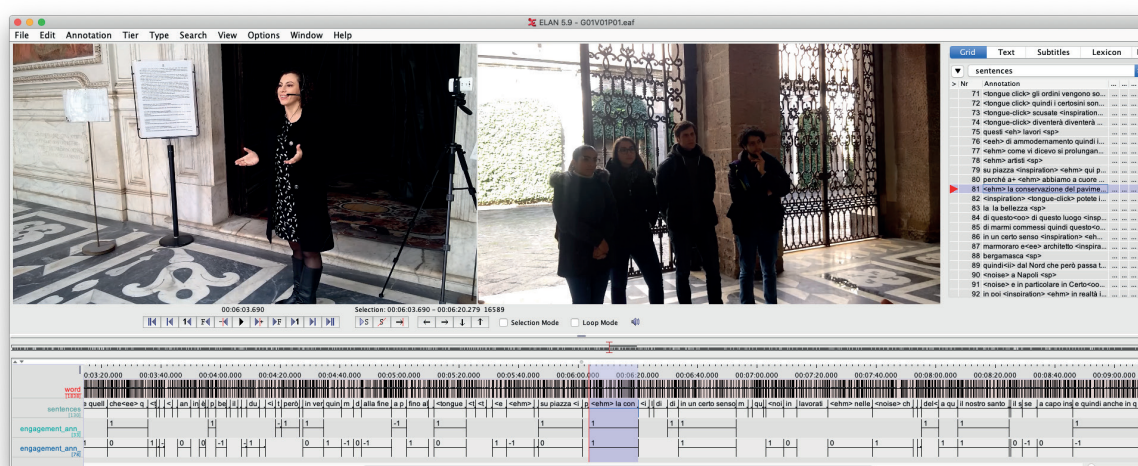


Figure 1: Elan project file with the alignment of the annotations.

asked to recognise signals of gain or loss of attention in the audience, and they recorded their observations through simple interactions with the up and down keys of the keyboard, where up stands for a gain and down for a loss in attention. Given the nature of the annotation (and the scope of this pilot work), no strict instructions have been delivered to the annotators. They based their judgement on visible cues of perceivable variation in the level of attention from the group of visitors, such as gaze following a deictic gestures, facial expressions as feedback to the guide’s speech, head movements, pose and so on. The interactions in PAGAN are recorded using RankTrace framework (Lopes et al., 2017), and the whole annotation session is exported as a tab-separated file containing continuous series of milliseconds and values for each interaction. In total, the set of videos consists of $\sim 3:20$ hours, with an average length of $\sim 8:40$ minutes per point of interest.

For 3 of these videos it was already available² the ELAN project file containing the orthographic transcription of the guide’s speech (more specifically, the speech from the visit in the POI 1 with the first three groups), thus it has been possible to automatically align the visually-derived annotation, using the pypmli Python Module (Lubbers and Torreira, 2018).

Figure 1 shows an example of the alignment for one of the videos in an ELAN project file. Using these alignments it has been possible to investigate

²The transcription and annotation of the whole corpus of the CHROME Project is still an ongoing work, thus completely annotated and aligned data is still limited.

if any correlation exists between linguistic features extracted from the guide’s speech and engagement from the visitors.

4 Evaluation of the corpus

Video	Length	Spearman’s rho
Group 1		
POI 1	00:11:33	0.94
POI 2	00:08:42	0.83
POI 3	00:05:17	0.70
POI 4	00:05:46	0.87
POI 5	00:06:47	0.72
POI 6	00:10:08	0.94
Group 2		
POI 1	00:13:12	0.98
POI 2	00:08:45	0.91
POI 3	00:05:24	0.39
POI 4	00:06:25	0.92
POI 5	00:08:09	0.83
POI 6	00:12:08	0.43
Group 3		
POI 1	00:16:18	0.98
POI 2	00:10:43	0.98
POI 3	00:07:38	0.98
POI 4	00:08:43	0.90
POI 5	00:05:40	0.89
POI 6	00:13:07	0.99
Group 4		
POI 1	00:02:35	0.93
POI 2	00:10:20	0.98
POI 3	00:07:17	0.89
POI 4	00:07:21	0.97
POI 5	00:05:52	0.98
POI 6	00:11:10	0.98
AVG	00:08:42	0.87

Table 1: Correlations on the annotations for each video.

To evaluate the agreement and thus the reliability of the annotation, we calculated the Spearman’s rho for the continuous series of values from the two annotators. Table 1 reports the results of the correlations: the overall agreement is significantly high, with a average correlation between the two series of 0.87. Figure 2 and 3 shows respectively the plot for highest and lowest correlation.

Linguistic Feature	Ann 1		Ann 2	
	Positive	Null	Positive	Null
	Avg (St.Dev)	Avg (St.Dev)	Avg (St.Dev)	Avg (St.Dev)
n.tokens	19.78 (14.63)**	10.42 (9.79)**	16.68 (13.78)**	5.68 (5.72)**
% NOUN	15.97 (9.69)	17.32 (14.32)	16.6 (10.09)	16.54 (16.98)
% PROPN	4.48 (11.7)*	4.24 (9.9)*	4.99 (11.24)**	4.12 (11.82)**
% PRON	7.65 (8.04)**	6.77 (11.85)**	8.13 (12.65)**	4.58 (9.88)**
% VERB	11.33 (9.2)*	12.2 (18.07)*	11.12 (9.71)*	16.91 (27.33)*
% AUX	5.87 (7.19)**	5.07 (12.12)**	5.73 (11.64)**	4.63 (13.38)**
% ADJ	3.94 (5.04)	5.06 (10.91)	4.69 (6.43)**	4.07 (14.18)**
% ADV	14.14 (13.49)**	13.55 (20.19)**	13.12 (14.75)**	12.97 (22.68)**
% DET	15.49 (13.99)	14.74 (12.73)	15.75 (10.42)**	13.66 (17.5)**
% NUM	0.32 (1.35)	0.42 (2.45)	0.52 (2.36)**	0.21 (2.29)**
% CCONJ	4.85 (15.34)**	2.48 (8.16)**	2.7 (5.77)**	4.21 (16.25)**
% SCONJ	2.48 (3.52)**	3 (11.46)**	2.05 (4.29)**	2.97 (11.97)**

Table 2: Average and Standard deviation of Profiling-UD linguistic features for POS distributions (in percentage) and number of tokens per sentence. * $p < 0.05$; ** $p < 0.01$.

Such information can be used to extract meaningful segments concerning the level of attention (e.g. for machine learning purposes).

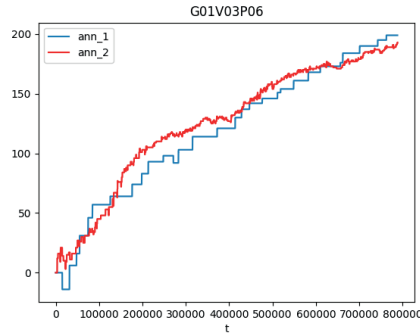


Figure 2: Plot of annotations for the video with the highest correlation (Spearman's rho: 0.99).

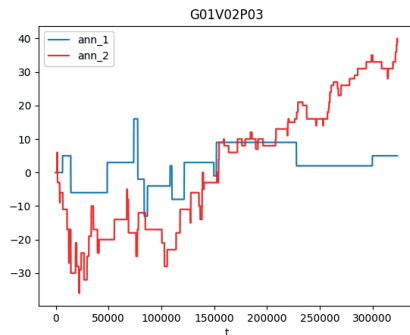


Figure 3: Plot of annotations for the video with the lowest correlation (Spearman's rho: 0.39).

4.1 Linguistic features correlation

As briefly mentioned before, we exploited the corpus composed of available orthographic transcriptions to carry out some analysis about the possible correlation between content of the speech and the perceivable engagement of the audience. To do so, we considered pause tags, i.e. short and long pauses (respectively, $\langle sp \rangle$ and $\langle lp \rangle$), as boundaries for sentence-like units of text to be processed along with the corresponding engagement value. We are aware that breath groups cannot be considered as reference units for the analysis of speech,³ and that applying written language methodologies and tools to spoken modality is biasing (Linell, 2005; Linell, 2019), but for the scope of the present work it has been necessary to make use of the available segmentation.

Even if we had few text available (3 transcriptions, for a total of 5,648 tokens in 464 sentences; ~ 12 tokens per sentence), we analysed the corpus using Profiling-UD⁴ (Brunato et al., 2020), a web-based application that performs linguistic profiling of a given text. The output of Profiling-UD is a tab-separated file, with one row per document (one for sentence, in this case) and one column for each of the 122 linguistic features analysed by the system. The objective is to investigate

³Segmentation of speech in basic units is still an open challenge in spoken language studies, as recently testified by Izre'el et al. (2020) and Mello et al. (2020).

⁴<http://linguistic-profiling.italianlp.it>

if any relation could be traced between the perceived attention from the audience and the linguistic features extracted from the guide's speech. We observed the scores for the sentences marked with a gain of attention against those for which annotators did not interact with the platform (i.e. those sentences that, aligned with time stamps to the series of the annotations, was not marked as gain or loss of attention). We performed the Wilcoxon rank sum test on features values for the two groups of sentences (positive vs. null) for both the annotators.

Table 2 reports average and standard deviation for the linguistic features with $p < 0.05$ for at least one annotator.⁵ It is possible to notice that, among positive and null marked sentences in both the annotator's data, the feature that significantly varies more than the other are the length of sentences (n_tokens) and the distribution of auxiliars.

The correlation between length and attention is not surprising, since longer sentences are likely to be more informative and thus probably more engaging. Even if sentence length is normally associated to a higher sentence complexity (Brunato et al., 2018), other typical features of complexity are not appreciably, given that subordinative conjunctions (SCONJ) are sensibly lower in higher attention marked sentences, while coordinative conjunctions (CCONJ) shows opposite trend in the two groups. For both the groups proper names (PROPN) and pronouns (PRON) seem to characterise engaging sentences.

5 Conclusions and Future Works

In this work we introduced a pilot annotation of visually perceivable attention, meant as a component of engagement, and its alignment in a multimodal corpus of guided tours in cultural sites. Moreover, we analysed the available speech transcription for 3 of the 24 videos and, notwithstanding the small dimension of the corpus (~5K tokens), some signal of the connection between attention and specific lexical features emerges, and it would be interesting to augment data in terms of annotations and alignment in order to extensively verify these correlations. Much more reliable analysis may be carried on by exploiting bet-

⁵In this analysis we consider exclusively features on sentence length and part-of-speech distributions. Profiling-UD is a tool designed for written text and not trained to work on speech transcriptions, thus any significance on syntactic features is not reliable.

ter textual segmentation, e.g. tonal units, and fine-tuning the feature extraction procedure in order to better handle spoken language. In this way, it would be possible to account also spoken-specific peculiarities and correlate them to audience engagement.

Finally, in the specific context of hosting and guiding visitors in cultural sites, the possibility to trace the level of engagement during tours can open up to interesting outcomes. In this regard, aligning speech transcription with attention tracking and other data, such as gaze, intonation, gesture, facial expression, body movement (for both the speaker and the addressee), would be particularly useful to train a classifier to recognise engaging information both in spoken language and in videos.

References

- Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.
- Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.
- Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109.
- Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2019. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review*, 35(1):463–23, January.
- Shlomo Izre'el, Heliana Mello, Alessandro Panunzi, and Tommaso Raso. 2020. *In Search of Basic Units of Spoken Language*, volume 94 of *A corpus-driven approach*. John Benjamins Publishing Company, Amsterdam, June.
- Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.
- Per Linell. 2005. *The Written Language Bias in Linguistics. Its Nature, Origins and Transformations*. Routledge.

- Per Linell. 2019. The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, 76:101230.
- Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE.
- Mart Lubbers and Francisco Torreira. 2018. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. <https://pypi.python.org/pypi/pympi-ling>. Version 1.69.
- David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. Pagan: Video affect annotation made easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE.
- Heliana Mello, Lúcia Ferrari, and Bruno Rocha. 2020. Multimodality, Segmentation and Prominence in Speech. *Journal of Speech Sciences*, 9:1–6.
- Antonio Origlia, Renata Savy, Isabella Poggi, Francesco Cutugno, Iolanda Alfano, Francesca D’Errico, Laura Vincze, and Violetta Cataldo. 2018. An Audiovisual Corpus of Guided Tours in Cultural Sites - Data Collection protocols in the CHROME Project. *JOWO*, 2091.
- Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. 2010. Recognizing engagement in human-robot interaction. In *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE.
- Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559.

A Case Study of Natural Gender Phenomena in Translation

A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish

Argentina Anna Rescigno¹, Eva Vanmassenhove², Johanna Monti¹, Andy Way³

¹UNIOR NLP Research Group, University of Naples L'Orientale

²Department of CSAI, Tilburg University ³ADAPT Centre Dublin City University

a.rescigno1@studenti.unior.it, e.o.j.vanmassenhove@tilburguniversity.edu, jmonti@unior.it, andy.way@adaptcentre.ie

Abstract

This paper presents the results of an evaluation of Google Translate, DeepL and Bing Microsoft Translator with reference to natural gender translation and provides statistics about the frequency of female, male and neutral forms in the translations of a list of personality adjectives, and nouns referring to professions and bigender nouns. The evaluation is carried out for English→Spanish, English→Italian and English→French.

1 Introduction

Gender manifests itself in a language in many ways, and different languages use different linguistic devices to mark (or sometimes 'not mark') gender. When dealing with language, three types of gender come into play: natural gender, grammatical gender and social gender. Natural gender is generally based on the sex of a person or an animal realised by means of the male/female polarity or on the absence of sex for neutral nouns. Grammatical gender, instead, is not always coherent with the semantic categorization of a word and can vary from language to language since it depends on the representation of objects in the world on the basis of specific properties attributed to them in a specific cultural context. Social gender is used in relation to the properties of a word on the basis of which the speakers of a language associate the natural gender of a person to a word (Hellinger and Motschenbacher, 2015): this mainly happens with names of professions, such as for instance *doctor* or *nurse* which are interpreted according to social stereotypes concerning the roles of males and females in the society. Gender is present in the data we use to train MT systems due to the demographic features of the human training data and because of the nature of stereotypes and biases we

communicate in our day-to-day communications. As most state-of-the-art MT systems handle translations on the sentence-level, gender phenomena are, usually, resolved on statistics inferred from the training data. Mistranslations of gender information occur more frequently when translating from gender-neutral languages, such as English¹, into morphological-rich languages, such as Italian or French, which explicitly mark gender and require additional information to correctly translate gender phenomena. When such additional information or context is not provided, the system will pick the most likely variant. A recent study by Prates et al. (2018) showed how Google Translate (GT) yields more male defaults than what ought to be expected when looking at demographic data on its own, alluding that there might be a phenomenon they refer to as machine bias (Prates et al., 2018; Vanmassenhove et al., 2019). In this paper, we systematically evaluate: (a) single-word queries, containing personality adjectives and profession nouns, and (b) bigender nouns² in an EN → IT, FR, ES translation setting for GT, DeepL (DL) and Bing Microsoft Translator (BMT) to verify the diversity in translations provided by these MT providers.

2 Related Work

As recent years have seen an increase in literature on bias in NLP, we focus particularly on work that has attempted controlling the seemingly random fluctuations in terms of gender in the translations provided by large MT providers, with a specific focus on neural approaches. Rabinovich et al. (2016) conducted a more elaborate series of experiments very similar to the work by Bawden et al. (2016). Their work on preserving original author traits fo-

¹Aside from pronouns such as 'she' and 'he' or some exceptions such as 'actress' vs 'actor'.

²Bigender nouns do not have a fixed grammatical gender; their gender is determined by the context and without any further context, they are valid for both male and female referents.

cuses particularly on gender. They treated personalizing PB-SMT systems as a domain-adaptation task where the female and male gender are two separate domains. In NMT, Vanmassenhove et al. (2018) experimented with the insertion of an artificial token at the beginning of the sentence, indicating the gender of the speaker (Vanmassenhove and Hardmeier, 2018). This approach is similar to Sennrich et al. (2016) who added an ‘informal’ or ‘polite’ tag indicating the level of politeness expressed to the training sentences.

The work by Elaraby et al. (2018) presents a technique for the translation of speech-like texts focusing particularly on English-to-Arabic. They train a baseline on generic data (4M sentences) and use a set of gender-labelled sentences (900k) in order to tune the system towards generating translations with correct gender agreement.

More recently, Moryossef et al. (2019) presented a simple yet effective black-box approach to control the NMT system’s translations in case of gender ambiguity. Instead of appending a token, they concatenate unambiguous artificial antecedents with information on the speaker and the interlocutors to ambiguous English sentences. Some recent studies have addressed the problem of the scarcity of publicly available corpora and create corpora specifically designed to evaluate or to test MT performance with respect to gender translation (Font and Costa-Jussa, 2019; Di Gangi et al., 2019). Finally, Monti (2020) provides an overview of outstanding issues and topics related to gender in MT and Sun et al. (2019) a literature review of work related to gender bias in the field of NLP.

3 Experimental setup

For the experiments, we compiled a dataset with the translation of both English single-word queries and short sentences into Italian (IT), French (FR) and Spanish (ES) with GT, BMT and DL. We experimented with both single-words and short sentences as e.g. GT provides, since end 2018³ gendered translations for single-word queries (limited to nouns and adjectives) for EN–FR, EN–ES and EN–IT. Similarly, BMT and DL provide users with alternative translations on the user interface. Differently from GT, these are simply alternative translations for the word in question. As such, GT is the

³<https://www.blog.google/products/translate/reducing-gender-bias-google-translate/>

only system that currently, to some extent, deals with gendered variants in a systematic way.

3.1 Compilation of Datasets

The datasets⁴ are compiled on the basis of a list of nouns and adjectives collected from different sources (see Table 1). The translations generated and their manual evaluations are also part of the datasets. The setup for this experiment consists of both words and sentences. We collected a set of 136 personality adjectives and 107 nouns of professions from three different sources and 30 of the most common bigender nouns in the Italian language. We tested these separate sets and analysed the behaviour of the major state-of-the-art MT systems, comparing the translations of the three language pairs. The first two sets of words have been assessed alone, without any context, while the last set has been examined within the sentence level.

Alongside with the number of adjectives and nouns retrieved, Table 1 provides more detailed information on the sources and the original language in which the data was retrieved.

	#	Sources
Adjectives	136	(I, 2019a); (II, 2019a); (III, 2019)
Professions	107	(I, 2019b); (II, 2019b)
Bigender	30	(Cacciari et al., 1997); (Cacciari et al., 2011) (Thornton and Anna, 2004)

Table 1: Overview of adjectives, profession and bigender nouns along with the sources from which they were retrieved

3.2 Description of the MT Systems

To evaluate the gender issues in translation we used three state-of-the-art, freely available, MT systems:

GT: Launched in 2003 as a statistical MT system, GT has switched to a NMT system in 2016 (Monti, 2017). The translations are generated at the sentence level. Since 2018, Google provides two alternatives when translating ‘ambiguous’ or underspecified English words into languages that have male\female alternatives for various languages (e.g. Italian, French and Spanish). The male\female variants are listed alphabetically (i.e. first the female variant, then the male one).

⁴available at https://github.com/argentina-res/gender_project.git

BMT: MT system owned by Microsoft that originally used a statistical approach (Monti, 2017) but more recently switched to a neural system (Almahasees, 2018). Unlike GT, BMT does not provide alternatives in the translation box itself. However, it does give synonyms in the “Other ways to say” section and provides examples of usage in the “How to use...” section, where sometimes the female form of a word is listed (by chance rather than in a consistent way).

DL: The most recent platform launched in August 2017 by a German company, DeepL GmbH. DL uses convolutional neural networks based on the Linguee database (Morán Vallejo and others, 2019). Even though it only supports nine languages (all Indoeuropean), DL is, according to a recent study, outperforming the other competitors (Morán Vallejo and others, 2019). The layout of the interface is similar to that of GT. Nevertheless, its suggestions for alternatives are not systematically morphological variants (although they are often somehow included in the alternatives provided). Underneath the actual translations, there is also a dictionary-like section when translating words in isolation.

4 Results

In this section, the results obtained with our experiments and our subsequent evaluation will be presented. A more in-depth analysis and a discussion of some concrete examples is provided in Section 4.1. All the manual evaluations were conducted in November, 2019⁵.

The **adjectives and profession nouns** were both evaluated in single-word settings. The **bi-gender nouns** were evaluated in short sentences. Bigender nouns in Italian, such as for example the nouns *giornalista*, *pianista* are by themselves not marked for gender. As such, a single-word query would not reveal any gender marking. However, the articles, adjectives, verbs, etc. that agree with these bigender nouns are (often) marked for gender. Therefore, these specific nouns were evaluated in short sentence settings.

We manually evaluated all the outputs and will report on the percentage of male (M), female (F) and ‘neutral’ (N) or ‘covered’ (C) (i.e. no explicit gender in the target language) variants for

the single-word and sentence queries. Additionally, we report on the errors (e.g. untranslated or mistranslated words).

As mentioned earlier, GT is currently the only system that provides both male and female variants when given a single-word ambiguous query (EN→FR, ES, IT). The alternatives are limited to adjectives and nouns⁶. DL and BMT provide the user with multiple alternatives underneath the primary translation for both nouns and adjectives but not in a systematic way. As such, these alternatives do not necessarily consist of morphological variants in terms of gender.

For our evaluation, only the main translations offered by the MT systems were considered, i.e. we evaluated both gendered forms for GT, but did not evaluate the list of alternative translations provided by BMT and DL as they can be alternatives of any kind and they differ depending on the query.

In the interest of clarity and order, we will separately consider the different test-sets, single words (i.e. adjectives and nouns), and sentences with bi-gender nouns. Each set has been tested in the previously stated systems and language pairs. In the tests, we were especially investigating – apart from any translation error – the occurrence of female forms, to see if there is somehow a bias towards the gender. From a practical point of view, we will consider only the first output as a valid result, since the systems also give “alternatives” for single-word translations. For Bing, it is quite straightforward: adjectives do not present alternatives, while nouns sometimes do. Google Translate produces now two different results, marked with the gender (feminine/masculine, in this sequence for alphabetical order). DeepL, in particular, provides at least three alternative translations, for both nouns and adjectives. However, even though we are not considering alternatives, we will explain or anyway mention the ones that are significative in this research. All the results have been recorded in November 2019. However, the systems continually improve, so results may vary also in a short time.

Table 2 presents the results for the single-word translations consisting of adjectives: for all three systems, the male variant was the most common. This was especially so for BMT, where only 1.5% of the adjectives were translated into a female variant. The ‘other’ category consists of translations:

⁵The translations were evaluated by a native Italian speaker/linguist for Italian, and by a professional linguist for French and Spanish.

⁶In the target languages certain verbs are also sometimes marked for gender, e.g. reciprocal verbs, passive constructions, past participles...

(a) that were ambiguous (e.g. *mean* was translated as a verb instead of as an adjective), (b) words that were not translated by the systems and (c) errors.

ADJ	GT	BMT	DL
F	37.3	1.5	22.8
M	39.2	58.8	45.6
N	20.7	33.1	26.5
Other	2.8	6.5	5.1
Total	100	100	100

Table 2: Results in % for male (M), female (F) and neutral (N) adjectives generated for EN → IT for GT, BMT and DL. The “Other” label includes all results obtained that do not correspond to the “adjective” category

We ought to note that, none of the single words were mistranslated by GT. The only single-word queries that caused a divergence between male and female forms were queries that were a compound noun in either the source or the target (e.g. *good-tempered*). They are not treated as a single unit by GT and thus the system fails to render both variants. From the Table 2, it is clear that GT performs best in terms of balanced single-word adjective translations. Table 3 presents a similar set of results but for the nouns indicating professions. Like Table 2 GT generated the most diverse translations, while BMT the least. As far as the set with sentences is concerned, we used bigender nouns from the Italian language. We used 30 common bigender nouns in two different contexts: (a) first, in a minimal sentence that would allow us to infer the gender based on the article in the target language “I am a(n).” and (b) with a referring adjective. We used *beautiful*, *efficient*, *intelligent*, *sad* and *famous*. In Table 4, our results are presented for the bigender nouns on minimal sentences (“I am a(n).”) and in combination with the aforementioned adjectives (“I am a(n) + adj...”). In the results, we oppose the translation where we added *beautiful* as an adjective as they differed considerably from the others. Table 4 presents the results for the translations generated by BMT for bigender nouns in sentences for Italian, French and Spanish. It can be noted that the male translation is the most common in the simple sentences that do not contain an adjective in all three languages. However, when adding the adjective *beautiful* to the phrase, the female forms are the most common for all three languages. An

example of such sentences is given below:

- (a) EN: “I am a pianist” (N)
IT: “Sono un pianista.” (M)
FR: “Je suis pianiste.” (N)
ES: “Soy pianista.” (N)
- (b) EN “I am a beautiful pianist.” (N)
IT “Sono **una bellissima** pianista.” (F)
FR “Je suis **une belle** pianiste.” (F)
ES “Soy **una hermosa** pianista.” (F)
- (c) EN “I am a famous pianist.” (N)
IT “Sono **un famoso** pianista.” (M)
FR “Je suis **un** pianiste célèbre.” (M)
ES “Soy un pianista **famoso**.” (M)

The results obtained for DL are very similar to the ones obtained with BMT except for the fact that DL generates overall more female forms than BMT. Interestingly, among all systems, GT is the most biased towards male forms when evaluating entire sentences for all three language pairs, with the male forms being the dominant ones for all categories and for several set-ups we observe more than 90% male variants.

NOUN	GT	BMT	DL
F	35.8	0.9	7.5
M	46.1	60.4	60.4
N	17.6	28.3	28.3
Other	0.6	10.5	3.7
Total	100	100	100

Table 3: Results in % for male (M), female (F) and neutral (N) nouns generated for EN → IT for GT, BMT and DL. The “Other” label includes all results obtained that do not correspond to the “noun” category

4.1 Analysis

In the analysis we compare the results of the three systems, comparing the occurrences of the female-gendered translated forms in terms of the different systems and the different languages.

GT: for the single nouns and adjectives, GT provides both male and female forms. The system, however, produces more male outputs as sometimes the alternative is not provided (e.g. for compound nouns). One of the provided nouns was ambiguous (*printer*) in English. The system translated this as the object instead of the profession.

BMT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	10.0	86.7	Q*	10.0	63.3	26.7	3.3	66.7	30.0
beautiful	63.3	36.7	0.0	43.3	56.7	0.0	66.7	33.3	0.0
other adj.	13.3	83.3	Q*	3.3	96.7	0.0	6.7	93.3	0.0
DL	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	30.0	70.0	0.0	20.0	63.3	16.7	3.3	76.6	20.0
beautiful	83.3	16.7	0.0	73.3	26.7	0.0	96.7	3.3	0.0
other adj.	53.3	43.3	Q*	13.3	83.3	3.3	6.7	93.3	0.0
GT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	6.7	93.3	0.0	6.7	90.0	3.3	3.3	66.7	30.0
beautiful	43.3	56.7	0.0	80.	20.0	0.0	80.0	20.0	0.0
other adj.	3.3	96.7	0.0	3.3	96.7	0.0	3.3	96.7	0.0

Table 4: Results in % for male (M), female (F) and neutral (N) forms generated for EN → IT, FR and ES for BMT, DL and GT

Whenever an ambiguous word was translated accurately, yet not in the way we intended it to be translated, we included it into the ‘other’ category. Considering the adjectives, we observed one incorrect translation where the adjective *supportive* was translated into an Italian noun (*‘supporto’* meaning *support*). Two other adjectives were ambiguous (*mean* and *kind*) and were translated into a verb and a noun respectively by GT. For the sentence-evaluation, GT has the strongest preference for translations using male-endings.

BMT: for the nouns, BMT has a strong tendency to output male variants. The only two exceptions are *nurse* for IT, FR and ES and *makeup artist* for FR and ES. Besides, words such as *newsreader*, *translator*, *warder* were not translated by BMT and others were mistranslated, e.g. *garbage man* and *window cleaner*, for which the translations provided were too literal in IT (*uomo spazzatura*) and *pulizia finestre* where “pulizia” is the equivalent of “cleaning”). Concerning the adjectives, unlike DL and GT, BMT rarely offers alternatives and the majority of the translations generated are in the male form. Exceptional female variants were found in: (a) Italian for: *devious/subdola* and *joyful/gioiosa*; (b) Spanish for: *artistic*, *bossy*, *calm*, *diplomatic*, *dynamic*, *extrovert*, *humorous*, *industrious*, *placid*. We ought to note that a considerable amount of the adjectives (33.1% for Italian and 29.4% for Spanish) have a translation with a bigender adjective. These words have the same form for both genders and are included in the ‘N’ (neutral) group. A small number of adjectives were translated with an expression, such as *good-tempered*

→ IT *di buon umore*, FR *de bonne humeur*, SP *de buen humor*. These expressions can be assigned to both male/female referents and are thus considered covered/neutral. We did not observe any errors except for the adjective *frank* (*without a capital letter*), which was left untranslated by DL (as if it were the first-name “Frank”). However, the most appropriate translation was among the alternatives suggested. Moreover, sometimes the adjectives are ambiguous, therefore the system has opted for a “non-gendered” alternative (e.g. *mean* was translated as a verb instead of an adjective).

DL: DL provides multiple options, for both nouns and adjectives, but for only 7.5% of nouns the female form is the first result, as, for example, *assistant* and *nurse* for French and Italian, *doctor*, *secretary* and *shop assistant* for Spanish and Italian, *soldier* and *teacher* for Italian. For the adjectives, instead, the number of female forms increases to 22.8%. Similarly to some of the observations for BMT and GT, some of our intended nouns were ambiguous (‘model’ can be a noun or a verb) and the system opted for the verb translation. The only error we observed is the translation of *tailor* which was incorrectly translated into the adjective *sartoriale*.

5 Conclusions and Future Work

In future work, we would like to conduct a larger evaluation comprising of more language pairs and a more diverse set of words. Furthermore, we aim to compile a challenge set focusing specifically on gender phenomena in language that can be used and automatically evaluated. We also envisage training our own state-of-the-art MT system to verify how and whether machine bias indeed influences the output of the translations generated.

References

- Zakaryia Mustafa Almahasees. 2018. Assessment of google and microsoft bing translation of journalistic texts.
- Rachel Bawden, Guillaume Wisniewski, and Hélène Maynard. 2016. Investigating gender adaptation for speech translation.
- Cristina Cacciari, Manuel Carreiras, and Cristina Barbolini Cionini. 1997. When words have two genders: Anaphor resolution for italian functionally ambiguous words. *Journal of memory and language*, 37(4):517–532.

- Cristina Cacciari, Paola Corradini, Roberto Padovani, and Manuel Carreiras. 2011. Pronoun resolution in Italian: The role of grammatical gender and context. *Journal of Cognitive Psychology*, 23(4):416–434.
- Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.
- Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to English-Arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6, April.
- Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.
- Marlis Hellinger and Heiko Motschenbacher. 2015. Gender across languages. the linguistic representation of women and men, volume 4. Amsterdam & Philadelphia.
- Adjectives Sources I. 2019a. Personality adjectives source i. <https://www.esolcourses.com/content/exercises/grammar/adjectives/personality/words-for-describing-personality.html>.
- Professions Source I. 2019b. Professions source i. <https://www.scribd.com/doc/82021393/List-of-Common-Jobs>.
- Adjectives Source II. 2019a. Personality adjectives source ii. <https://www.esolcourses.com/content/exercises/grammar/adjectives/personality/more-words-for-describing-personality.html>.
- Professions Source II. 2019b. Professions source ii. <https://www.vocabulary.cl/Basic/Professions.htm>.
- Adjectives Sources III. 2019. Personality adjectives source iii. <https://7esl.com/adjectives-that-describe-personality/>.
- Johanna Monti. 2017. Questioni di genere in traduzione automatica. In *Al femminile. Scritti linguistici in onore di Cristina Vallini*, pages 411–431. Cesati.
- Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge.
- Alberto Morán Vallejo et al. 2019. The translation of Spanish agri-food texts into English and Italian using machine translation engines: A contrastive study.
- Amit Moryossef, Roei Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. *arXiv preprint arXiv:1903.03467*.
- Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2018. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, pages 1–19.
- Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July.
- ANNA Thornton and M Anna. 2004. Mozione. *La Formazione Delle Parole in Italiano*, pages 218–225.
- Eva Vanmassenhove and Christian Hardmeier. 2018. Europarl datasets with demographic speaker information.
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October–November.
- Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume I: Research Track*, pages 222–232, Dublin, Ireland, August.

Multifunctional ISO standard Dialogue Act tagging in Italian

Gabriel Roccabruna¹, Alessandra Cervone^{2*}, Giuseppe Riccardi¹

¹Signals and Interactive Systems Lab, University of Trento, Italy, ²Amazon Alexa AI

`gabriel.roccabruna@studenti.unitn.it`, `giuseppe.riccardi@unitn.it`

Abstract

English. The task of Dialogue Act (DA) tagging, a crucial component in many conversational agents, is often addressed assuming a single DA per speaker turn in the conversation. However, speakers' turns are often multifunctional, that is they can contain more than one DA (i.e. "I'm Alex. Have we met before?" contains a 'statement', followed by a 'question'). This work focuses on multifunctional DA tagging in Italian. First, we present iLISTEN2ISO, a novel resource with multifunctional DA annotation in Italian, created by annotating the iLISTEN corpus with the ISO standard. We provide an analysis of the corpus showing the importance of multifunctionality for DA tagging. Additionally, we train DA taggers for Italian on iLISTEN (achieving State of the Art results) and iLISTEN2ISO. Our findings indicate the importance of using a multifunctional approach for DA tagging.

1 Introduction

Dialogue Acts (DAs), a linguistically motivated model of speakers' intentions in a conversation, play a crucial role for several conversational AI tasks. DAs have been successfully used as part of conversational agents components, for example for Spoken Language Understanding (Zhao and Feng, 2018) or Natural Language Generation, and for response generation (Hedayatnia et al., 2020). Moreover, DAs have been shown to be important

* Work done while at University of Trento, prior to joining Amazon.

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

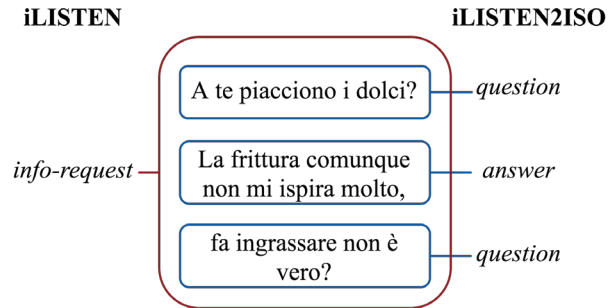


Figure 1: Example of the same turn with iLISTEN annotation versus our iLISTEN2ISO *multifunctional* annotation following the ISO standard. In this example, without the multifunctional approach a Conversational Agent would not understand that two different questions are asked.

features to learn the intentional structure of conversations (Allen and Perrault, 1980; Cervone and Riccardi, 2020; Cervone et al., 2018).

One of the bottlenecks for current research on DAs is the lack of publicly available resources with DA annotation. While this is true also for English, it is even more important for languages with fewer resources, such as Italian. For Italian, the only publicly available resource with DA annotation is currently the iLISTEN corpus (Basile and Novielli, 2018), released for EVALITA in 2018.

While useful, this resource relies on an annotation scheme which assumes only one single DA per conversational turn (see Figure 1). However, ISO 24617-2 (Bunt et al., 2010; Bunt et al., 2020), the latest accepted standard for DA annotation, posits that conversational turns can be multifunctional in a sequential way, i.e. speakers' turns can be composed of multiple DAs in sequence (Huang, 2017).

In this work, we investigate the task of multifunctional DA tagging in Italian. The contributions of this paper are: (1) we create iLISTEN2ISO, to the best of our knowledge the first

publicly available resource with DA annotation in Italian which uses a *multifunctional* approach and is *ISO-standard compliant*; (2) we present an analysis of iLISTEN2ISO showing the importance of multifunctional DA annotation; (3) we propose baseline DA tagging models for Italian trained on iLISTEN (achieving, to the best of our knowledge, SOTA results) and iLISTEN2ISO.¹

2 Related work

Dialogue act corpora Most publicly available resources with DA annotation are hardly compatible, given that each resource is typically tagged with its own different scheme tailored for a given domain (Carletta et al., 1997). This prevents both meaningful comparisons among different resources, and the possibility of experimenting with cross-corpora training of DA taggers. ISO 24617 (Bunt et al., 2010), the latest universally accepted standard for DA annotation, represents an attempt to overcome this fragmentation by providing a domain- and task-independent taxonomy, useful for both task- and non-task-oriented dialogue. Compared to previous schemes, the ISO standard is *multifunctional*, both from a sequential perspective (the same turn can contain multiple DAs in sequence) and from a simultaneous perspective (a text span can have multiple DA tags at once). Moreover, the ISO standard is a hierarchical taxonomy, rather than a flat one, which enables it to capture similarities among different tags. Sequential multifunctionality is also present in the DAMSL schema (Core and Allen, 1997), although this definition is not commonly applied to corpora that adopted DAMSL (Chowdhury et al., 2016), with the consequent possibility of introducing ambiguities and a lack of precision in understanding the communicative functions of text spans.

While for English there have recently been successful attempts to create publicly available resources mapped to ISO 24617-2 (Mezza et al., 2018); datasets mapped to ISO are scarcely available for other languages, see for example (Ngo et al., 2018) for Vietnamese and (Yoshino et al., 2018) for Japanese. For the Italian language, the only corpus with a subset of dialogues tagged with ISO in a multifunctional way is LUNA (Chowdhury et al., 2016), which is currently not publicly

available.

Dialogue Act tagging DA tagging is the task of assigning a DA tag to a given utterance in a dialogue. The definition of utterance depends on the schema used: in some schemes (Dinarelli et al., 2009), the utterance corresponds to a turn, while in others (Jurafsky, 1997) to segments of a turn. DA tagging is usually framed as text classification (Lee and Dérnoncourt, 2016; Mezza et al., 2018) or as a sequence tagging problem (Quarteroni et al., 2011; Chen et al., 2018; Colombo et al., 2020).

3 iLISTEN2ISO: Mapping iLISTEN to ISO standard

The iLISTEN corpus (Basile and Novielli, 2018) is a dataset of dyadic dialogues about food and dietary issues in Italian annotated with DAs, used during the 2018 EVALITA competition for a DA classification task. The corpus consists of 60 dialogues, with 1576 user turns and 1611 system turns. Dialogues were collected with a Wizard of Oz procedure using either written (30) or spoken (30) interactions. The system side mimics a diet therapist, asking questions about users diets or answering to users' questions. The DA schema adopted is a refined version of DAMSL (Core and Allen, 1997). As reported in Table 1, the number of DAs in the schema is 15, where 7 are reserved only to users, 6 only to the system and the remaining 2 are in common.

In iLISTEN the turn DA annotation is not multifunctional, i.e. each turn is assigned one single DA. However, not tackling the turn DAs with a multifunctional approach could result in loss of information, with the DA tag capturing only the most dominant function of a turn. In Figure 1, for example, tagging the entire turn with one DA would prevent the system from understanding that two different questions are asked.

In order to create the iLISTEN2ISO annotation, each turn from iLISTEN was annotated with a *multifunctional* approach following the ISO standard. This process involved first segmenting turns into functional units (FUs), defined as minimal stretches of communicative behaviour that have a communicative function (Bunt et al., 2010); and then annotating each FU with a DA tag. The subset of ISO schema used for mapping iLISTEN to ISO was build incrementally, since an a-priori definition was impossible due to the fact that many communicative functions were hidden by the lack

¹iLISTEN2ISO and the code of our experiments are available at: <https://github.com/BrownFortress/Multifunctional-Dialogue-Act-tagging-in-Italian>.

of segmentation. This annotation process involved user and system turns, since system turns are used as context in the prediction phase. The annotation of system turns was done only on unique turns, given the repetitiveness of system turns (only 430 of 1611 are unique).

Because of the lack of resources, the segmentation and mapping process of iLISTEN was conducted by one single annotator, under the supervision of a second annotator with previous training in ISO standard annotation. In order to ensure a reliable annotation process, after the creation of the guidelines, the second annotator repeatedly assessed a sample (100 utterances) of the annotated data. This sample was built through a stratified random sample, where for each DA tag, 20% of examples of that class was randomly sampled. This evaluation and reassessment was performed twice. In the first round, performed after the first annotation of the data, some issues regarding the usage of some DAs arose and were discussed; in the second examination, performed after the second phase of annotation of the data, no problem was found.

4 Analysis of iLISTEN2ISO

The annotation layer of iLISTEN2ISO does not change only the legacy iLISTEN schema, but also the internal structure of turns due to the segmentation process. On average in iLISTEN2ISO we have 1.61 FUs per turn, which become 1.81 on system side and 1.5 on user side if we consider them separately (this difference is justified by the fact that the system turns are on average longer than user turns). In Figure 2, we report for each legacy DA (user side), the number of segments per turn on average. Furthermore, inside the bars we list the 3 most common sequences of ISO DAs to which each legacy DA is mapped. In Table 1, we compare the number of DA tags between iLISTEN and iLISTEN2ISO. We notice that iLISTEN2ISO has a larger number of DAs in total, compared to iLISTEN. Additionally, while in iLISTEN the number of DAs in common (2) is much lower compared to either user or system DAs, in iLISTEN2ISO the common DAs between system and user are larger than the independent ones, with the advantage of potential better generalization across the two. Looking at the distribution of the ISO DAs regarding user turns, it can be noticed that the four most common DAs are: *inform* 24.5%, *ques-*

	User	System	Common	Total
iLISTEN	7	6	2	15
iLISTEN2ISO	10	2	15	27

Table 1: Number of Dialogue Acts (DAs) used by the system and the user in iLISTEN and iLISTEN2ISO (multifunctional). “Common” reports the number of DAs used by both system and user.

tion 21.3%, *answer* 15.3% and *auto-positive* 7%. Moreover, the DA distribution has a tail composed of 19 DAs with a frequency below the 5%. However, this is not a drawback of the scheme since it gives us a fine-grained representation of the actions performed by the user. Additionally, iLISTEN2ISO can be used in conjunction with any other corpus annotated with ISO standard thus, giving the possibility of augmenting the samples for a specific low-represented class.

5 Models

In this section, we describe the two baseline models for Dialogue Act (DA) classification used in our experiments. The first model is a Support Vector Machine (SVM) (Vapnik, 1995) with linear kernel, with One versus One strategy. The features used are: FastText word embeddings, Part-Of-Speech (POS) and dependency parsing tags (DEP) (retrieved using Spacy), and the previous DA tag. For word embeddings, the utterance representation is computed using the average of the relative word embeddings. The model was implemented using scikit-learn (Pedregosa et al., 2011). Hyperparameters and features selection was performed using 3 folds cross-validation. The feature vector that gave the best results for iLISTEN is the concatenation of word embeddings, POS tags, DEP tags and the previous DA. For iLISTEN2ISO, the feature vector that gave the best performances is the concatenation of word embeddings and the previous DA.

Our second model is a Convolutional Neural Network (CNN), following (Lee and Dernoncourt, 2016). The utterance representation is computed using a CNN taking as input FastText word embeddings. This representation is then concatenated with the previous DA and passed through a linear and a softmax output layer. We use cross entropy loss optimized with Adam and early stopping according to best Macro F1 on a randomly generated development set (6 dialogues), chosen for the

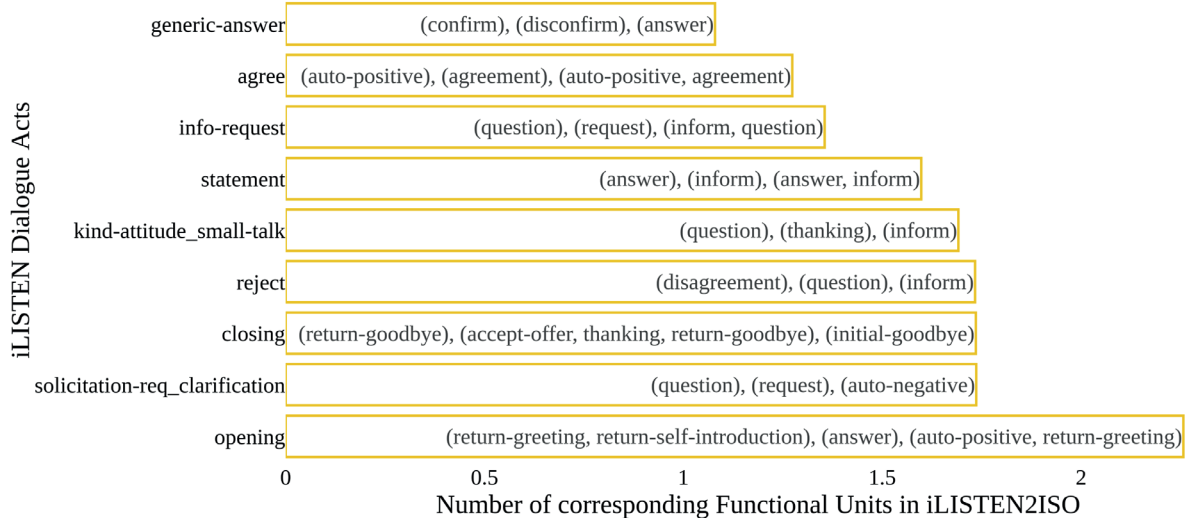


Figure 2: For each iLISTEN user DA, we report the corresponding average number of corresponding Functional Units and (inside the bars) the three most common DA sequences in iLISTEN2ISO.

lowest tags distribution difference compared to the full dataset. The learning rate is set to 10^{-3} and the batch size to 128. The number of filters is 200 and the filters sizes are 1,2,3 and 4 times the word embedding dimension (300).

6 Experiments

In this section, we report the results of Dialogue Act (DA) tagging experiments, using our proposed baselines on both legacy (using iLISTEN) and *multifunctional* ISO standard DA schemes (using iLISTEN2ISO).

Experimental setup For comparison with previous work, we follow the competition rules and report results considering only user DAs, using official splits. Additionally, we do not assume gold DAs for the context for testing (which might not be available at inference time), rather we use predicted ones. In order to do this we train a separate model for tagging system DAs used only during inference. The performances of system models are: Micro F1 96.1% and Macro F1 96.6% on iLISTEN; Micro F1 97.5% and Macro F1 96.3% on iLISTEN2ISO. For iLISTEN, the obtained classification results are compared with Unitor, the winner of the EVALITA competition (Basile and Novielli, 2018) and to the best of our knowledge the SOTA on iLISTEN (we could not perform comparisons for iLISTEN2ISO, as the code is not publicly available). Given the larger number of DA tags with few examples in iLISTEN2ISO, for comparison with the legacy scheme

Dataset	Model	Macro F1	Micro F1
iLISTEN	Unitor	63.7	73.2
	SVM	67.3	75.1
	CNN	68.0	75.0
iLISTEN2ISO	SVM	69.3	74.8
	CNN	71.4	74.9

Table 2: Results of Dialogue Act (DA) tagging using iLISTEN legacy annotation and iLISTEN2ISO multifunctional annotation.

we group the least frequent DA tags to the label “Other”. The final DA scheme for iLISTEN2ISO consists of 7 DAs. In iLISTEN the number of examples in training and testing is 1097 and 479 respectively; in iLISTEN2ISO we have 1609 and 777 respectively.

Results As shown in Table 2, our proposed models yield comparable results on both non-multifunctional (iLISTEN) and multifunctional (iLISTEN2ISO) DA tagging. On iLISTEN, our models even overcome previous SOTA performances (Unitor) on both Micro and Macro F1. We observe that while in terms of Micro F1 our models achieve very similar results on both corpora, in terms of Macro F1 they perform better on multifunctional DA tagging.

Error analysis To better understand the performance of our models on iLISTEN and iLISTEN2ISO, we look at the confusion matrices depicted in Figure 3 and Tables 3 and 4 reporting the performances computed for each DA.

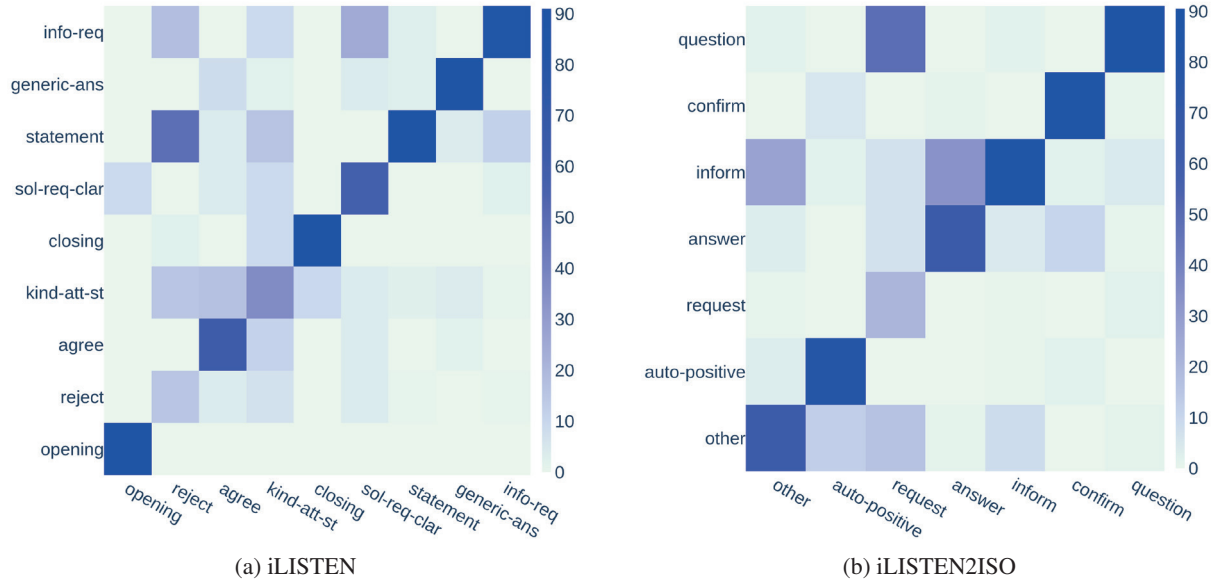


Figure 3: Confusion matrices of the CNN model on iLISTEN (a) and iLISTEN2ISO (b). The presented values are in percentage. To improve the readability of Figure (a) we used some abbreviations: sol-req-clar corresponds to *solicitation-req-clarification* and kind-att-st corresponds to *kind-attitude-small-talk*.

Dialogue Acts	F1 scores			Freq.
	Unitor	SVM	CNN	
statement	83.6	83.2	83.8	34%
info-request	80.1	81.3	82.4	23.4%
generic-ans	88.8	89.5	87.0	10.9%
kind-att-st	43.8	55.8	40.5	9.2%
reject	13.0	13.0	23.0	8.1%
agree-accept	53.6	66.7	65.2	5%
sol-req-clar.	48.9	52.0	59.6	5%
opening	100.0	90.9	95.2	2.3%
closing	73.6	73.7	75.0	2.1%
Macro F1	63.7	67.3	68.0	
Micro F1	73.2	75.1	75.0	

Table 3: This table reports the F1 results for each **iLISTEN** Dialogue Act achieved by Unitor, SVM and CNN models. All the values reported are in percentage. The last column (*Freq.*) reports the frequencies of the Dialogue Acts in the test set.

Dialogue Acts	F1 scores		Freq.
	SVM	CNN	
inform	76.6	76.3	30.5%
other	63.4	67.4	19.8%
question	84.2	85.7	17.5%
answer	77.8	68.7	13.4%
auto-positive	80.0	83.0	7.1%
confirm	83.7	87.2	6.2%
request	22.2	31.6	5.5%
Macro F1	69.9	71.4	
Micro F1	74.8	74.9	

Table 4: This table reports the F1 scores for each **iLISTEN2ISO** Dialogue Act achieved by SVM and CNN models. All the values reported are in percentage. The last column (*Freq.*) reports the frequencies of the Dialogue Acts in the test set.

Considering the CNN performance, looking at confusion matrices in Figure 3, we notice that on iLISTEN the worst class is *reject* where 48.7% of examples are predicted as *statement*. This is probably due to the similar structure of *reject* utterances to *statement* ones, while the discriminant is the semantic content that model fails to detect. This problem can be seen also in Table 3, where the *reject* DA is predicted with the worst performances among other tags. An example of error is given by the following interaction: the system says “Mangiare ad orari fissi e’ un modo per evitare di saltare i pasti e di trascurare sostanze che spesso non vengono compensate nei pasti successivi.” and the user responds “purtroppo spesso il lavoro limita la possibilità di fare una dieta sana e regolare.”. This user’s turn is tagged with *reject* but it is predicted by the model as *statement*. As it can be seen, the structure of the user’s turn is similar to a statement because the user expresses her or his opinion, in this case regarding the difficulty to follow an healthy diet.

Another interesting mismatch in iLISTEN regards *info-request*, 11.6% of which are predicted as *statement*. This is interesting because the class *info-request* is usually composed of questions, however analyzing heuristically the examples we notice that some of them contains other tags, such as answers or statements, which are hidden in the legacy annotation. In this regard, another potential source of error is the lack of punctuation as it can be seen in the utterance “è necessario fare sport per mantenersi in forma”. This utterance can be interpreted as a statement, but if a question mark is added at the end of the utterance it can be interpreted as a question. This also highlights the importance of punctuation or prosodic features in order to detect the right DA.

Another problem, that can be identified looking at the iLISTEN confusion matrix in Figure 3, is that the *kind-attitude-smalltalk* DA is confused with many different others DAs. This is due to lack of segmentation since analysing the ISO DAs distribution of the turns tagged with this tag, it emerged there is not a predominant DA. In fact, the four most common ISO DAs are: *inform* 21.3%, *question* 20.9%, *thanking* 13.5% and *auto-positive* 10.8%.

Regarding the iLISTEN2ISO confusion matrix, it can be seen that *request* is the most confused class. Indeed, 48.8% of examples are predicted

as *question*, 16.3% as *other* and only 20.9% are predicted correctly. The reason behind this performance is that the model fails to distinguish a request from a question since both of them are in a question style.

Another frequently mispredicted DA in iLISTEN2ISO is *answer*, often confused with *inform*. This is due to the fact that the model has difficulties in representing and then distinguishing the semantic content. Moreover, as it can be noticed in Table 4 this problem is more highlighted in the CNN’s rather than in SVM’s performances.

Finally, comparing the iLISTEN2ISO results presented in Table 4 with iLISTEN results presented in Table 3, it can be seen that the *question* DA is better predicted than *info-request*. In this case, only 4.4% of *question* examples are confused with *inform*. The reason of this improvement is probably the segmentation process that highlighted the multifunctionality of the utterances augmenting the specificity of the classes.

Interestingly, if we compare confusion matrices for SVM (which we decided not to include in the paper for lack of space) and CNN, shown in figure 3, we notice that the most confused classes are the same for both models across both datasets.

7 Conclusions

We presented iLISTEN2ISO, a resource for Italian multifunctional DA tagging using ISO 24617-2. We argued the importance to consider turns as a composition of multiple communicative functions, in order to preserve important semantic information. Moreover, we presented different baseline DA tagging models, on both iLISTEN and iLISTEN2ISO.

We believe the presented resource could be useful to the research community for experimenting with multifunctional DA tagging in Italian, as well as cross-corpora DA tagging. As future work, we plan to explore joint DA segmentation and classification in Italian, for example taking inspiration from the work presented by Zhao and Kawahara (2019).

Acknowledgements

The research leading to these results has received funding from the European Union – H2020 Programme under grant agreement 826266: COAD-APT.

References

- James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.
- Pierpaolo Basile and Nicole Novielli. 2018. Overview of the evalita 2018 italian speech act label (iliste n) task. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:44.
- Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. *Seventh conference on International Language Resources and Evaluation (LREC’10)*.
- Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 549–558.
- Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.
- Alessandra Cervone and Giuseppe Riccardi. 2020. Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174.
- Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *Proc. Interspeech*, pages 1011–1015.
- Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234.
- Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. 2016. Transfer of corpus-specific dialogue act annotation to ISO standard: Is it worth it? In *LREC*.
- Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *AAAI*, pages 7594–7601.
- Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*.
- Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of the 2nd Workshop on Semantic Representation of Spoken Language*, pages 34–41. Association for Computational Linguistics.
- Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *arXiv preprint arXiv:2005.12529*.
- Yan Huang. 2017. *The Oxford handbook of pragmatics*. Oxford University Press.
- Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function. *Annotation, Technical Report, 97-02, University of Colorado, CO, USA*.
- Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520.
- Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551.
- Thi-Lan Ngo, Pham Khac Linh, and Hideaki Takeda. 2018. A vietnamese dialog act corpus based on ISO 24617-2 standard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.
- Silvia Quarteroni, Alexei V Ivanov, and Giuseppe Riccardi. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5596–5599. IEEE.
- V.N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.
- Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. 2018. Japanese dialogue corpus of information navigation and attentive listening annotated with extended ISO 24617-2 dialogue act tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

- Lin Zhao and Zhe Feng. 2018. Improving slot filling in spoken language understanding with joint pointer and attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 426–431.
- Tianyu Zhao and Tatsuya Kawahara. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127.

Tracing Metonymic Relations in T-PAS: An Annotation Exercise on a Corpus-based Resource for Italian

Emma Romani
University of Pavia
Department of Humanities
emma.romani01@
universitadipavia.it

Elisabetta Ježek
University of Pavia
Department of Humanities
jezek@unipv.it

Abstract

In this paper we address the main issues and results of a research thesis (Romani, 2020) dedicated to the annotation of metonymies in T-PAS, a corpus-based digital repository of Italian verbal patterns (Ježek et al., 2014). The annotation was performed on the corpus instances of a selected list of 30 verbs and was aimed at both implementing the resource with metonymic patterns and identifying and creating a map of the metonymic relations that occur in the verbal patterns. The annotated corpus data (consisting of 1218 corpus instances), the patterns, and the relations can be useful for NLP tasks such as metonymy recognition.¹

1 Introduction

Metonymy is a language phenomenon for which one referent is used to denote another referent associated with it (Lakoff & Johnson, 1980; Fauconnier, 1985; Ježek, 2016). For example, in the sentence ‘he drank a glass at the pub’, *glass* (the *metonymic* or *source type* denoting a container) refers to its content (the *target type*, a beverage). In our research, we investigated metonymy from a corpus-based perspective, through the analysis of corpus data and the annotation performed in T-PAS, a corpus-based resource for Italian verbs. T-PAS consists of a repository of typed predicate argument structures (Ježek et al., 2014), i.e. verbal patterns together with semantically-specified arguments, linked to manually annotated corpus instances (see Section 3.1). An example of a pattern (or *t-pas*) for the verb

bere ‘to drink’ is reported in Figure 1:

1 [Animate] bere [Beverage]
[Animate] ingerisce, assume [Beverage]

Figure 1. Pattern 1 of the verb *bere* (‘to drink’) in T-PAS together with its sense description

where [Animate] and [Beverage] are the semantic types specifying the subject and object positions.

The annotation of metonymies was performed on the corpus instances of a list of 30 verbs contained in T-PAS (taken from Ježek & Quochi, 2010). As emerged from this background study, the semantic properties of those verbs were likely to convey metonymies in their argument structures. Starting from this list, our work was intended as an implementation of the resource; specifically, we annotated metonymic corpus instances and created metonymic sub-patterns linked to them.

The research had several aims. First, we were interested in studying qualitatively the phenomenon in and through the corpus instances and in implementing the annotation tool of the resource with a specific feature that allowed us to encode metonymic arguments in the verbal patterns. For the latter purpose, we collaborated with the Faculty of Informatics at Masaryk University of Brno (CZ): they gave us the technical support for the implementation of the annotation tool.

Second, our intention was to conceive a general theoretical framework to represent the metonymies found through the qualitative corpus analysis, by designing a map of metonymies and by drafting a list of the metonymic *relations* that occur in the verbal patterns (see Section 4).

The paper is organized as follows. In Section 2 we present related studies. In Section 3 we describe the methodology we followed in annotating the corpus instances for metonymies, together with a brief introduction to T-PAS. In Section 4 we present the results of our annotation: a generalization of the

¹ Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

metonymic relations found, and a map which visually highlights the semantic and cognitive connections between the semantic types. Further developments of the project are described in Section 5; our intention is to enrich the number of analysed verbs and eventually add new types of metonymic relations.

2 Related works

Corpus-based studies on metonymy are often intended for NLP tasks. Markert & Nissim (2006), provide a corpus-based annotation scheme for metonymies with the aim of improving automatic metonymy recognition and resolution. Related to it, Markert and Nissim (2007) present the results of a supervised task on metonymy resolution; an analogous task has been addressed by Pustejovsky et al. (2010) within the scope of SemEval-2010. A recent study elaborated a computational model based on the dataset of Pustejovsky et al. (2010) for the detection of metonymies (McGregor et al., 2017).

Corpus-based studies on metonymies do not necessarily address NLP tasks. An attempt to implement corpus-based resources to display metonymies is described in Ježek & Frontini (2010). Also, Pustejovsky & Ježek (2008) present a corpus investigation aimed at identifying metonymic mechanisms in predicate-argument constructions from a theoretical perspective. Finally, Marini & Ježek (2020) performed an equivalent corpus-based metonymy annotation on a sample of 101 Croatian verbs within the scope of CROATPAS (Marini & Ježek, 2019), sister project of T-PAS.

3 Resource and methodology

3.1 The resource: T-PAS

T-PAS is the corpus-based resource used in this research. It consists of a repository of Typed Predicate Argument Structures (T-PAS) (Ježek et al., 2014) for Italian verbs. The resource consists of three components:

- 1) a repository of corpus-derived predicate argument structures for verbs with semantic specification of the arguments, linked to lexical units (verbs);
- 2) an inventory of about 200 corpus-derived semantic classes for nouns, relevant for the disambiguation of the verb in context;
- 3) a corpus² of sentences that instantiate T-

PAS, tagged with lexical unit (verb) and pattern number.

Typed predicate argument structures (or *t-pass*) are patterns which display the syntactic and semantic properties of verbs: for each meaning of a verb a specific t-pas is provided. Verb sense is determined by the arguments it combines with (subject, object, etc.), which are defined through a specific Semantic Type.³

T-pass are corpus-derived: patterns were acquired through the manual clustering and annotation of corpus instances for each verb following the CPA procedure (Hanks, 2013). Each t-pas in the resource is labelled with a number and connected to a list of corpus instances, realizing the specific verb meaning. Each pattern is associated with a *sense description*, a brief definition of the meaning of the verb (see the second line in Figure 1). Each pattern can have *sub-patterns* created by annotators, for corpus instances that do not reflect the prototypical semantic behaviour of the verb or of its arguments, as in metonymic uses. Like their patterns, sub-patterns are connected to annotated instances from the corpus. In our work, we implemented the annotation tool by adding a new label ('.m'), which we used to annotate metonymic uses in sub-patterns (see Figure 2).

1	[Animate] bere [Beverage] [Animate] ingerisce, assume [Beverage]
1.m	[Animate] bere [Container (bicchiere bottiglia)] [Animate] ingerisce, assume [Container] (che contiene [Beverage])

Figure 2. Metonymic sub-pattern for t-pas 1 of the verb *bere* ('to drink') in T-PAS

3.2 Methodology

We conceived an empirical methodology in order to get significant results from the corpus analysis: we manually extracted significant instances from the corpus and annotated them as metonymic instances under their specific pattern. In order to annotate the instances, we exploited the Sketch Engine functions available for analysing the corpus. The annotation scheme can be summarized as follows:

- 1) Random sampling of about 200 corpus instances for each of the 30 verbs (the sample allowed to reduce the time spent in skimming the instances, still providing a balanced overview of the kind of instances contained in the corpus);

² The corpus is a reduced and cleaned version of It-WaC (Baroni et al., 2009), a corpus of Italian texts, available in the Sketch Engine tool (Kilgariff et al., 2014).

³ Semantic Types are expressed through square brackets (e.g. [Animate], [Beverage]) and are organized in a hierarchy, called the System of Semantic Types (see Ježek, 2019 for a more detailed account).

<i>verbs</i>	<i>n. of annotated instances</i>	<i>verbs</i>	<i>n. of annotated instances</i>	<i>verbs</i>	<i>n. of annotated instances</i>
accusare	34	concludere	39	parcheggiare	19
annunciare	40	contattare	27	raggiungere	51
arrivare	47	continuare	21	recarsi	81
ascoltare	103	divorare	0	rimbombare	28
atterrare	76	echeggiare	24	sentire	27
avvisare	22	finire	66	sorseggiare	32
bere	93	informare	16	udire	35
chiamare	16	interrompere	39	venire	63
cominciare	19	leggere	84	versare	12
completare	42	organizzare	13	visitare	49

Table 1. List of the Italian verbs with number of annotated instances in ItWaC corpus

- 2) Manual annotation of the metonymic instances through the sublabel (signalled with “.m”);
- 3) Implementation of the sub-pattern in the resource by adding metonymic semantic types (see 1.m in Figure 1);
- 4) Definition of the metonymic relation (see Table 2) between the source and the target semantic type (e.g. [Container] ‘contains’ [Beverage]), with its encoding in the sense description, translated in Italian (see Figure 2).

In Table 1, we show the number of instances annotated for each of the 30 verbs. Overall, the dataset consists of 1218 annotated instances. The number of instances from the random sample can vary from a verb to another, because verbs have different frequencies in the corpus and metonymic phenomena can be more or less pervasive according to the verb under examination. Some cases (e.g. *divorare* – ‘to devour’ – in Table 1) did not provide any metonymic instance at all (for an explanation and further discussion on this point, see Romani, 2020).⁴

The annotation procedure was conducted manually by one single annotator (the first author) and, so far, it was not possible to evaluate our annotation procedure as we focused on the qualitative analysis and the retrieval of the relations: it is our intent for the future, as it is essential for further progresses in the research.

Currently, the adopted annotation scheme did not provide ambiguous cases, as metonymies were usually clear-cut and the shift of referent from the source to the target semantic type easily identifica-

ble. This may differ from metaphors, for example, where the shift between literal and non-literal meaning may be perceived as more gradual. However, further investigation needs to be done through the annotation of a higher number of instances (expanding the list of verbs) and the comparison and the evaluation of the annotation results of more than one annotator.

4 Results

The overall aim of the research was to give a theoretical account of the metonymic relations found through the corpus analysis and pattern annotation. Therefore, the main results of the study are a list of metonymic relations between the *target* and the *metonymic* (source) *semantic type* (Table 2, Appendix) and a map where the target semantic types are connected to the metonymic types, and the relation between the two is expressed (Figure 3).

The second column in Table 2 (Appendix) reports the 37 *relations* we identified and encoded (the relations are grouped according to their target type, following this order: [Human], [Location], [Document], [Beverage], [Vehicle], [Sound]). The *relation* is a short description that illustrates how the metonymic semantic type is connected to the target semantic type; for example (see the highlighted line of the table), [Container] is the metonymic semantic type (first column) and ‘contains’ is the relation (second column) which links [Container] to the target semantic type [Beverage] (third column).⁵ An instance for this is: ‘we went out to drink *a glass*’ (*glass* ‘contains’ *something to drink*). The fourth column contains an instance in Italian

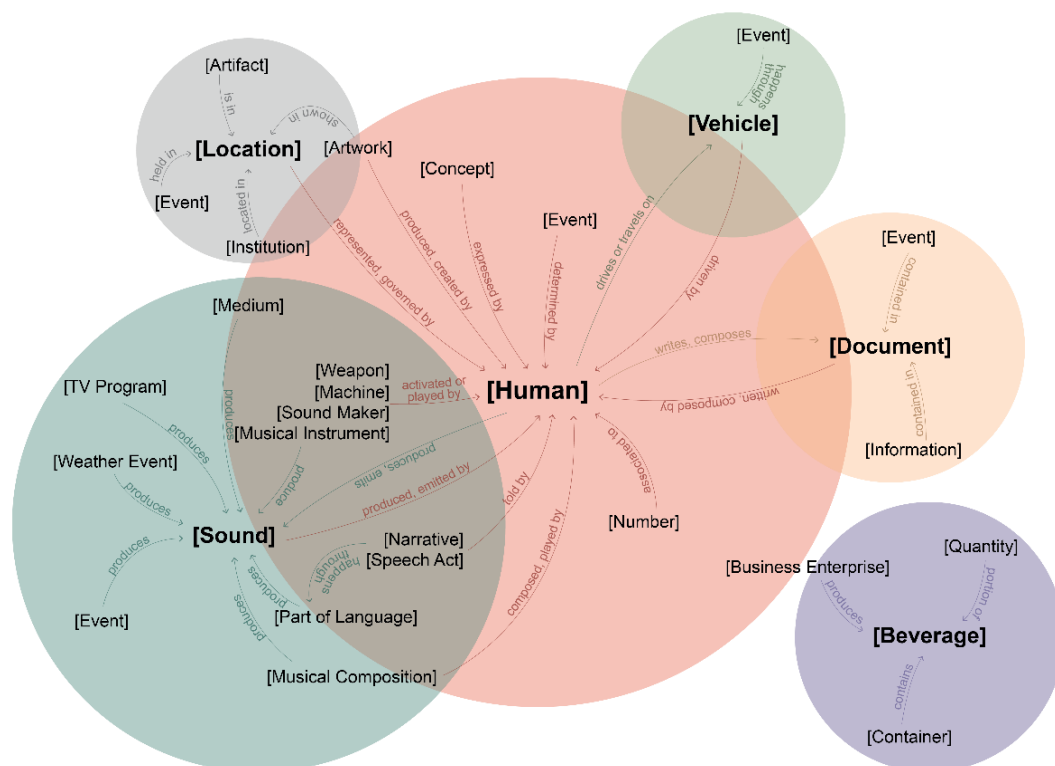
⁴ In some cases, additional instances were included, if the number of metonymic instances provided by the sample was not sufficient to exemplify a specific relation. Instances with arguments and semantic types analogous to the ones already tagged were selected. To do so, we exploited other Sketch Engine functions (see Romani, 2020 for further details).

⁵ Highlighted in bold are the metonymic semantic types that are also target types (for example, [Sound] is the metonymic semantic type in “[Sound] is emitted by [Human]”, but also the target semantic type in “[Medium] produces [Sound]”).

from ItWaC reduced corpus, for each relation found. For each instance, the metonymic argument (exemplifying the source-metonymic semantic type) is highlighted in bold, and the verb triggering the metonymy is in *italics*.

Our results show that metonymic semantic types are fluid; target types can also be metonymic types,

The structure of the map we conceived draws attention to two main aspects. First, it depicts the complexity of the metonymic relations between semantic types and highlights how metonymy is not a unidirectional phenomenon but, conversely, it is fluid and changeable. Second, from a cognitive point of view, [Human] is at the centre of most of the relations and each target type area is connected to it by multiple relations. In particular, in our data, [Human] is deeply connected and involved within the [Sound] area (for more details, see Romani, 2020).



5 Conclusions and future works

In this paper, we approached the study of metonymy from a corpus-based perspective. The research was conducted on a selected list of verbs, taken from a background study (Ježek & Quochi, 2010). Our aim was to search for metonymic phenomena inside a corpus of Italian language and to register them in a resource for Italian verbs, T-PAS. To do so, we conceived an annotation scheme and procedure that gave us relevant results and allowed us to register a variety of metonymic relations.

We also attempted to make some theoretical generalizations based on the metonymic relations we found through the corpus analysis. We therefore created a list of metonymic relations and we designed a map in which the relations are connected to the semantic types they involve. Both the map and the list depict the complexity and variety of the phenomenon, in terms of number of possible metonymic relations and of the semantic types interested.

In future perspectives, we intend to enrich the map and the list with new relations by extending the number of verbs investigated and to evaluate the annotation procedure. For future annotations, we will provide the current version of the list and of the map on the online public version of T-PAS (upcoming). We are also interested in comparing our results with those in Marini & Ježek (2020), in a cross-linguistic perspective.

In line with previous studies (Section 2), we believe that the annotated corpus data, as well as the relations in Table 2, could be useful for automatic detection of metonymies. To our knowledge, little work has been done on this for Italian language: it would be therefore intriguing to test our data in NLP tasks.

References

- Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Fauconnier, G. (1985). *Mental Spaces*. Cambridge, USA: MIT Press.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.
- Ježek, E. (2016). *The Lexicon: An Introduction*. Oxford: Oxford University Press.
- Ježek, E. (2019). Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies. In Barton, A., Seppälä, S. & Porello, D. (eds.), *Proceedings of the Joint Ontology Workshops 2019*. Graz, Austria.
- Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. In *Proceedings of LREC*, 890-895.
- Ježek, E. & Frontini, F. (2010). From Pattern Dictionary to PatternBank. In de Schryver G.M. (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Menha Publishers, Kampala. 215-239.
- Ježek, E. & Quochi, V. (2010). Capturing Coercions in Texts: A First Annotation Exercise. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. La Valletta, Malta. (LREC-2010), 1464-1471.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we Live By*. Chicago, USA: University of Chicago Press.
- Marini, C. & Ježek, E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it)*. Bari, Italy.
- Marini C. & Ježek, E. (2020). Annotating Croatian Semantic Type Coercions in CROATPAS. In *Proceedings of the 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16)*, pages 49–58. *Language Resources and Evaluation Conference (LREC 2020)*, Marseille, May 2020.
- Markert, K. & Nissim, M. (2006). Metonymic Proper Names: A Corpus-Based Account. In Gries, S.T. & Stefanowitsch, A. (eds.), *Corpora in Cognitive Linguistics. Vol. 1: Metaphor and Metonymy*. Mouton-De Gruyter: Berlin & New York.
- Markert, K. & Nissim, M. (2007, June). Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 36-41).
- McGregor, S., Ježek, E., Purver M. & Wiggings G. (2017). A Geometric Model for Detecting Semantic Coercion. In Gardent, C. and C. Retoré (eds) *Proceedings of the 12th International Workshop on Computational Semantics (IWCS)*, ACL Anthology, Association for Computational Linguistics, pp. 1-14.
- Pustejovsky, J. & Ježek, E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In *Italian Journal of Linguistics/Rivista Italiana di Linguistica* 20(1).
- Pustejovsky, J., Rumshisky, A., Plotnick, A., Ježek, E., Batiukova, O. & Quochi, V. (2010). SemEval-2010

Task 7: Argument Selection and Coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation (ACL-2010)*. Uppsala, Sweden. 27-32.

Romani, E. (2020). *Searching for Metonymies in Natural Language Texts. A Corpus-based Study on a Resource for Italian Verbs*. BA Thesis, University of Pavia, Pavia, Italy.

Appendix

<i>metonymic (source) semantic type</i>	<i>relation</i>	<i>target semantic type</i>	<i>corpus example (ItWaC reduced)</i>
[Vehicle]	is driven by	[Human]	L'agente scese velocemente in strada, <i>chiamò un taxi</i> e dette l'indirizzo segreto.
[Document]	is written, composed by	[Human]	Il gioco, <i>informa un comunicato</i> , sarà lanciato contemporaneamente in Italia e Regno Unito.
[Location]	is represented, governed by	[Human]	Dissidenti e rifugiati <i>accusano la Corea del Nord</i> di tortura politica e [...] chiedono di includere nei colloqui anche il tema dei diritti umani e delle libertà fondamentali.
[Sound]	is produced, emitted by	[Human]	A questo punto <i>una voce interrompe</i> Gesù.
[Event]	is determined by	[Human]	Ricordo la <i>telefonata</i> che mi <i>raggiunse</i> la mattina presto nella mia abitazione di Milano, la corsa in ufficio, il viaggio dell'indomani nei luoghi della catastrofe [...]
[Projectile]	is shot by	[Human]	L' uomo viene <i>raggiunto</i> da cinque <i>proiettili</i> e muore mentre viene trasportato in ospedale.
[Sound Maker]	is activated by	[Human]	Una <i>campana</i> <i>annuncia</i> l'inizio della messa.
[Machine]	is activated by	[Human]	<i>L'altoparlante</i> <i>annunciava</i> l'arrivo di un treno.
[Musical Instrument]	is played by	[Human]	Oltre al Flauto d'oro, <i>lo zufolo</i> pastorale <i>annuncia</i> ed accompagna Papageno.
[Concept]	is expressed by	[Human]	Alcuni studiosi <i>accusavano</i> la <i>psicologia</i> di naturalismo, altri di non essere una scienza naturale.
[Number]	is associated to	[Human]	L'iniziativa consiste nella possibilità per gli anziani di <i>contattare un numero</i> messo a disposizione gratuitamente dal Comune, [...] che attiverà uno degli oltre mille volontari.
[Part of Language]	is pronounced by	[Human]	La <i>frase</i> <i>venne interrotta</i> dal suono di sirene, quelle della Polizia.
[Event]	is held in	[Location]	Una sera, mentre <i>si sta recando ad una cena</i> dove dovrà tenere un discorso, Henry riceve l'invito a presentarsi al commissariato.
[Institution]	is located in	[Location]	Giovanni Paolo II <i>ha visitato</i> il <i>Parlamento</i> italiano, su invito dei Presidenti della Camera dei Deputati.
[Artwork]	is shown in	[Location]	[...] raggiungiamo piazza Pio IX dove sorge la Pinacoteca Ambrosiana, entriamo per <i>visitare le opere</i> di Caravaggio, Leonardo e Botticelli.
[Artifact]	is in	[Location]	La mia peste la sento tre volte al giorno, anche se non vuole <i>venire al telefono</i> a parlarmi [...]
[Human]	writes, composes	[Document]	<i>Ho letto Dante, Moravia, Calvino.</i>
[Information]	is contained in	[Document]	Vi raccomandiamo, prima di procedere nella consultazione, di <i>leggere le avvertenze</i> .
[Event]	is contained in	[Document]	Consiglio di <i>leggere</i> senza paraocchi ideologici questa <i>intervista</i> del prof. Dallapiccola sulla diagnosi preimpianto.
[Container]	contains	[Beverage]	Al pub Orange Paolo <i>aveva bevuto un bicchiere</i> di troppo e alcuni clienti [...] hanno chiesto l'intervento dei carabinieri

			perché venisse allontanato.
[Quantity]	is a portion of	[Beverage]	Occorre portarsi le sedie e il fuoco, e mettere ciascuno due soldi, se si vuole <i>bere</i> un goccio .
[Business Enterprise]	produces	[Beverage]	Anche noi della Nazionale <i>beviamo</i> Uliveto!
[Human]	drives or travels on	[Vehicle]	Il presidente del Consiglio è <i>atterrato</i> a mezzogiorno sul campo sportivo di Sant'Agnello a Sorrento.
[Fantasy Character]	drives or travels on	[Vehicle]	Una coppia di alieni <i>atterra</i> sulla Terra, precisamente in una campagna.
[Event]	happens through	[Vehicle]	Alle 16:50 è <i>atterrato</i> il volo speciale Parigi-Beirut della linea di bandiera libanese.
[Machine]	activated by [Human], produces	[Sound]	Ma molti non hanno voluto <i>ascoltare</i> la sirena d' allarme e sono rimasti nelle loro abitazioni [...]
[Weapon]	activated by [Human], produces	[Sound]	I fulci <i>echeggiano</i> in lontananza mentre tutto intorno continua a muoversi e girare.
[Sound Maker]	activated by [Human], produces	[Sound]	Le campane non risuoneranno i rintocchi della morte, ma <i>echeggeranno</i> a festa per celebrare la Vita.
[Musical Instrument]	played by [Human], produces	[Sound]	Le trombe non si <i>udivano</i> più, ma dalla parte della vallata si udivano ad intervalli dei lontani fragori.
[Human]	produces, emits	[Sound]	Se io fossi una persona che non ha mai <i>ascoltato</i> Patty Smith [...] magari mi passerebbe anche la voglia di andarla a scoprire.
[Weather Event]	produces	[Sound]	<i>Ascolta</i> la pioggia , se hai sonno ti tengo con me.
[Part of Language]	pronounced by [Human], produces	[Sound]	Avete <i>ascoltato</i> tutti le parole di Romano: sono sicuro che tanti tra noi pensano che le sue idee siano una buona base per governare il Paese.
[Narrative]	told by [Human] through [Part of Language], produces	[Sound]	Per <i>ascoltare</i> un racconto , una storia , occorre restare in silenzio.
[Speech Act]	told by [Human] through [Part of Language], produces	[Sound]	Low Key <i>udi</i> a stento la domanda di Eric mentre tornava a concentrarsi sul presente.
[Event]	involves	[Sound]	In una grotta dedicata alla Madonna di Lourdes è possibile, oltre che <i>ascoltare</i> la Santa Messa la domenica, celebrare matrimoni [...]
[Medium]	produces	[Sound]	Roberto Landi sta seduto dentro il camper e <i>ascolta</i> la televisione .
[TV Program]	emits	[Sound]	L'autista stava <i>ascoltando</i> un notiziario della Bbc su quanto è accaduto qualche giorno fa a Madaen.

Table 2. Metonymic relations (column 2) identified between the source (metonymic) semantic type (column 1) and the target semantic type (column 3), with an instance from ItWaC for each relation found (column 4)

Datasets and Models for Authorship Attribution on Italian Personal Writings

Gaetana Ruggiero[•], Albert Gatt[•], Malvina Nissim[◊]

[•]Institute of Linguistics and Language Technology, University of Malta, Malta

[◊]Center for Language and Cognition, University of Groningen, The Netherlands

garuggiero@gmail.com, albert.gatt@um.edu.mt, m.nissim@rug.nl

Abstract

Existing research on Authorship Attribution (AA) focuses on texts for which a lot of data is available (e.g. novels), mainly in English. We approach AA via Authorship Verification on short Italian texts in two novel datasets, and analyze the interaction between genre, topic, gender and length. Results show that AV is feasible even with little data, but more evidence helps. Gender and topic can be indicative clues, and if not controlled for, they might overtake more specific aspects of personal style.

1 Introduction and Background

Authorship Attribution (AA) is the task of identifying authors by their writing style. In addition to being a tool for studying individual language choices, AA is useful for many real-life applications, such as plagiarism detection (Stamatatos and Koppel, 2011), multiple accounts detection (Tsikerdekis and Zeadally, 2014), and online security (Yang and Chow, 2014).

Most work on AA focuses on English, on relatively long texts such as novels and articles (Juola, 2015) where personal style could be mitigated due to editorial interventions. Furthermore, in many real-world applications the texts of disputed authorship tend to be short (Omar et al., 2019).

The PAN 2020 shared task was originally meant to investigate multilingual AV in fanfiction, focusing on Italian, Spanish, Dutch and English (Bevendorff et al., 2020). However, the datasets were eventually restricted to English only, to maximize the amount of available training data (Kestemont et al., 2020), emphasizing the difficulty in compiling large enough datasets for less-resourced languages.

AA research in Italian has largely focused on the single case of Elena Ferrante (Tuzzi and Cortelazzo, 2018)¹. The present work seeks a more realistic take, using more diverse, user-generated data namely web forums comments and diary fragments, thereby introducing two novel datasets for this task: *ForumFree* and *Diaries*.

We cast the AA problem as *authorship verification* (AV). Rather than identifying the specific author of a text (the most common task in AA), AV aims at determining whether two texts were written by the same author or not (Koppel and Schler, 2004; Koppel et al., 2009).

The GLAD system of Hürlimann et al. (2015) was specifically developed to solve AV problems, and has been shown to be highly adaptable to new datasets (Halvani et al., 2018). GLAD uses an SVM with a variety of features including character level ones, which have proved to be most effective for AA tasks (Stamatatos, 2009; Moreau et al., 2015; Hürlimann et al., 2015), and is freely available. Moreover, Kestemont et al. (2019) show that many of the best models for authorship attribution are based on Support Vector Machines. Hence we adopt GLAD in the present study.

More specifically, we run GLAD on our datasets and study the interaction of four different dimensions: topic, gender, amount of evidence per author, and genre. In practice, we design intra-topic, cross-topic, and cross-genre experiments, controlling for gender and amount of evidence per author. The focus on cross-topic and cross-genre AV is in line with the PAN 2015 shared task (Stamatatos et al., 2015); this setting has been shown to be more challenging than the task definitions of previous editions (Juola and Stamatatos, 2013; Stamatatos et al., 2014).

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.newyorker.com/culture/cultural-comment/the-unmasking-of-elena-ferrante>

Contributions We advance AA for Italian introducing two novel datasets, *ForumFree* and *Diaries*, which contribute to enhance the amount of available Italian data suitable for AA tasks.²

Running a battery of experiments on personal writings, we show that AV is feasible even with little data, but more evidence helps. Gender and topic can be indicative clues, and if not controlled for, they might overtake more specific aspects of personal style.

2 Data

For the present study, we introduce two novel datasets, *ForumFree* and *Diaries*. Although already compiled (Maslennikova et al., 2019), the original ForumFree dataset was not meant for AA. Therefore, we reformat it following the PAN format³. The dataset contains web forum comments taken from the ForumFree platform⁴, and the subset used in this work covers two topics, *Medicina Estetica* (“Aesthetic Medicine”) and *Programmi Tv* (“Tv Programmes”; *Celebrities* in the original dataset). A third subset, *Mix*, is the union of the first two. The Diaries dataset is originally assembled for the present study, and contains a collection of diary fragments included in the project *Italiani all'estero: i diari raccontano* (“Italians abroad: the diaries narrate”).⁵ For Diaries, no topic classification has been taken into account. Table 1 shows an overview of the datasets.

Subset	# Authors			# Docs	W/A	D/A	W/D
	F	M	Tot				
Med Est	33	44	77	56198	63	661	48
Prog TV	78	71	149	153019	32	812	22
Mix	111	115	276	209217	41	791	29
Diaries	77	188	275	1422	462	5	477

Table 1: Overview of the datasets. W/A = Avg words per author; D/A = Avg docs per author; W/D = Avg words per doc.

2.1 Preprocessing

For the ForumFree dataset, comments which only contained the word *up*, commonly used on the internet to give new visibility to a post that was writ-

ten in the past, were removed from the dataset, together with their authors when this was the only text associated with them.

The stories narrated in the diaries are of a very personal nature, which means that many proper nouns and names of locations are used. To avoid relying on these explicit clues, which are strong but not indicative of personal writing style, we perform Named Entity Recognition (NER), using *spaCy* (Honnibal, 2015). Person names, locations and organizations were replaced by their corresponding labels, namely *PER*, *LOC*, *ORG*. The fourth label used by *spaCy*, *MISC* (miscellany), was not considered; dates were also not normalized. Moreover, a separate set of experiments was performed by *bleaching* the diary texts prior to their input to the GLAD system. The bleaching method was proposed by van der Goot et al. (2018) in the context of cross-lingual Gender Prediction, and consists of transforming tokens into an abstract representation that masks lexical forms while maintaining key features. We only use 4 of the 6 original features. *Shape* transforms uppercase letters into ‘U’, lowercase ones into ‘L’, digits into ‘D’, and the rest into ‘X’. *PunctA* replaces emojis with ‘J’, emoticons with ‘E’, punctuation with ‘P’ and one or more alphanumeric characters with a single ‘W’. *Length* represents a word by the number of its characters. *Frequency* corresponds to the *log* frequency of a token in the dataset. The features are then concatenated. The word ‘House’ would be rewritten as ‘ULLLL W 05 6’.

2.2 Reformatting

We reformat both datasets in order to make them suitable for AV. The data is divided into so-called *problems*: each problem is made of a known and an unknown text of equal length.

To account for the shortness of the texts and to avoid topic biases that would derive by taking consecutive text as known and unknown fragments, all the documents written by the same author are first shuffled and then concatenated into a single string. The string is split into two spans containing the same number of words, so that the words contained in the unknown span come from subsets of texts which are different from the ones that form the known one. An example of this process is displayed in Figure 1. Rather than being represented by individual productions, each author is therefore represented by a *set* of texts, whose original se-

²Further information about the datasets can be found at <https://github.com/garuggiero/Italian-Datasets-for-AV>

³<https://pan.webis.de/clef15/pan15-web/authorship-verification.html>

⁴<https://www.forumfree.it/>

⁵<https://www.idiariiraccontano.org>

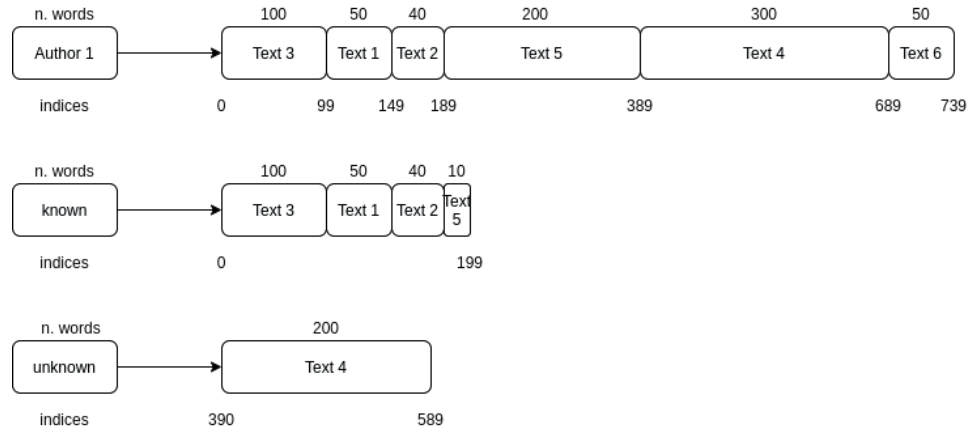


Figure 1: Example of the creation of known and unknown documents for the same author when considering 400 words per author.

quential order has been altered. Each known text is paired with an unknown text from the same author. To create negative instances, given a dataset with multiple problems, one can (i) make use of external documents (*extrinsic* approach (Seidman, 2013; Koppel and Winter, 2014)), or (ii) use fragments collated from all authors in the training data, except the target author (*intrinsic* approach). We create negative instances with an intrinsic approach. More specifically, following Dwyer (2017), the second half of the unknown array is shifted by one, so that the texts of the second half of the known array are paired with a *different-author* text in the unknown array. In this way, the label distribution is balanced.

3 Method

Given a pair of known and unknown fragments (KU pair), the task is to predict whether they are written by the same author or not. In designing our experiments, we control for topic, gender, amount of evidence, and genre. The latter is fostered by the diverse nature of our datasets.

Topic Maintaining the topic roughly constant should allow stylistic features to gain more discriminative value. We design intra-topic (IT) and cross-topic experiments (CT). In IT, we distinguish same- and different-topic KU pairs. In same-topic, we train and test the system on KU pairs from the same topic. In different-topic, we include the Mix set and the diaries. Since we train and test on a mixture of topics and there can be topic overlap, these are not truly cross-topic, and we do not consider them as such.

Given that no topic classification is available for the diaries, the CT experiments are only performed on the ForumFree dataset. We train the system on Medicina Estetica and test it on Programmi Tv, and vice versa.

Gender Previous work has shown that similarity can be observed in writings of people of the same gender (Basile et al., 2017; Rangel et al., 2017).⁶ In order to assess the influence of same vs different gender in AA, we consider three gender settings: only female authors and only male authors (*single-gender*), and *mixed-gender*, where the known and unknown document can be either written by two authors of the same gender, or by a male and a female author. In dividing the subsets according to the gender of the authors, we consider gender implicitly. However, we also perform experiments adding gender as feature to the instance vectors, indicating both the gender of the known and unknown documents’ authors and whether or not the gender of the authors is the same.

Evidence Following Feiguina and Hirst (2007), we experiment with KU pairs of different sizes, i.e. with 400, 1 000, 2 000 and 3 000 words per author. Each element of the KU pair is thus made up of 200, 500, 1 000 and 1 500 words respectively. To observe the effect of the different text sizes on the classification, we manipulate the number of instances in training and test, so that the same authors are included in all the different word settings of a single topic-gender experiment.

⁶Binary gender is a simplification of a much more nuanced situation in reality. Following previous work, we adopt it for convenience.

Genre We perform cross-genre experiments (CG) by training on ForumFree and testing on the Diaries, and vice versa.

Splits and Evaluation We train on 70% and test on 30% of the instances. However, since we are controlling for gender and topic, the number of instances contained in the training and test sets varies in each experiment. We keep the test sets stable across IT, CT and CG experiments, so that we can compare results. Following the PAN evaluation settings (Stamatatos et al., 2015), we use three metrics. $c@1$ takes into account the number of problems left unanswered and rewards the system when it classifies a problem as unanswered rather than misclassifying it.

Probability scores are converted to binary answers: every score greater than 0.5 becomes a positive answer, every score smaller than 0.5 corresponds to a negative answer and every score which is exactly 0.5 is considered as an unanswered problem. The *AUC* measure corresponds to the area under the ROC curve (Fawcett, 2006), and tests the ability of the system to rank scores properly, assigning low values to negative problems and high values to positive ones (Stamatatos et al., 2015). The third measure is the product of $c@1$ and *AUC*.

Model We run all experiments using GLAD (Hürlimann et al., 2015). This is an SVM with *rbf* kernel, implemented using Python’s *scikit-learn* (Pedregosa et al., 2011) library and NLTK (Bird et al., 2009). GLAD was designed to work with 24 different features, which take into account stylometry, entropy and data compression measures. We compare GLAD to a simple baseline which randomly assigns a label from the set of possible labels (i.e. ‘YES’ or ‘NO’) to each test instance.

Our choice fell on GLAD for a variety of reasons. As a general observation, even in later challenges, SVMs have proven to be the most effective for AA tasks (Kestemont et al., 2019). More specifically, in a survey of freely available AA systems, GLAD showed best performance and especially high adaptability to new datasets (Halvani et al., 2018). Lastly, de Vries (2020) has explored fine-tuning a pre-trained model for AV in Dutch, a less-resourced language compared to English. He found that fine-tuning BERTje (a Dutch monolingual BERT-model, (de Vries et al., 2019)) with PAN 2015 AV data (Stamatatos et al., 2015),

failed to outperform a majority baseline (de Vries, 2020). He concluded that Transformer-encoder models might not be suitable for AA tasks, since they will likely overfit if the documents contain no reliable clues of authorship (de Vries, 2020).

4 Results and Discussion

The number of experiments is high due to the interaction of the dimensions we consider.

Tables 2 and 3 only include the mixed-gender results of the IT experiments on Mix (which corresponds to the entire ForumFree dataset used for this study) and Diaries, respectively. Results concerning all dimensions considered are anyway discussed in the text. We refer to the combined score. Since the baseline results are different for each setting, we do not include them. However, all models perform consistently above their corresponding baseline.

For the Mix topic, we achieved 0.966 with 96 authors in total and 3 000 words (Table 2). For the diaries, we achieved 0.821 with 46 authors in total and 3 000 words each (Table 3).⁷ Although the training and test sets are of different sizes for both datasets, more evidence seems to help the model to solve the problem.

In the IT experiments, the highest score for Medicina Estetica is 0.923, with 41 authors in total and 1 000 words per author, and for Programmi Tv 0.944, with 59 authors and 3 000 words each. In the CT setting, the scores stay basically the same in both directions. In CG, when training on the diaries and testing on Mix, we obtain the same score when training on Mix with 3 000 words. When training on Mix and testing on Diaries, we achieved 0.737 on the same test set, and 0.748 with 1 000 words per instance.

Discussion When more variables interact in the same subset, as in mixed-gender sets of the ForumFree and Diaries dataset, we found that the classifier uses the implicit gender information. Indeed, it achieves slightly better scores in mixed-gender settings than in female- and male-only ones, suggesting that the classifier might be using internal clustering of the data rather than writing style characteristics. This also explains why results are higher in Mix than in separate topics, because the classifier can use topic information.

⁷Using a bleached representation of the texts, the score increased by 0.36

# W/A	# Auth	# Problems		Eval					
		Train	Test	C	I	U	c@1	AUC	*
400	127	88	39	33	6	0	0.846	0.947	0.801
1 000	109	76	33	30	3	0	0.909	0.926	0.842
2 000	100	70	30	29	1	0	0.967	0.995	0.962
3 000	96	67	29	28	1	0	0.966	1.000	0.966

Table 2: Training and test set configurations and IT evaluation scores on Mix texts written by female and male authors. *C, I* and *U* are Correct, Incorrect, Unanswered problems.

# W/A	# Auth	# Problems		Eval					
		Train	Test	C	I	U	c@1	AUC	*
400	229	160	69	47	21	1	0.691	0.725	0.500
1 000	180	126	54	43	11	0	0.796	0.891	0.709
2 000	98	68	30	25	5	0	0.833	0.905	0.754
3 000	46	32	14	12	2	0	0.857	0.958	0.821

Table 3: Training and test configurations and IT evaluation scores on diaries made of NE converted text written by both genders. *C, I* and *U* are Correct, Incorrect, Unanswered problems.

We also observe that by adding gender as an explicit feature in topic- and gender-controlled subsets, GLAD uses this information to improve classification, especially in mixed-gender scenarios.

Although previous research demonstrated that CT and CG experiments are harder than IT ones (Sapkota et al., 2014; Stamatatos et al., 2015), in our case the scores for the three settings are comparable. However, since we only performed CT and CG experiments on mixed-gender subsets, the gender-specific information might have also played a role in this process (see above).

Overall, the experiments show that using a higher number of words per author is preferable. Although 3 000 words seems to be optimal for most settings, in the large number of experiments that we carried out (not all included in this paper) we also observed that lower amounts of words also led to comparable results. This aspect will require further investigation.

5 Conclusion

We experimented with AV on Italian forum comments and diary fragments. We compiled two datasets and performed experiments which considered the interaction among topic, gender, length and genre. Even when the texts are short and present more individual variation than traditional texts used in AA, AV is a feasible task, but having more evidence per author improves classification.

While making the task more challenging, controlling for gender and topic ensures that the system prioritizes authorship over different data clusters. Although the datasets used are intended for AV problems, they can be easily adapted to other AA tasks. We believe this to be one of the major contributions of our work, as it can help to advance the up-to-now limited AA research in Italian.

Acknowledgments

The ForumFree dataset was a courtesy of the Italian Institute of Computational Linguistics “Antonio Zampolli” (ILC) of Pisa.⁸

References

- Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *CEUR Workshop Proceedings*, volume 1866.
- Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Shared Tasks on Authorship Analysis at PAN 2020. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*,

⁸<http://www.ilc.cnr.it/>

- pages 508–516, Cham. Springer International Publishing.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.
- Wietse de Vries. 2020. Language Models are not just English Anymore: Training and Evaluation of a Dutch BERT-based Language Model Named BERTje. Master Thesis in Information Science, University of Groningen, The Netherlands.
- Gareth Terence Bryan Dwyer. 2017. Novel approaches to authorship attribution. Master Thesis in Language and Communication Technologies, Information Science, University of Groningen, The Netherlands.
- Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- Olga Feiguina and Graeme Hirst. 2007. Authorship attribution for small texts: Literary and forensic experiments. In *Proceedings of the SIGIR’07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*.
- Oren Halvani, Christian Winter, and Lukas Graner. 2018. Unary and binary classification approaches and their implications for authorship verification. *arXiv preprint arXiv:1901.00399*.
- Matthew Honnibal. 2015. spacy: Industrial-strength natural language processing (nlp) with python and cython.
- Manuela Hürlimann, Benno Weck, Esther van den Berg, Simon Suster, and Malvina Nissim. 2015. Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*.
- Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. *CLEF (Working Notes)*, 1179.
- Patrick Juola. 2015. The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(suppl_1):i100–i113.
- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF (Working Notes)*.
- Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névél, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, September.
- Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62.
- Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.
- Aleksandra Maslennikova, Paolo Labruna, Andrea Cimino, and Felice Dell’Orletta. 2019. Quanti anni hai? Age Identification for Italian. In *Proceedings of 6th Italian Conference on Computational Linguistics (CLiC-it), 13-15 November, 2019, Bari, Italy*.
- Erwan Moreau, Arun Jayapal, Gerard Lynch, and Carl Vogel. 2015. Author verification: basic stacked generalization applied to predictions from a set of heterogeneous learners-notebook for pan at clef 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.
- Abdulfattah Omar, Basheer Ibrahim Elghayesh, and Mohamed Ali Mohamed Kassem. 2019. Authorship attribution revisited: The problem of flash fiction a morphological-based linguistic stylometry approach. *Arab World English Journal (AWEJ) Volume*, 10.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, pages 1613–0073.
- Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237.
- Shachar Seidman. 2013. Authorship verification using the impostors method. In *CLEF 2013 Evaluation labs and workshop–Working notes papers*, pages 23–26. Citeseer.

- Efstathios Stamatatos and Moshe Koppel. 2011. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation*, 45(1):1–4.
- Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*, Sheffield, UK, 2014, pages 1–21.
- Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the author identification task at pan 2015. clef 2015 evaluation labs and workshop, online working notes, toulouse, france. In *CEUR Workshop Proceedings*, pages 1–17.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.
- Michail Tsikerdekis and Sherali Zeadally. 2014. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Transactions on Information Forensics and Security*, 9(8):1311–1321.
- Arjuna Tuzzi and Michele A Cortelazzo. 2018. *Drawing Elena Ferrante’s Profile: Workshop Proceedings, Padova, 7 September 2017*. Padova UP.
- Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.
- Min Yang and Kam-Pui Chow. 2014. Authorship attribution for forensic investigation with thousands of authors. In *IFIP International Information Security Conference*, pages 339–350. Springer.

The Archaeo-Term Project: Multilingual Terminology in Archaeology

Giulia Speranza, Raffaele Manna, Maria Pia di Buono, Johanna Monti

UniOr NLP Research Group
“L’Orientale” University of Naples
Italy

{gsperanza, rmanna, mpdibuono, jmonti}@unior.it

Abstract

In this paper, we present the Archaeo-Term Project, along with one of its first efforts in enhancing multilingual access to Archaeological data, making available a resource of Archaeological terms within the framework of YourTerm CULT project. In order to enhance and promote the use of a terminological common ground across different languages the Archaeo-Term multilingual Glossary is intended both for scholars, experts in the field, translators and the general public. Its first release contains terms in Italian, English, German, Spanish and Dutch together with PoS, definitions and other linguistic information. This paper presents the data and the methodology adopted to create the glossary as well as the evaluation of the first results.

1 Introduction

Languages for Special Purposes (LSP) have their roots in the need of communicating specialised and technical knowledge within a restricted group of domain experts.

From a linguistic perspective, LSP are mainly characterised by the use of specialised terminology, which is usually monosemous for the principle of clearly defining concepts and avoiding miscommunication and can often result opaque and unintelligible to laypeople (Gotti, 2008; Cabré, 1999; Faber and Rodríguez, 2012; Crystal, 1997). In fact, for these reasons, it is often necessary to modulate specialised languages when both oral and written communication takes place between expert and non-experts, in order to ease the di-

dactic and informative functions of communication (Cortelazzo, 1994).

The language used in the domain of Cultural Heritage (CH), and its sub-domains, such as Archaeology, shares many points with other LSPs, such as the presence of technical terminology, terms of Greek and Latin origins, re-semanticisation of common words into specialised domains of knowledge, complex multiword expressions, to mention a few. Nonetheless, it has been traditionally less investigated if compared to, for example, the language of medicine or law, which are considered soft disciplines too. As a consequence, except for a few felicitous examples (see Section 2), language resources and especially terminological resources, in this domain, are still needed.

Language resources such as glossaries, thesauri, dictionaries and term-banks are invaluable sources for language experts, translators, learners, among others. Their development can often be demanding and time-consuming, especially when carried out manually.

Specialised domain resources are even more challenging because their creation also needs the validation of experts in the domain of knowledge.

In this paper we present our work aimed at the creation of a multilingual glossary of archaeological terms, which is useful in many application scenarios from Machine Translation (MT) to Natural Language Processing (NLP).

The remainder of the paper is organized as follows: Section 2 describes related work and, following this, Section 3 presents the Archaeo-Term Project’s aims and the creation of the multilingual glossary of archaeological terms, along with the description of the starting data used so far, namely the ICCD Thesaurus, and the methodology applied to extract multilingual data from the Getty AAT. To complete this section, we illustrate the first results together with their evaluation. Finally, the paper ends with the conclusions and the future

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

work.

2 Related Work

Terminology, as several scholars pointed out (Wright et al., 2010; Melby, 2012), may sometimes result in a heterogeneous activity involving different formats, data models and practices; therefore, in order to support the sharing and the reuse of terminological resources, several standard formats have been developed, such as TermBase eXchange (TBX) (Melby, 2015).

More recently, with the spreading of the Semantic Web Technologies, many language resources are being released in compliance with the Linked Open Data (LOD) principles, using formalisms such as SKOS and Ontolex-Lemon, which are based on the Resource Description Framework (RDF), for representing glossaries, vocabularies and taxonomies (Chiarcos et al., 2013).

In the field of CH some language resources have been released during the years, both monolingual and multilingual. Among the multilingual resources, the most referred one in this domain is the Art & Architecture Thesaurus (AAT)², developed and maintained by The Getty Research Institute. It is a multilingual thesaurus used to describe art, architecture, decorative arts, material culture, and archival materials, which can be accessed through a web interface or via its LOD version (JSON, RDF, N3/Turtle, N-Triples), as well as XML and relational tables.

Another multilingual terminological project on CH is the iDAI.vocab³, a controlled vocabulary specifically designed for archaeological terms available in several languages, developed by the German Archaeological Institute (DAI).

Many other glossaries and thesauri have been created as monolingual resources for cataloguing purposes. Such as the vocabularies developed by the FISH (Forum on Information Standards in Heritage)⁴ and maintained as LOD resources by the Heritage Data⁵ for English, or the thesauri and controlled vocabularies developed by the Italian Institute for Cataloguing (Istituto Centrale per Il

Catalogo e La Documentazione - ICCD)⁶.

The ICCD has also started, in 2017, the ArCo project⁷ together with l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del CNR, in order to make available data from the General Catalogue of Cultural Heritage according to the LOD principles (Carriero et al., 2019b; Carriero et al., 2019a).

Some glossaries are also released by the museums or cultural institutions such as the British Museum's Object Names Thesaurus⁸.

In the field of Cultural Heritage in general, and particularly, in archaeology, it is worth mentioning the ARIADNE Project (Meghini et al., 2017) which provides a portal for the collection of data and resources in order to overcome the fragmentation of archaeological data repositories of all types.

3 Archaeo-Term project

The Archaeo-Term project of the UNIOR NLP Research Group⁹ of the University of Naples "L'Orientale" is part of the YourTerm CULT initiative¹⁰ in partnership with the Terminology Without Borders program fostered by the Terminology Coordination Unit (TermCoord)¹¹ of the European Parliament's Directorate-General for Translation (DG TRAD). Among the different projects, YourTerm CULT is specifically designed to operate in all aspects of culture.

The Archaeo-Term project has been launched to fill the gap in an important field which takes us back to the roots of European culture and history, namely Archaeology.

The project aims at improving the accessibility of the archaeological information available in various sources (scientific papers, texts addressed to general audiences, web sites, structured databases, etc.) by creating language resources useful to NLP and MT tasks across languages. This will ease the availability of the information that can be used to structure and connect different types of knowledge bases together, both structured databases and un-

²<https://www.getty.edu/research/tools/vocabularies/aat/about.html>

³<https://archwort.dainst.org/it/vocab/index.php>

⁴<http://www.heritage-standards.org.uk/terminology/>

⁵<https://www.heritagedata.org/blog/vocabularies-provided/>

⁶<http://www.iccd.beniculturali.it/it/strumenti-terminologici>

⁷<http://stlab.istc.cnr.it/stlab/project/arco/>

⁸<http://terminology.collectionstrust.org.uk/British-Museum-objects/>

⁹<https://sites.google.com/view/unior-nlp-research-group>

¹⁰<https://yourterm.org/yourterm-cult/>

¹¹<https://termcoord.eu/>

structured text collections.

Indeed, although some scientific communities felt the need to structure their knowledge by means of thesauri or ontologies, the scenario is still very fragmented as posed by Felicetti et al. (2018). Nowadays, European archaeological documentation consists of a multifaceted series of information, produced in different and independent ways by each of the various national and international institutions active in this discipline, by means of tools and methods that are often very different from each other. Thus, there is still the need to establish a terminological common core shared across languages.

In this scenario, the Archaeo-Term project tries to contribute to the improvement of scientific cooperation and advancements by attracting both academia and museums from different countries in the creation of a wide multilingual terminological resource in Archaeology. With this aim in mind, one of the first results of this project is a multilingual Glossary of archaeological terms which is mainly useful for the multilingual digitalisation efforts of the museums, but also to scholars, translators and the general public.

3.1 Data and Methodology

For the creation of the Archaeo-Term multilingual glossary, we start from the RDF/SKOS version of the Italian ICCD Thesaurus¹², one of the best practices adopted by the Italian Ministry of Cultural Heritage (MiBAC) to publish institutional information as LOD, in order to be easily findable, reused and freely shared. It contains 1,059 Italian terms which are linked to the LOD version of the Getty AAT¹³, by means of the `skos:closeMatch` property pointing to the Getty URIs (Figure 1). This property is used to link two similar concepts that can be used interchangeably in some information retrieval applications (Cfr. SKOS Recommendation 18 August 2009). We choose to extract the information stored into the Getty AAT because it is a valuable and trustworthy resource, created by experts in the field.

¹²https://github.com/ICCD-MiBACT/Standard-catalografici/blob/master/strumenti-terminologici/beni%20archeologici/ICCD_Thesaurus_definizione%20del%20bene_reperti%20archeologici.rdf

¹³For the mapping process see the ARIADNE project described in Felicetti et al. (2015)

The exploitation of the ICCD resource to read URIs pointing to Getty AAT contributes to build our multilingual glossary of archaeological terms along with the corresponding definitions and sources in other languages, namely English, Spanish, German and Dutch. Among the many languages available in the Getty AAT, we decide to use for our glossary those mentioned above since they show the best coverage in terms of linguistic equivalence (translations) starting from the Italian terms in the ICCD thesaurus.

In order to perform this, we use the Getty AAT SPARQL Endpoint¹⁴ to access term related information by means of setting queries. In detail, the querying process consists of a matching operation between the results of integrated queries in the AAT SPARQL Endpoint.

We first use a query capable of parsing the ICCD resource and reading each URI which refers to the corresponding English archaeological term. In fact, in the Getty AAT, English terms and other available corresponding terms in different languages are represented as equivalent terms by means of the `skos:prefLabel` property¹⁵ and as alternative terms in `skos:altLabel` property¹⁶. Both properties carry one lexical value and one language tag, associated with the lexical value, for each URI.

Since we try to extract corresponding terms in different languages, we then perform a further query able to extract archaeological equivalent terms along with their language tags and alternative terms along with language tags for each available language per URIs.

In addition to this, we set another query able to read URIs and collect corresponding definitions and sources along with their language tags, (both contained in the `skos:scopeNote` property)¹⁷. As a first result of such a query looping over ICCD URIs, we collect archaeological terms, definitions and sources. These queries guarantee the exploitation of the Getty AAT resource but, regardless of the language tags, also a combination of each term value associated with each definition and source value (present in the `skos:scopeNote`).

¹⁴<http://vocab.getty.edu/sparql>

¹⁵https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_prefLabel.html

¹⁶https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_altLabel.html

¹⁷https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_scopeNote.html

To the best of our knowledge, in the AAT we did not find a direct link between the different language terms values (stored in `skos:prefLabel` and `skos:altLabel` and the different language literal values (definitions and sources in `skos:scopeNote`) represented for the same URI. Therefore, to build our multilingual glossary we rely on a matching operation between URIs and language tags related to term values (represented in `skos:prefLabel` and `skos:altLabel`), definitions and sources (both represented in `skos:scopeNote`).

In particular, starting from a combination of all term values and literal values (definitions and sources) per language present for an URI, we apply a matching operation able to select only the terms, definitions and sources concerning the same language based on the reference URI. This matching operation allows us to recognise and organise archaeological terms and their literal values, that is definitions and sources, pertaining to the same language for each archaeological term identified by URI.

3.2 Results and Evaluation

Once the queries steps are performed, we first replace retrieved URIs with numeric IDs in order to provide an identification code for each entry of our glossary; then we build monolingual tables for each language mentioned above and a multilingual synoptic table.

For monolingual tables, we automatically classify in separated tables all retrieved data based on the language tag for each term entry. On the other hand, we align the terms in the different languages based on the shared ID to build the multilingual synoptic table.

In detail, the Glossary first release¹⁸ is organised as follows:

- For each language foreseen in the glossary (Italian, English, Spanish, German and Dutch) there is a dedicated monolingual table, named after the corresponding language locale (e.g., IT for Italian, EN for English) which contains 8 fields (ID, Singular Term, Plural Term, Qualifier¹⁹, PoS, Alternative

Terms, Alternative Terms Qualifier, Definition and Source) as shown in figure 2.

- a multilingual synoptic table contains all the languages singular terms, which are linked to one another by means of the IDs. This multilingual table aims at providing a comprehensive overview on the equivalent terms across the languages.

During the evaluation phase, we noticed that 9 Italian terms had two equivalent English terms in the Getty AAT, marked by two `closeMatch` URIs to the AAT instead of just one.

A manual evaluation revealed that one URI leads to a more generic term and the other one to a more specific term. For example the Italian term *letto* is linked both to the Getty AAT ‘Bed’ (generic) and to ‘Canopy Bed’ (specific). In these cases, instead of following the URI pointing to the specific reference, we choose to follow the most generic one, in accordance with the Italian term meaning. We opt for a manual evaluation due to the low presence of this phenomenon, but, alternatively, it could have been performed automatically making use of an external resource such as a dedicated dictionary.

Furthermore, the evaluation phase revealed a difference in the granularity of terms between the Italian ICCD Thesaurus and the other languages coming from the Getty AAT. Indeed, while the Italian terms result to be highly specific and fine-grained, many equivalents in the other languages are more in a relation of hyperonymy/hyponymy. For example, in the Italian Thesaurus there are several semantically and linguistically different types of relieves: their meanings change according to the following adjectives (e.g., *Rilievo + storico, funerario, votivo*, could be in English historical, funerary, votive + Relief). Nonetheless, the retrieved equivalent in English extracted from the Getty AAT is always ‘Relief’, as well as in Spanish is always ‘Relieve’ and in Dutch is ‘Reliëf’.

Finally, some terms in the different languages, as well as some definitions, are missing and we plan to implement the missing fields in the future. Table 1 shows the total number of terms for each language in the terminological database. Missing fields are due to data sparsity, since for each Italian term there are not always equivalent terms in

¹⁸<https://drive.google.com/file/d/1cKvZPd6bdh7lrZ6pljlgGKatWvopqFo4/view>
The Glossary is released under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

¹⁹The ‘Qualifier’ field, enclosed between brackets, indi-

cates the subfield the term belongs to, thus allowing the disambiguation in case of homographs (e.g. *Ax (weapon)* vs. *Ax (tool)*)

```

</rdf:Description>
<rdf:Description rdf:about="009.005.000.126">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <rdfs:label xml:lang="it">stamnos</rdfs:label>
  <skos:narrower rdf:resource="009.005.000.126.003"/>
  <skos:narrower rdf:resource="009.005.000.126.004"/>
  <skos:narrower rdf:resource="009.005.000.126.001"/>
  <skos:narrower rdf:resource="009.005.000.126.002"/>
  <skos:narrower rdf:resource="009.005.000.126.007"/>
  <skos:narrower rdf:resource="009.005.000.126.008"/>
  <skos:narrower rdf:resource="009.005.000.126.005"/>
  <skos:narrower rdf:resource="009.005.000.126.006"/>
  <skos:broader rdf:resource="009.005"/>
  <skos:prefLabel xml:lang="it">stamnos</skos:prefLabel>
  <skos:definition xml:lang="it">Recipiente capace, col collo breve, corpo espanso, a
  lte spalle, due anse quasi orizzontali e spesso fornito di coperchio; serviva per contene
  re olio, vino e anche monete.</skos:definition>
  <skos:editorialNote xml:lang="it">Vedi anche: Vasellame metallico - ICCD [In rete]
  iccd.beniculturali.it/getFile.php?id=179 (05 marzo 2018); Dizionario oggetto (OGTD-
  OGTT): Vetri [In rete] iccd.beniculturali.it/getFile.php?id=175 (05 marzo 2018)</skos:edi
  torialNote>
  <foaf:depiction rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_arc
  heologici/immagini/th-ra_009.005.000.126.jpg"/>
  <skos:editorialNote xml:lang="it">Immagine tratta da: http://www.metmuseum.org/toah
  /images/h2/h2_06.1021.178.jpg</skos:editorialNote>
  <skos:closeMatch rdf:resource="http://vocab.getty.edu/page/aat/300198881"/>
  <skos:inScheme rdf:resource=""/>
</rdf:Description>

```

Figure 1: Sample of the Italian term entry “stamnos” in the ICCD RDF/SKOS formalism.

ID	Singular Term	Plural Term	Qualifier	PoS	Alternative Terms	Alternative Terms Qualifier	Definition	Source
18	patera	patterae	(container)	Noun	pateras	(containers)	Ancient Roman containers in the form of a shallow bowl without handles, often with a base whose center is pushed up into the body; used for offering libations at religious ceremonies or for drinking. For similar ancient Greek containers, use "phiae."	Legacy Art & Architecture Thesaurus (AAT) data. Compiled without citing sources. Warranted by AAT staff. 1983-1995.
140	fish plate	fish plates	(ancient dish)	NP (Noun + Noun)	fish-plates	(ancient dishes)	Plates of a special form used by the ancient Greeks, having a central depression and sometimes a turned-down rim, used for serving fish. The central depression was used to collect the juice or sauce in which the fish was served. [...]	J. Paul Getty Museum. [online] Los Angeles: J. Paul Getty Trust, 2-. http://www.getty.edu/art/collections/ (1 January 23).
186	tympanum	tympanums	(wall component)	Noun	tympan	(wall component)	Architectural elements comprising stone or masonry enclosed by an arch, usually supported by a lintel. Tympana are normally set above doors, but also occur in windows and wall arcades. They may be ornamented with sculptural or painted decoration.	Harris, Cyril M., ed. Dictionary of Architecture and Construction. New York: McGraw-Hill Book Co., 1975. Grove Art Online. Oxford University Press, 28-. http://www.oxfordartonline.com (1 July 28).
521	aryballos	aryballoi	(Greek vessels)	Noun	aryballas aryballes aribalos aribalo	(Greek vessels)	Relatively small ancient Greek vessels with a globular body, a short neck, a flat disk-shaped mouth with a small orifice, and a handle (or sometimes two) extending from the shoulder to the rim; used for holding oils, perfumes, and ointments. They are usually made of terracotta. Uses of the aryballoi included in funeral rituals and by athletes who wore them on their wrists, suspended by thongs or strings.	Cook, R. M. Greek Painted Pottery. London: Methuen and Co., Ltd., 1966.

Figure 2: Example of the English monolingual table.

all the other languages.

Language	Terms
Italian (IT)	1059
English (EN)	1026
Dutch (NL)	900
Spanish (ES)	593
German (DE)	376

Table 1: Number of terms for each language in the termbase.

4 Conclusions and future works

In this paper we present our Archaeo-Term Project aimed at the creation of a multilingual glossary on archaeology. The Glossary is the result of an extraction and merging process from two already available resources released according to the RDF Data Model, namely the RDF/SKOS version of the Italian ICCD Thesaurus and the LOD version of the multilingual Getty AAT.

The Archaeo-Term glossary is an ongoing project which will address, as future steps, the completion of missing data (terms, definitions, correspondences, examples, etc.) for English, Dutch, Spanish and German, as well as the enlargement of the glossary on the basis of the semi-automatic extraction of terminology from specialised corpora and other existing glossaries for the languages currently foreseen.

Furthermore, we also plan to implement the glossary with other languages such as French, Swedish, Chinese and Russian.

As future work we also plan to convert the result of Archaeo-Term project into more formalised formats, i.e., both TBX format (TermBase eXchange) to be used in connection with CAT-Tools and Ontolex-Lemon Model (McCrae et al., 2017), following the Linguistic Linked Open Data (LLOD) principles.

Finally, when we achieve a more complete version of the glossary we plan to publish it also on a Research Infrastructure Repository such as CLARIN.

Acknowledgments

This work has been partially supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018

and by POR Campania FSE 2014-2020 “Dottorati di Ricerca a Caratterizzazione Industriale”.

We would like to thank Michele Stefanile for his support as expert in the domain of Archaeology. Authorship Attribution is as follows: Giulia Speranza is author of Section 2 and 3.2, Raffaele Manna is author of Section 3.1, Maria Pia di Buono is author of Section 1 and Johanna Monti is author of Section 3 and 4.

References

- Maria Teresa Cabré. 1999. *Terminology: Theory, methods, and applications*, volume 1. John Benjamins Publishing.
- Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019a. Arco ontology network and lod on italian cultural heritage. In *ODOCH@ CAiSE*, pages 97–102.
- Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019b. Arco: The italian cultural heritage knowledge graph. In *International Semantic Web Conference*, pages 36–52. Springer.
- Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.
- Michele Cortelazzo. 1994. *Lingue speciali. La dimensione verticale*, Padova.
- David Crystal. 1997. *The cambridge encyclopedia of language*, wyd. 2. New York.
- Pamela Faber and Clara Inés López Rodríguez. 2012. 2.1 terminology and specialized language. *A cognitive linguistics view of terminology and specialized language*, 20:9.
- Achille Felicetti, Ilenia Galluccio, Cinzia Luddi, Maria Letizia Mancinelli, Tiziana Scarselli, and Antonio Davide Madonna. 2015. Integrating terminological tools and semantic archaeological information: the iccd ra schema and thesaurus. In *EMF-CRM@ TPDF*, pages 28–43.
- Achille Felicetti, Daniel Williams, Ilenia Galluccio, Douglas Tudhope, and Franco Niccolucci. 2018. Nlp tools for knowledge extraction from italian archaeological free text. In *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, pages 1–8. IEEE.

- Maurizio Gotti. 2008. *Investigating specialized discourse*. Peter Lang.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Carlo Meghini, Roberto Scopigno, Julian Richards, Holly Wright, Guntram Geser, Sebastian Cuy, Johan Fihn, Bruno Fanini, Hella Hollander, Franco Nicolucci, et al. 2017. Ariadne: A research infrastructure for archaeology. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(3):1–27.
- Alan K Melby. 2012. Terminology in the age of multilingual corpora. *The Journal of Specialised Translation*, 18:7–29.
- Alan Melby. 2015. Tbx: A terminology exchange format for the translation and localization industry. 201), *Handbook of Terminology*, pages 393–424.
- Sue Ellen Wright, Nathan Rasmussen, Alan K Melby, and L Warburton. 2010. Tbx glossary: a crosswalk between termbase and lexbase formats. In *Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop*.

Exploiting Distributional Semantics Models for Natural Language Context-aware Justifications for Recommender Systems

Giuseppe Spillo (*), Cataldo Musto, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro

University of Bari - Dip. di Informatica

(*) Bachelor Student - email: giuseppe.spillo@studenti.uniba.it - others: name.surname@uniba.it

Abstract

In this paper¹ we present a methodology to generate *context-aware natural language justifications* supporting the suggestions produced by a recommendation algorithm. Our approach relies on a natural language processing pipeline that exploits *distributional semantics models* to identify the most relevant *aspects* for each different context of consumption of the item. Next, these aspects are used to identify the most suitable pieces of information to be combined in a natural language justification. As information source, we used a corpus of reviews. Accordingly, our justifications are based on a combination of *reviews' excerpts* that discuss the aspects that are particularly relevant for a certain context.

In the experimental evaluation, we carried out a user study in the *movies* domain in order to investigate the validity of the idea of adapting the justifications to the different contexts of usage. As shown by the results, all these claims were supported by the data we collected.

1 Introduction

Recommender Systems (RSs) (Resnick and Varian, 1997) are now recognised as a very effective mean to support the users in decision-making tasks (Ricci et al., 2015). However, as the importance of such technology in our everyday lives grows, it is fundamental that these algorithms support each suggestion through a *justification* that allows the user to understand the *internal mechanisms* of the recommendation process and to more easily discern among the available alternatives.

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To this end, several attempts have been recently devoted to investigate *how* to introduce explanation facilities in RSs (Nunes and Jannach, 2017) and to identify the most suitable explanation styles (Gedikli et al., 2014). Despite such a huge research effort, none of the methodologies currently presented in literature diversifies the justifications based on the different *contextual situations* in which the item will be consumed. This is a clear issue, since context plays a key role in every decision-making task, and RSs are no exception. Indeed, as the *mood* or the company (*friends, family, children*) can direct the choice of the movie to be watched, so a justification that aims to convince a user to enjoy a recommendation should contain different concepts depending on whether the user is planning to watch a movie with her *friends* or with her *children*.

In this paper we fill in this gap by proposing an approach to generate a *context-aware* justification that supports a recommendation. Our methodology exploits distributional semantics models (Lenci, 2008) to build a *term-context* matrix that encodes the importance of terms and concepts in each context of consumption. Such a matrix is used to obtain a vector space representation of each context, which is in turn used to identify the most suitable pieces of information to be combined in a justification. As information source, we used a corpus of reviews. Accordingly, our justifications are based on a combination of *reviews' excerpts* that discuss with a *positive* sentiment the aspects that are particularly relevant for a certain context. Beyond its *context-aware nature*, another distinctive trait of our methodology is the fact that we generate *post-hoc justifications* that are completely independent from the underlying recommendation models and completely separated from the step of generating the recommendations.

To sum up, we can summarize the contributions of the article as follows: (i) we propose a method-

ology based on *distributional semantics models* and *natural language processing* to automatically learn a vector space representation of the different *contexts* in which an item can be consumed; (ii) We design a pipeline that exploits *distributional semantics models* to generate *context-aware natural language justifications* supporting the suggestions returned by any recommendation algorithm;

The rest of the paper is organized as follows: first, in Section 2 we provide an overview of related work. Next, Section 3 describes the main components of our workflow and Section 4 discusses the outcomes of the experimental evaluation. Finally, conclusions and future work of the current research are provided in Section 5.

2 Related Work

The current research borrows concepts from *review-based explanation strategies* and *distributional semantics models*. In the following, we will try to discuss relevant related work and to emphasize the hallmarks of our methodology.

Review-based Explanations. According to the taxonomy discussed in (Friedrich and Zanker, 2011), our approach can be classified as a *content-based explanation strategy*, since the justifications we generate are based on descriptive features of the item. Early attempts in the area rely on the exploitation of tags (Vig et al., 2009) and features gathered from knowledge graphs (Musto et al., 2016). With respect to classic content-based strategies, the novelty of the current work lies in the use of *review data* to build a natural language justification. In this research line, (Chen and Wang, 2017) Chen et al. analyze users' reviews to identify relevant features of the items, which are presented on an *explanation* interface. Differently from this work, we did not bound on a fixed set of static aspects and we left the explanation algorithm deciding and identifying the most relevant concepts and aspects for each contextual setting. A similar attempt was also proposed in (Chang et al., 2016). Moreover, as previously emphasized, a trait that distinguishes our approach with respect to such literature is the adaptation of the justification based on the different setting in which the item is consumed. The only work exploiting *context* in the justification process has been proposed by Misztal et al. in (Misztal and Indurkha, 2015). However, differently from our work, they did not diversify the justifi-

cations of the same items on varying of different contextual settings in which the item is consumed, since they just adopt features inspired by context (e.g., "*I suggest you this movie since you like this genre in rainy days*") to explain a recommendation. **Distributional Semantics Models.** Another distinctive trait of the current work is the adoption of *distributional semantics models* (DSMs) to build a vector space representation of the different *contextual situations* in which an item can be consumed. Typically, DSMs rely on a *term-context matrix*, where rows represent the terms in the corpus and columns represents *contexts of usage*. For the sake of simplicity, we can imagine a context as a fragment of text in which the term appears, as a sentence, a paragraph or a document. Every time a particular term is used in a particular context, such an information is encoded in this matrix. One of the advantages that follows the adoption of DSMs is that they can learn a vector space representation of terms in a totally *unsupervised* way. These methods, recently inspired methods in the area of *word embeddings*, such as WORD2VEC (Mikolov et al., 2013) and contextual word representations (Smith, 2020). Even if some attempts evaluating RSs based on DSMs already exists (Lops et al., 2009; Musto et al., 2011; Musto et al., 2012; Musto et al., 2014), in our attempt we used DSMs to build a *vector-space representation of the different contextual dimensions*. Up to our knowledge, the usage of DSMs for justification purposes this is a completely new research direction in the area of explanation.

3 Methodology

Our workflow to generate context-aware justifications based on users' reviews is shown in Figure 1. In the following, we will describe all the modules that compose the workflow.

Context Learner. The first step is carried out by the CONTEXT LEARNER module, which exploits DSMs to learn a vector space representation of the *contexts*. Formally, given a set reviews R and a set of k contextual settings $C = \{c_1 \dots c_k\}$, this module generates as output a matrix $C_{n,k}$ that encodes the importance of each term t_i in each contextual setting c_j . In order to build such a representation, we first split all the reviews $r \in R$ in sentences. Next, let S be the set of previously obtained sentences, we *manually annotated* a subset of these sentences in order to obtain a

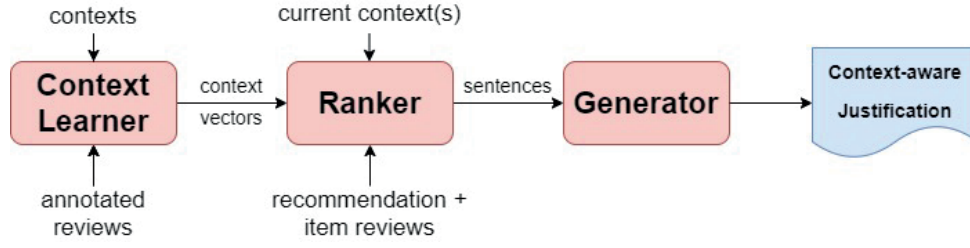


Figure 1: Workflow to generate Context-aware Justifications by Exploiting DSMs

set $S' = \{s_1 \dots s_m\}$, where each s_i is labeled with one or more contextual settings, based on the concepts mentioned in the review. Of course, each s_i can be annotated with more than one context. As an example, a review including the sentence 'a very romantic movie' is annotated with the contexts *company=partner*, while the sentence 'perfect for a night at home' is annotated with the contexts *day=weekday*. After the annotation step, a *sentence-context* matrix $A_{m,k}$ is built, where each a_{s_i,c_j} is equal to 1 if the sentence s_i is annotated with the context c_j (that is to say, it mentions concepts that are relevant for that context), 0 otherwise.

Next, we run *tokenization* and *lemmatization* algorithms (Manning et al., 1999) over the sentences in S to obtain a *lemma-sentence* matrix $V_{n,m}$. In this case, v_{t_i,s_j} is equal to the TF/IDF of the term t_i in the sentence s_j . Of course, IDF is calculated over all the annotated sentences. In order to filter out non-relevant lemmas, we maintained in the matrix V just *nouns* and *adjectives*. Nouns were chosen due to previous research (Nakagawa and Mori, 2002), which showed that descriptive features of an item are usually represented using nouns (e.g., service, meal, location, etc.). Similarly, adjectives were included since they play a key role in the task of catching the characteristics of the different contextual situations (e.g., romantic, quick, etc.). Moreover, we also decided to take into account and extract combinations of nouns and adjectives (bigrams) such as *romantic location*, since they can be very useful to highlight specific characteristics of the item.

In the last step of the process annotation matrix $A_{n,k}$ and vocabulary matrix $V_{m,n}$ are multiplied to obtain our *lemma-context* matrix $C_{n,k}$, which represents the final output returned by the CONTEXT LEARNER module. Of course, each $c_{i,j}$ encodes the importance of term t_i in the context c_j . The whole process carried out by this component is de-

scribed in Figure 2.

Given such a representation, two different outputs are obtained. First, we can directly extract *column vectors* \vec{c}_j from matrix C , which represents the vector space representation of the context c_j based on DSMs. It should be pointed out that such a representation perfectly fits the principles of DSMs since contexts discussed through the same lemmas will share a very similar vector space representation. Conversely, a poor overlap will result in very different vectors. Moreover, for each column, lemmas may be ranked and those having the highest TF-IDF scores may be extracted. In this way, we obtain a *lexicon* of lemmas that are relevant for a particular contextual setting, and this can be useful to empirically validate the effectiveness of the approach. In Table 1, we anticipate some details of our experimental session and we report the top-3 lemmas for two different contextual settings starting from a set of *movie* reviews.

Ranker. Given a recommended item (along with its reviews) and given the context in which the item will be consumed (from now on, defined as 'current context'), this module has to identify the most relevant *review excerpts* to be included in the justification. To this end, we designed a ranking strategy that exploits DSMs and similarity measures in vector spaces to identify suitable excerpts: given a set of n reviews discussing the item i , $R_i = \{r_{i,1} \dots r_{i,n}\}$, we first split each r_i in sentences. Next, we processed the sentences through a sentiment analysis algorithm (Liu, 2012; Petz et al., 2015) in order to filter out those expressing a *negative* or *neutral* opinions about the item. The choice is justified by our focus on review excerpts discussing *positive* characteristics of the item. Next, let c_j be the current contextual situation (e.g., *company=partner*), we calculate the *cosine similarity* between the context vector \vec{c}_j returned by the CONTEXT LEARNER and a vector

$$\begin{pmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,m} \\ v_{2,1} & v_{2,2} & \dots & v_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n,1} & v_{n,2} & \dots & v_{n,m} \end{pmatrix}_{V_{n,m}} \times \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,k} \end{pmatrix}_{A_{m,k}} = \begin{pmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,k} \\ c_{2,1} & c_{2,2} & \dots & c_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n,1} & c_{n,2} & \dots & c_{n,k} \end{pmatrix}_{C_{n,k}}$$

Figure 2: Building a *lemma-context* matrix C by exploiting distributional semantics models

	<i>Attention=high</i>	<i>Attention=low</i>
Unigrams	engaging, attentive, intense	simple, smooth, easy
Bigrams	intense plot, slow movie, life metaphor	easy vision, simple movie, simple plot

Table 1: Top-3 lemmas returned by the CONTEXT LEARNER module for two couples of different contextual settings in the MOVIE domain.

space representation of each sentence \vec{s}_i . The sentences having the highest cosine similarity w.r.t. to the context of usage c_j are selected as the most suitable excerpts and are passed to the GENERATOR.

Generator. Finally, the goal of GENERATOR is to put together the compliant excerpts in a single *natural language justification*. In particular, we defined a *slot-filling* strategy based on the principles of Natural Language Generation (Reiter and Dale, 1997). Such a strategy is based on the combination of a *fixed* part, which is common to all the justifications, and a *dynamic* part that depends on the outputs returned by the previous steps. In our case, the top-1 sentence for *each current* contextual dimension is selected, and the different excerpts are merged by exploiting simple connectives, such as adverbs and conjunctions. An example of the resulting justifications is provided in Table 2.

4 Experimental Evaluation

The experimental evaluation was designed to identify the best-performing configuration of our strategy, on varying of different combinations of the parameters of the workflow (*Research Question 1*), and to assess how our approach performs in comparison to other methods (both context-aware and non-contextual) to generate *post-hoc justifications* (*Research Question 2*). To this end, we designed a user study involving 273 subjects (male=50%, degree or PhD=26.04%, age \geq 35=49.48%, already used a RS=85.4%) in the *movies* domain. Interest in *movies* was indicated as *medium* or *high* by 62.78% of the sample. Our sample was obtained through the *avail-*

ability sampling strategy, and it includes students, researchers in the area and people not skilled with computer science and recommender systems. As in (Tintarev and Masthoff, 2012), whose protocol was took as a reference in several subsequent research in the area of explanation (Musto et al., 2019), we evaluated the following metrics: *transparency*, *persuasiveness*, *engagement* and *trust* through a post-usage questionnaire.

Experimental Design. To run the experiment, we deployed a web application² implementing the methodology described in Section 3. Next, as a first step, we identified the relevant *contextual dimensions* for each domain. Contexts were selected by carrying out an analysis of related work of context-aware recommender systems in the MOVIE domain. In total, we defined 3 contextual dimensions, that is to say, *mood* (*great*, *normal*), *company* (*family*, *friends*, *partner*) and *level of attention* (*high*, *low*). To collect the data necessary to feed our web application, we selected a subset of 300 popular movies (according to IMDB data) discussed in more than 50 reviews in the Amazon Reviews dataset³. This choice is motivated by our need of a large set of sentences discussing the item in each contextual setting. These data were processed by exploiting lemmatization, POS-tagging and sentiment analysis algorithms available in CoreNLP⁴ and Stanford Sentiment Analysis algorithm⁵. tool. Some statistics about the final dataset are provided in Table

²<http://193.204.187.192:8080/filmando-eng>

³<http://jmcauley.ucsd.edu/data/amazon/links.html> - Only the reviews available in the 'Movies and TV' category were downloaded.

⁴<https://stanfordnlp.github.io/CoreNLP/>

⁵<https://nlp.stanford.edu/sentiment/>

Movie	Justification
Company=Partner	You should watch 'Stranger than Fiction'. It is a good movie to watch with your partner because <i>it has a very romantic end. Moreover, plot is very intense.</i>
Company=Friends	You should watch 'Stranger than Fiction'. It is a good movie to watch with friends since <i>the film crackles with laughter and pathos and it is a classy sweet and funny movie.</i>

Table 2: Context-aware justifications for the MOVIE domain. Automatically extracted review excerpts are reported in *italics*.

3.

In order to compare different configurations of the workflow, we designed several variant obtained by varying the *vocabulary* of lemmas. In particular, we compared the effectiveness of simple *unigrams*, of *bigrams* and their merge. In the first case, we encoded in our matrix just single lemmas (*e.g.*, service, meal, romantic, etc.) while in the second we stored combination of nouns and adjectives (*e.g.*, romantic location). Due to space reasons, we can't provide more details about the lexicons we learnt, and we suggest to refer again to Table 1 for a qualitative evaluation of some of the resulting representations. Our representations based on DSMs were obtained by starting from a set of 1,905 annotations for the *movie* domain, annotated by three annotators by adopting a *majority vote* strategy. To conclude, each user involved in the experiment carried out the following steps:

1. *Training, Context Selection and Generation of the Recommendation.* First, we asked the users to provide some basic demographic data and to indicate their interest in *movies*. Next, each user indicated the *context of consumption* of the recommendation, by selecting a context among the different contextual settings we previously indicated (see Figure 3-a). Given the current context, a suitable recommendation was identified and presented to the user. As recommendation algorithm we used a content-based recommendation strategy exploiting users' reviews.
2. *Generation of the Justification.* Given the recommendation and the current context of consumption, we run our pipeline to generate a context-aware justification of the item *adapted to that context*. In this case, we designed a *between-subject* protocol. In particular, each user was randomly assigned to one of the *three* configurations of our pipeline and the output was presented to the user along with the recommendation (see Figure 3-b).

Clearly, the user was not aware of the specific configuration he was interacting with.

3. *Evaluation through Questionnaires.* Once the justification was shown, we asked the users to fill in a post-usage questionnaire. Each user was asked to evaluate *transparency*, *persuasiveness*, *engagement* and *trust* of the recommendation process through a five-point scale (1=strongly disagree, 5=strongly agree). The questions the users had to answer follow those proposed in (Tintarev and Masthoff, 2012). Due to space reasons, we can't report the questions and we suggest to interact with the web application to fill in the missing details.
4. *Comparison to baselines.* Finally, we compared our method to two different baselines in a *within-subject* experiment. In this case, all the users were provided with two different justifications styles (*i.e.*, our context-aware justifications and a baseline) and we asked the users to choose the one they preferred. As for the baselines, we focused on other methodologies to generate post-hoc justifications and we selected (i) a context-aware strategy to generate justifications, which is based on a set of *manually defined relevant terms* for each context; (ii) a method to generate *non-contextual* review-based justifications that relies on the automatic identification of relevant aspects and on the selection of compliant reviews excerpts containing such terms. Such approach partially replicates that presented in (Musto et al., 2020).

Discussions of the Results Results of the first experiment, that allows to answer to *Research Question 1*, are presented in Table 4. The values in the tables represent the average scores provided by the users for each of the previously mentioned questions. As for the *movie* domain, results show that the overall best results are obtained by us-

	#Items	#Reviews	#Sentences	#Positive Sent.	Avg. Sent./Item	Avg. Pos. Sent./Item
MOVIES	307	153,398	1,464,593	560,817	4,770.66	1,826.76

Table 3: Statistics of the dataset



Figure 3: Interaction with the web application.

ing a vocabulary based on *unigrams* and *bigrams*. This first finding provides us with an interesting outcome, since most of the strategies to generate explanations are currently based on *single keywords and aspects*. Conversely, our experiment showed that both *adjectives* as well as couples of co-occurring terms are worth to be encoded, since they catch more fine-grained characteristics of the item that are relevant in a particular contextual setting. Overall, the results we obtained confirmed the validity of the approach. Beyond the increase in TRANSPARENCY, high evaluations were also noted for PERSUASION and ENGAGEMENT metrics. This outcome confirms how the identification of relevant reviews' excerpts can lead to satisfying justifications. Indeed, differently from feature-based justifications, that typically rely on very popular and well-known characteristics of the movie, as the *actors* or the *director*, more specific aspects of the items emerge from users' reviews.

Next, in order to answer to **Research Question 2**, we compared the best-performing configurations emerging from Experiment 1 to two different baselines. The results of these experiments are reported in Table 5 which show the percentage of users who preferred our context-aware methodology based on DSMs to both the baselines. In particular, the first comparison allowed us to assess the effectiveness of a *vector space* representation of contexts based on DSMs with respect to a simple context-aware justification method based

on a *fixed* lexicon of relevant terms, while the second comparison investigated how valid was the idea of diversifying the justifications based on the different contextual settings in which the items is consumed. As shown in the table, our approach was the preferred one in both the comparisons. It should be pointed out that the gaps are particularly large when our methodology is compared to a *non-contextual* baseline. In this case, we noted a statistically significant gap ($p \leq 0.05$) for all the metrics, with the exception of *trust*. This suggests that diversifying the justifications based on the *context of consumption* is particularly appreciated by the users. This confirms the validity of our intuition, which led to a completely new research direction in the area of justifications for recommender systems.

5 Conclusions and Future Work

In this paper we presented a methodology that exploits DSMs to build post-hoc *context-aware* natural language justifications supporting the suggestions generated by a RS. The hallmark of this work is the diversification of the justifications based on the different contextual settings in which the items will be consumed, which is a new research direction in the area. As shown in our experiments, our justifications were largely preferred by users. This confirms the effectiveness of our approach and paves the way to several future research directions, such as the definition of *personalized* justi-

<i>Metrics / Configuration</i>	<i>Unigrams</i>	<i>Bigrams</i>	<i>Uni+Bigrams</i>
<i>Transparency</i>	3.38	3.81	3.64
<i>Persuasion</i>	3.56	3.62	3.54
<i>Engagement</i>	3.54	3.72	3.70
<i>Trust</i>	3.44	3.66	3.61

Table 4: Results of Experiment 1 for the MOVIE domain. The best-performing configuration is reported in **bold** and underlined

<i>Metrics / Choice</i>	vs. Context-aware Static Baseline			vs. Non-Contextual Baseline		
	<i>CA+DSMs</i>	<i>Baseline</i>	<i>Indifferent</i>	<i>CA+DSMs</i>	<i>Baseline</i>	<i>Indifferent</i>
<i>Transparency</i>	52.38%	38.10%	19.52%	53.21%	34.47%	12.32%
<i>Persuasion</i>	54.10%	36.33%	19.57%	55.17%	32.33%	12.50%
<i>Engagement</i>	49.31%	39.23%	11.56%	44.51%	32.75%	22.74%
<i>Trust</i>	42.86%	39.31%	17.83%	42.90%	42.11%	14.99%

Table 5: Results of Experiment 2, comparing our approach (*CA+DSMs*) to a context-aware baseline that does not exploit DSMs (*CA Static*) and to a non-contextual baseline that exploit users’ reviews (*review-based*). The configuration preferred by the higher percentage of users is reported in **bold**.

fication as well as the generation of *hybrid* justifications that combine elements gathered from user-generated content (as the reviews) with descriptive characteristics of the items. Finally, we will also evaluate the use of ontologies and rules (Laera et al., 2004) in order to implement reasoning mechanisms to better identify the most relevant aspects in the reviews.

References

- [Chang et al.2016] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based Personalized Natural Language Explanations for Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 175–182. ACM.
- [Chen and Wang2017] Li Chen and Feng Wang. 2017. Explaining Recommendations based on Feature Sentiments in Product Reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 17–28. ACM.
- [Friedrich and Zanker2011] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32(3):90–98.
- [Gedikli et al.2014] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382.
- [Laera et al.2004] Loredana Laera, Valentina Tamma, Trevor Bench-Capon, and Giovanni Semeraro. 2004. Sweetprolog: A system to integrate ontologies and rules. In *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pages 188–193. Springer.
- [Lenci2008] Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.
- [Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- [Lops et al.2009] Pasquale Lops, Marco de Gemmis, Giovanni Semeraro, Cataldo Musto, Fedelucio Narducci, and Massimo Bux. 2009. A semantic content-based recommender system integrating folksonomies for personalized access. In *Web Personalization in Intelligent Environments*, pages 27–47. Springer.
- [Manning et al.1999] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.
- [Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- [Misztal and Indurkha2015] Joanna Misztal and Bipin Indurkha. 2015. Explaining contextual recommendations: Interaction design study and prototype implementation. In *IntRS@ RecSys*, pages 13–20.
- [Musto et al.2011] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. 2011. Random indexing and negative user preferences for enhancing content-based

- Recommender Systems. In *EC-Web 2011*, volume 85 of *Lecture Notes in Business Inf. Processing*, pages 270–281. Springer.
- [Musto et al.2012] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. De Gemmis, M. Barbieri, J. Korst, V. Pronk, and R. Clout. 2012. Enhanced semantic tv-show representation for personalized electronic program guides. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7379 LNCS:188–199. cited By 19.
- [Musto et al.2014] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. 2014. Combining distributional semantics and entity linking for context-aware content-based recommendation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8538:381–392. cited By 19.
- [Musto et al.2016] Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2016. Explod: A framework for explaining recommendations based on the linked open data cloud. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys ’16, pages 151–154, New York, NY, USA. ACM.
- [Musto et al.2019] Cataldo Musto, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2019. Justifying recommendations through aspect-based sentiment analysis of users reviews. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 4–12.
- [Musto et al.2020] Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2020. Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction*, pages 1–45.
- [Nakagawa and Mori2002] Hiroshi Nakagawa and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14*, pages 1–7. Association for Computational Linguistics.
- [Nunes and Jannach2017] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5):393–444.
- [Petz et al.2015] Gerald Petz, Michał Karpowicz, Harald Fürschuß, Andreas Auinger, Václav Střiteský, and Andreas Holzinger. 2015. Reprint of: Computational approaches for mining user’s opinions on the web 2.0. *Information Processing & Management*, 51(4):510–519.
- [Reiter and Dale1997] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.
- [Resnick and Varian1997] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Communications of the ACM*, 40(3):56–58.
- [Ricci et al.2015] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer.
- [Smith2020] Noah A Smith. 2020. Contextual word representations: putting words into computers. *Communications of the ACM*, 63(6):66–74.
- [Tintarev and Masthoff2012] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *UMUAI*, 22(4-5):399–439.
- [Vig et al.2009] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56. ACM.

MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion Analysis for Italian

Rachele Sprugnoli

CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
rachele.sprugnoli@unicatt.it

Abstract

English. This paper¹ presents a new linguistic resource for Italian, called **MULTIEMOTIONS-IT**, containing comments to music videos and advertisements posted on YouTube and Facebook. These comments are manually annotated according to four different dimensions: i.e., relatedness, opinion polarity, emotions and sarcasm. For the annotation of emotions we adopted the Plutchik's model taking into account both basic and complex emotions, i.e. dyads.

1 Introduction

Emotions play an influential role in consumer behaviour affecting the decision to purchase goods and services of different types, including music (Mizerski and White, 1986; Lacher, 1989). Both positive and negative emotions have an influence and this is why marketing strategies have always focused on both rational and emotional aspects (Cotte and Ritchie, 2005).

With the advent of social media, platforms such as YouTube and Facebook have gained importance in the marketing industry because they allow to connect and engage consumers (Kujur and Singh, 2018). The progressive consolidation of social media as marketing spaces has highlighted the need to monitor unstructured data written by social media users. In this context, the application of Sentiment Analysis techniques have flourished with the aim of tracking customers' opinions and attitudes by analysing comments or reviews posted on social media channels (Micu et al., 2017).

In this paper we present a new linguistic resource for Italian, called **MULTIEMOTIONS-IT**, con-

taining comments to music videos and advertisement posted on YouTube and Facebook. Comments are manually annotated according to four different dimensions: relatedness, opinion polarity, emotions and sarcasm. Particular attention is devoted to the annotation of emotions for which we adopted the model proposed by Plutchik (1980). Following Plutchik, we take into consideration both the eight basic emotions (joy, sadness, fear, anger, trust, disgust, surprise, anticipation) and the dyads, that is feelings composed of two basic emotions (e.g., love is a blend of joy and Trust). At the time of writing, **MULTIEMOTIONS-IT** is the only freely available manually annotated dataset for emotion analysis for Italian.²

2 Related Works

The computational study of opinions and emotions falls within the scope of the Sentiment Analysis research field (Liu, 2012). Opinion polarity identification is a task aiming at understanding whether a text is expressing positive, negative or neutral sentiment towards the subject of the text. As for emotions, their analysis follows two main approaches (Buechel and Hahn, 2017): in the first one emotions are classified into discrete categories based on the theories of psychologists such as those of Ekman (Ekman, 1992) and Plutchik whereas in the second approach emotions are represented in a dimensional form using continuous values such as *valence*, *arousal* and *dominance* (the so called VAD model).

Survey papers like the ones by Hakak et al. (2017), Bostan & Klinger (2018) and Kim & Klinger (2019) report on studies that focus on different text genres, mainly news (Strapparava and Mihalcea, 2007), social media (Mohammad, 2012) and literary works (Alm et al., 2005).

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²https://github.com/RacheleSprugnoli/Esercitazioni_SA/tree/master/dataset

Among social media, Twitter is the most studied platform and datasets of annotated tweets are available for different Sentiment Analysis tasks. For emotion analysis see, among others, EmpaTweet (Roberts et al., 2012) and EmoTweet (Liew et al., 2016). The literature also reports works on Facebook posts and YouTube comments with corpora and systems developed for various languages such as English (Preoŧiuc-Pietro et al., 2016), Thai (Sarakit et al., 2015), Bangla (Tripto and Ali, 2018) and Indonesian (Savigny and Purwarianti, 2017). As for Italian, there are several emotion lexicons, for example (Araque et al., 2019; Passaro and Lenci, 2016; Mohammad and Turney, 2013; Mohammad, 2018), but, at the moment, no dataset with annotated emotions has been released yet.³

Similarly to SenTube (Uryupina et al., 2014), MULTIEMOTIONS-IT includes YouTube comments and contains the annotation of opinion polarity: however, we also include comments to Facebook posts and we pay particular attention to the categorical annotation of emotions. More specifically, our emotion annotation is inspired by that proposed by Phan et al. (2016) that goes beyond the classification of only the basic emotions to include Plutchik’s dyads so to better capture the spectrum of human emotional experience.

3 Dataset Development

3.1 Data Collection

Comments were scraped from YouTube and Facebook around mid-April 2020 using “Web Scraper”⁴, an extension for browsers. We focused on two genres of media contents: music videos (MVs) on YouTube and advertisements (Ads) both in the form of short videos (on YouTube and Facebook) and pictures (only on Facebook).

We chose 9 music videos of the songs presented during Sanremo Music Festival 2020 selecting both songs that reached the top of the chart in the contest and those that ranked in the last positions. All those videos had thousands of comments: we downloaded the most recent ones, at least one hundred comments per video. Finding advertising videos with lots of comments on YouTube was more complicated because many brands disable

	YT MVs	YT Ads	FB Ads	Avg
unrelated	0.43	0.43	0.41	0.42
neutral	0.30	0.50	0.34	0.38
positive	0.59	0.78	0.77	0.71
negative	0.49	0.71	0.64	0.61
joy	0.50	0.61	0.63	0.58
trust	0.33	0.42	0.35	0.37
sadness	0.45	0.28	0.43	0.39
anger	0.47	0.67	0.49	0.54
fear	0.13	0.11	0.10	0.11
disgust	0.48	0.53	0.27	0.43
surprise	0.33	0.12	0.12	0.19
anticipation	0.35	0.11	0.15	0.20
sarcasm	0.49	0.34	0.24	0.36

Table 1: Inter-annotator agreement in terms of Krippendorff’s Alpha for YouTube music videos (YT MVs), YouTube advertisements (YT Ads), Facebook advertisements (FB Ads). Last column reports the average across the three categories of comments.

the possibility of adding comments to their channel. In the end, we managed to select 20 videos of various products, mostly of food and services, such as telecommunication and banking. Similar products and services were also chosen on Facebook by downloading the comments from 13 different posts.

3.2 Data Annotation

The annotation was performed in the context of the “Sentiment Analysis” seminar held within the ‘Comunicazione per l’impresa, i media e le organizzazioni complesse’⁵ master’s degree at Università Cattolica del Sacro Cuore in Milan. The annotation process lasted 1 week and involved thirty six students: each student annotated 30 comments for each category (i.e., YouTube MVs, YouTube Ads, Facebook Ads) for a total of 90 comments. Each comment was annotated by two students. It is important to note that students had no previous experience in linguistic annotation but had specific training in the strategic management of communication flows on various media platforms.

Annotation Guidelines. Students were required to annotate the following four dimensions for each comment; a comment may consist of more than one sentence but was analysed as a single unit:

1. Relatedness: does the comment refer to the media content? Is the comment written in a

³Annotated datasets for emotion analysis have been mainly developed in enterprises and are not public, see for example (Bolioli et al., 2013).

⁴<https://webscraper.io/>

⁵EN: “Communication for the enterprise, the media and complex organizations”

COMMENT	UNR	NEU	POS	NEG	JOY	TRU	SAD	ANG	FEA	DIS	SUR	ANT	SAR	EMOTIONS
<i>Saludos desde Argentina!!!</i>	1	0	0	0	0	0	0	0	0	0	0	0	0	
<i>Ha superato diodato con le vius</i>	0	1	0	0	0	0	0	0	0	0	0	0	0	
<i>Frizzante come una Coca light scaduta</i>	0	0	0	1	0	0	0	0	0	1	0	0	1	disgust
<i>Questa canzone mi ha rovinato l'esistenza</i>	0	0	0	1	0	0	1	0	1	0	0	0	0	despair
<i>Idea interessante! Meno l'esecuzione...</i>	0	0	1	1	0	1	1	0	0	0	1	0	0	trust - disappointment

Table 2: Examples of annotation.

	YT MV's	YT Ads	FB Ads
# COMMENTS	1,080	1,080	1,080
# WORDS	17,762	19,603	20,844
unrelated	65	59	51
neutral	84	96	115
positive	896	509	597
negative	46	434	373
joy	397	172	149
trust	797	463	574
sadness	188	213	292
anger	30	211	117
fear	5	29	32
disgust	32	155	48
surprise	174	140	199
anticipation	28	56	42
sarcasm	7	30	8

Table 3: Dataset Statistics.

language other than Italian? Comments that are not related to the media content or that are not written in Italian are to be annotated as **unrelated**.

2. Opinion Polarity: is the comment positive, negative or neutral with respect to the media content? Positive and negative polarities are not mutually exclusive: a comment can have a mixed polarity containing both positive and negative opinions on different aspects of the media content.
3. Emotions: what emotions are expressed in the comment? This dimension applies only to comments with positive or negative opinion polarity. Each comment can be annotated with one or more emotions at the same time: the list of emotions to assign includes Plutchik's basic emotions and dyads. Conflict or mixed emotions can appear in the same comment.
4. Sarcasm: are emotions expressed using sarcasm? Following Gibbs (2000), we define sarcasm as a language device that conveys the

opposite of its literal meaning (Cignarella et al., 2018).

Annotation was carried on using spreadsheets where the aforementioned dimensions were converted into 13 fields: **unrelated**, **neutral**, **positive**, **negative**, **joy**, **trust**, **sadness**, **anger**, **fear**, **disgust**, **surprise**, **anticipation**, **sarcasm**. Each field had to be filled in with a binary value: 0 (the dimension is absent) or 1 (the dimension is present). Spreadsheets contained 4 additional metadata fields: **type**, **title**, **URL**, **comment**. For the annotation, students were provided with the images of Plutchik's "Wheel of Emotions"⁶ and of the combination of emotions in dyads⁷.

Inter-Annotator Agreement. Table 1 reports the results of the inter-annotator agreement (IAA): we measured the Krippendorff's Alpha for each label and for each pair of annotators and then we computed the average for each type of comment. The average across the three type of comments is reported in the table as well. For all the labels, IAA is below the 0.8 threshold usually considered as good reliability for content analysis research (Klaus, 1980; Artstein and Poesio, 2008), however these results are in line with the ones obtained in similar works presenting a multi-label annotation of emotions or the annotation of mixed emotions (Aman and Szpakowicz, 2007; Phan et al., 2016). The analysis of the cases of disagreement revealed several interesting issues: i) labels **unrelated** and **neutral** tended to be confused with each other. For example, the comment *Qualcuno mi sa dire dove si trova il porticato della quinta immagine?* (Can anyone tell me where the portico in

⁶<https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg>

⁷<https://i.pinimg.com/originals/83/93/d6/8393d660082c3124a684edc3cade4607.jpg>

DYADS	love	<i>amo questa musica</i> EN: I love this music
	disappointment	<i>Io non capisco come faccia ad essere fra le ultime questa canzone.</i> EN: I don't understand how this song is ranked so low.
	sentimentality	<i>Mi veniva da piangere.... Ricordavo la vecchia pubblicità</i> EN: It makes me want to cry...I remembered the old advertisement
MIX	trust - disappointment	<i>Bellissima!!! Come possa essere ultima! Mah...</i> EN: Gorgeous!!! How can it be the last! Mah ...
	trust - sentimentality	<i>Io ho pianto. Complimenti a Barilla</i> EN: I cried. Congratulations to Barilla
	love - sentimentality	<i>A te la manina tremava e io piangevo.. ♥</i> EN: Your hand was shaking and I was crying ..♥

Table 4: Top 3 dyads and mixes of emotions in the dataset with associated examples. Dyads mentioned in the table are composed by two basic emotions as follows: love = joy + trust; disappointment = surprise + sadness; sentimentality = trust + sadness.

the fifth image is located?) is related to the content of the video but it is neutral; ii) sarcasm was confused with other forms of figurative language such as metaphors, e.g. *È l'Ibrahimovic dei biscotti: perfetto* (EN: it is the Ibrahimovic of biscuits: perfect); iii) the assignment of positive and negative labels registered the highest scores (average Alpha across the 3 categories: 0.71 for positive and 0.61 for negative). Nevertheless, sometimes annotators failed to distinguish between the annotation of opinion polarity and the annotation of emotions by assigning a negative polarity to comments containing negative emotions. However, the two dimensions do not always match: for example, the comment *sta canzone meritava molto di più* (EN: this song deserved much more) expresses disappointment but also an implicit appreciation for the song and thus a positive opinion polarity. iv) the IAA on the single emotion labels varies greatly: a similar wide variability is reported also in previous works even when dealing with non multi-label annotation (Strapparava and Mihalcea, 2008; Aman and Szpakowicz, 2007).

Creation of the Ground Truth. All comments were manually revised and disagreement were reconciled so to assign gold labels. In this way, we generated a ground truth dataset where the noise coming from the annotation of non-expert annotators was minimized. Moreover, the field emotions was added to the spreadsheets so to make explicit the name of the emotions conveyed by the comments. Table 2 shows the structure of the final dataset (metadata fields are not displayed due to space limitation) and some examples of annotation. In particular, the table reports: an unrelated comment, a neutral comment, a com-

ment with a negative polarity, a basic emotion (i.e. disgust) and sarcasm, a comment with a negative polarity and a dyad (i.e., disgust which is made of sadness and fear), a comment with mixed polarity and mixed emotions.

4 Dataset Analysis

Table 3 summarizes the statistics of our final dataset showing the distribution of labels in the three categories of media content. MULTIEMOTIONS-IT contains 3,240 comments for a total of more than 58,000 tokens. Only 470 comments (14.5% of the whole dataset) have no associated emotions because annotated as unrelated or neutral. Comments with positive opinion polarity are more than those with negative polarity: this is especially evident for YouTube MVs that are mostly commented by supporters of the artists performing in the video. Sarcasm is not a pervasive phenomenon: the number of comments annotated with the corresponding label is marginal, covering 1.6% of the total number of comments with an affective content, i.e. annotated with at least one emotion. More specifically, sarcasm co-occurs with two basic emotions: that is, anger (10 comments) and disgust (9 comments).

As for emotions, trust is the most frequent one: indeed, many comments express admiration towards the media content in different ways, for example by thanking the brand, declaring loyalty to a product or expressing appreciation for a specific feature of the media content (e.g. the location of the video). The emotion trust does not appear in the dataset only as a basic emotion but also in several combinations: indeed, 36.5% of the comments with an affective content

are annotated with a dyad and 18.3% with a mix of emotions. Table 4 reports the 3 most frequent dyads and mixes of emotions in the dataset together with an example. As shown in the table, sentimentality (that is a combination of trust and sadness) plays an important role in Ads that try to induce a deep, overwhelming emotional response. Indeed, sentimentality is an emotion that marketing research has identified as a fundamental purchase decision variable (Morton et al., 2013).

Optimism (anticipation + joy) and pessimism (anticipation + sadness) are not very frequent in the dataset with 65 and 16 occurrences respectively. However, it is interesting to note that they are mainly associated with comments on advertisements related to the COVID-19 pandemic, for example:

- optimism: *All'Italia che, ancora una volta, resiste!* EN: To Italy that, once again, resists!
- pessimism: *mamma mia quanta retorica spicciolafinita l'epidemia staremo tutti ad odiarci e ad insultarci come sempreun paese che non ha senso più di esistere* EN: oh my gosh, how much rhetoric once the epidemic is over we will all be hating and insulting each other as always a country that no longer makes sense to exist

5 Baseline System

To establish a baseline on our data, we developed a simple multi-label classification model using the fastText library (Joulin et al., 2016).⁸ The aim of the model is to assign the correct emotion labels to comments. To this end, we randomly split comments and their annotated emotion labels into train and validation following an 80:20 ratio, thus having 2,592 comments for training and the remaining 648 for testing the performance of the learned classifier on new data. Texts have been lower-cased and punctuation removed. We trained the model with the following parameters:

- learning rate: 0.5
- epochs: 25
- word n-grams: 2
- loss function: one-vs-all

⁸<https://fasttext.cc/docs/en/supervised-tutorial.html>

With the previous setting, we obtained 0.57 Precision, 0.43 Recall and 0.49 F-measure. Only four labels registered a F-measure above 0.5: i.e., trust (0.68), love (0.54), delight (0.53), sentimentality (0.50).⁹

6 Conclusion and Future Work

This paper describes MULTIEMOTIONS-It, a new manually annotated dataset for opinion polarity and emotion analysis made of more than 3,000 comments on music videos and advertisements published on YouTube and Facebook.

As for future work, we plan to: (i) extend the annotation guidelines to distinguish the specific object towards which the opinion is directed (e.g. the product, the actor, the location of the video) following the work by Severyn et al. (2016), (ii) extend the dataset with new comments taken also from Instagram and Twitter, (iii) extract a new word-emotion association lexicon from MULTIEMOTIONS-It using vector space models (Passaro et al., 2015) in order to cover complex emotions.

Acknowledgements

The author wants to thank the students of the “Sentiment Analysis” seminar held within the “Comunicazione per l’impresa, i media e le organizzazioni complesse” master’s degree at Università Cattolica del Sacro Cuore (Milan) for the annotation they performed.

References

- Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.
- Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.
- Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE transactions on affective computing*.

⁹Love is a blend of joy and trust; delight is a dyad made of love and surprise; sentimentality is made of trust and sadness.

- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Andrea Bolioli, Federica Salamino, and Veronica Porzionato. 2013. Social media monitoring in real life with blogmeter platform. *ESSEM@ AI* IA*, 1096:156–163.
- Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.
- Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the EVALITA 2018 task on irony detection in italian tweets (IronITA). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.
- June Cotte and Robin Ritchie. 2005. Advertisers’ theories of consumers: Why use negative emotions to sell? *ACR North American Advances*.
- Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.
- Raymond W Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.
- Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (COMPTHELIX)*, pages 397–402. IEEE.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Evgeny Kim and Roman Klinger. 2019. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift fuer Digitale Geisteswissenschaften*, 4.
- Krippendorff Klaus. 1980. Content analysis: An introduction to its methodology.
- Fedric Kujur and Saumya Singh. 2018. Emotions as predictor for consumer engagement in youtube advertisement. *Journal of Advances in Management Research*.
- Kathleen T Lacher. 1989. Hedonic consumption: Music as a product. *ACR North American Advances*.
- Jasy Suet Yan Liew, Howard R Turtle, and Elizabeth D Liddy. 2016. EmoTweet-28: a fine-grained emotion corpus for sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1149–1156.
- Bing Liu. 2012. *Sentiment analysis and opinion mining*, volume 5 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers.
- Adrian Micu, Angela Eliza Micu, Marius Geru, and Radu Constantin Lixandroi. 2017. Analyzing user sentiment in social media: Implications for online marketing strategy. *Psychology & Marketing*, 34(12):1094–1100.
- Richard W Mizerski and J Dennis White. 1986. Understanding and using emotions in advertising. *Journal of Consumer Marketing*.
- Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.
- Saif Mohammad. 2012. # emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.
- Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.
- Anne-Louise Morton, Cheryl Rivers, Stephen Charters, and Wendy Spinks. 2013. Champagne purchasing: the influence of kudos and sentimentality. *Qualitative Market Research: an international journal*.
- Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. European Language Resources Association (ELRA).
- Lucia Passaro, Laura Pollacci, and Alessandro Lenci. 2015. ItEM: A vector space model to bootstrap an Italian emotive lexicon. In *Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 215–220. Academia University Press.
- Duc-Anh Phan, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Multiple emotions detection in conversation transcripts. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 85–94.

- Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.
- Daniel Preotjuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 9–15.
- Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. 2012. EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, volume 12, pages 3806–3813. Citeseer.
- Phakhawat Sarakit, Thanaruk Theeramunkong, Choochart Haruechaiyasak, and Manabu Okumura. 2015. Classifying emotion in thai youtube comments. In *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, pages 1–5. IEEE.
- Julio Savigny and Ayu Purwarianti. 2017. Emotion classification on youtube comments using word embedding. In *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, pages 1–5. IEEE.
- Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2016. Multi-lingual opinion mining on youtube. *Information Processing & Management*, 52(1):46–60.
- Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.
- Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560.
- Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.
- Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249.

Becoming JILDA

Irene Sucameli

Dipartimento di Informatica
Università di Pisa

irene.sucameli@phd.unipi.it

Alessandro Lenci

Dipartimento di Filologia, Letteratura, Linguistica
Università di Pisa

alessandro.lenci@unipi.it

Bernardo Magnini

Fondazione Bruno Kessler
Trento

magnini@fbk.eu

Maria Simi

Dipartimento di Informatica
Università di Pisa

simi@di.unipi.it

Manuela Speranza

Fondazione Bruno Kessler
Trento

manspera@fbk.eu

Abstract

English. The difficulty in finding useful dialogic data to train a conversational agent is an open issue even nowadays, when chatbots and spoken dialogue systems are widely used. For this reason we decided to build JILDA, a novel data collection of chat-based dialogues, produced by Italian native speakers and related to the job-offer domain. JILDA is the first dialogue collection related to this domain for the Italian language. Because of its collection modalities, we believe that JILDA can be a useful resource not only for the Italian research community, but also for the international one.

Italiano. Negli ultimi anni l'utilizzo di chatbot e sistemi dialogici è diventato sempre più comune; tuttavia, il reperimento di dati di apprendimento adeguati per addestrare agenti conversazionali costituisce ancora una questione irrisolta. Per questo motivo abbiamo deciso di produrre JILDA, un nuovo dataset di dialoghi relativi al dominio della ricerca del lavoro e realizzati via chat da parlanti nativi italiani. JILDA costituisce la prima collezione di dialoghi relativi a questo dominio, in lingua italiana. Per gli aspetti metodologici e la modalità di raccolta dei dati, riteniamo che una simile risorsa possa essere utile ed interessante non solo per la comunità di ricerca italiana ma anche per quella internazionale.

1 Introduction

Chatbots and spoken dialogue systems are now widespread; however, there is still a main issue

connected to their development: the availability of training data. Finding useful data to train a system to interact as human-like as possible is not a trivial task. This problem is even more critical for the Italian language, where only few datasets are available. To supplement this deficiency of data, we decided to develop **JILDA** (*Job Interview Labelled Dialogues Assembly*), a new collections of chat-based mixed-initiative, human-human dialogues related to the job offer domain. Our work offers different elements of novelty. First of all, it constitutes, to the best of our knowledge, the first dialogue collection for this domain for the Italian language. Moreover, our dataset was not built using a Wizard of Oz approach, usually adopted in the realization of dialogues. Instead, we used an approach similar to the Map Task one, as we will describe in the next section. This allowed us to obtain more complex, mixed-initiative dialogues.

2 Background

Few dialogic datasets are available for Italian, including the NESPOLE dialogues related to the tourism domain (Mana, 2004), QA datasets related to the movie or the customer care domains (Bentivogli, 2014), and a recent dataset derived from the translation of the English SNIPS (Castellucci, 2019). However, the resources currently available are still limited and, to the best of our knowledge, none of the existing ones is related to the domain of job-offer. For what concerns the English language, although there are more dialogic resources that can be used to train conversational agents (Lowe, 2015; Yu, 2015; El Asri, 2017; Budzianowski, 2018; Li, 2018), as far as we know there are no relevant and freely accessible datasets related to job-matching. Moreover, these

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

datasets usually record simplified conversations, which do not represent the effective complexity that characterises human-human interactions. To fill this gap, we decided to produce a new dialogic dataset for the job domain, for the Italian language. To collect data representative of the linguistic naturalness of native speakers, we had to detect the best approach to fulfil our aim.

The WoZ approach. One of the common approaches used to build full-scale datasets is Wizard of Oz (WoZ) (Kelley, 1984), where a human (the wizard) covers the role of the computer within a simulated human-computer conversation. The other participants in the conversation, however, are not aware that they are talking to a human rather than a conversational system (Rieser, 2008). This method has pros and cons: it may allow to collect conversations written in natural language in a short time (Wen, 2017); however, the dialogues built in this way may not record the noisy conditions experienced in real conversations (e.g. repetitions, errors) and do not show much variation from the syntactic and semantic point of view (Budzianowski, 2018). Due to the limitations of WoZ, we decided to adopt other methods to build our dataset. The first method used in an initial phase of experimentation, was the template-based approach.

The template-based approach. In this solution, it is asked to a volunteer to paraphrase template dialogues using natural language in order to create a simulated dialogue (Shah, 2018). We experienced this modality during an initial experimental phase, in which we used templates for creating task-oriented dialogues. In this first experiment, as previously done by Shah et al. (Shah, 2018), we used Amazon Mechanical Turk¹ and we asked Italian native speakers to cover the role of both the computer and the user, paraphrasing templates of dialogues between a recruiter and a job seeker. We proposed three different templates, with 15-20 recruiter-user interactions each and, to ensure greater lexical variety, we inserted some random variables into the templates (for example, user's skills and the type of job requested). With this experimental set up, we built a first dataset of 220 dialogues. However, despite the attempts to ensure linguistic variety, we noticed that in the MTurk dataset the conversation was strongly guided by

the templates provided and that the dialogues were little diversified from a lexical point of view.

The Map Task approach. To overcome the limits of the WoZ and of the template-based approach, and to produce a set of mixed-initiative dialogues which reflect the naturalness typical of human-human interaction, we decided to organise a new experiment. In this second phase of experimentation, we used as guideline the methodology adopted for the Map Task experiment (Brown, 1984), in which two participants collaborate to achieve a common purpose. For example, Anderson et al. adopted the Map Task to build the HCRC Corpus (Anderson, 1991), a corpus of dialogue recordings and transcriptions. Realized in a similar way, but for the Italian language, there is the CLIPS² corpus, a dataset containing speech recordings.

In Anderson's Map Task, one speaker (the Instruction Giver) has a route marked on the map while the other speaker (the Instruction Follower) has the map without the route and, talking with the Instruction Giver, has to reproduce the route. However, the two maps are not identical and the participants have to discover how they differ.

In our experiment, the two parts involved had to collaborate in a conversation to find the best match between job-offer and candidate profile. The participants covered the role of the *navigator*³, who had a set of possible job offers, and of the *applicant*, who was provided with a job profile to impersonate (a short CV). While in the HCRC Map Task the two parts had to interact in order to figure out the route on the blind map, in this case the two participants had to chat to find the best job-offer match possible for both parts. In the next section, both the framework and the set up of our experiment are described in detail.

3 Experimental setup

To create the JILDA dialogues collection for job-offer, we asked 50 Italian native speakers to simulate a conversation between a "navigator" and an applicant. At the end of the experiment, all the volunteers received an economical reward for their participation. We randomly assigned to 25 volun-

²Available here: <http://www.clips.unina.it/it/corpus.jsp>

³The navigator plays a role similar to the recruiter's one, who is in charge of reviewing candidate's skills and past experiences in order to find a suitable job.

¹Available here: <https://www.mturk.com/>

teers the role of navigator, providing 5 job offers each. The other 25 volunteers had to pretend to be applicants and describe themselves on the basis of the information contained in a curriculum we provided. The navigators' goal was to help applicants to find a job offer (among the offers available) best suited to their curriculum and interests by asking questions. Applicants, on the other side, had to interact with the navigator describing the skills and competencies included in their curricula.

Similarly to the Map Task framework, the two parties had to collaborate in order to reach their goal and were engaged in creating a mixed initiative spontaneous dialogue without a strict guidance. Navigators and applicants were free to lead the conversation as they preferred; in fact, we did not use any dialogue template (although we provided some examples) and both applicants and navigators were allowed to ask questions to their interlocutor, in order to reach the best possible match between applicant's needs and the job offers available to the navigator. The only compulsory requirements we imposed to participants was to converse only about topics related to the experiment. In addition to this, we provided as guideline an indicative length of 15/20 (overall) utterances per dialogue.

Both navigators and applicants were not allowed to interact with the same interlocutor twice. Each navigator interacted with 21 different applicants and, in a similar way, each applicant had to interact with 21 navigators. With this strategy we wanted not only to obtain dialogues as linguistically diversified as possible, but also to ensure that navigators with different offers interacted with applicants with different curricula and needs.

To make the navigator interact with the applicant, we used the Slack platform⁴, which allowed the volunteers to interact with each other in an easy way, maintaining anonymity through the use of nicknames. Moreover, it allowed us to monitor multiple conversations at the same time and to easily download the dialogues' output in a json format suitable for the future annotations. Neither the applicants nor the navigators knew with whom they had to chat.

We asked the volunteers to realise 21 chat-based dialogues distributed in five days, so they had to produce 4 or 5 dialogues per day.

⁴Available at <https://slack.com/intl/en-it/>

4 Results and Discussion

At the end of the experiment, we collected 525 chat-based, mixed initiative dialogues⁵. In order to have a first evaluation of the data produced, we asked our volunteers to assess the quality of the dialogues. More specifically, we asked to evaluate the degree of naturalness, the linguistic variety of the dialogues (Table 1), and the difficulties detected in the experiment (Table 2). Among the 50 participants, 29 completed the evaluation questionnaire. The results obtained are reported below.

Rating Scale	Realism	Linguistic variety
1 (very low)	0%	0%
2	7%	14%
3	14%	55%
4	62%	21%
5 (very high)	17%	10%

Table 1: *Evaluation of the degree of realism and linguistic variety of JILDA dialogues.*

Rating scale	Difficulty in understanding	Difficulty in the description
Very low	0%	0%
Medium	17%	48%
Very high	83%	52%

Table 2: *Evaluation on the degree of difficulty in understanding the interlocutor's requests and in describing the job offers/CV available.*

The volunteers' evaluation is in line with what can be observed directly from the dialogues. In fact, from a preliminary analysis, the dialogues produced exhibit a good linguistic variety and capture complex phenomena of the Italian language, such as co-reference. Since they are task oriented dialogues, the data follow a certain pattern of questions/answers but, within this common structure, the navigator-applicant interaction varies in an extremely interesting way. For instance, we noticed the presence of asynchronous messages with respect to the context, as shown in the example reported in Appendix A. This is due to the fact that users have the tendency to type fast while they are chatting, and this may lead to overlapping messages, where the answer to a question is not immediate but comes in a later turn. Furthermore,

⁵Both JILDA and MTurk datasets are available here: <http://dialogo.di.unipi.it/jilda/>

applicants do not passively answer to navigators but they often take the initiative, formulating questions and proactively giving unsolicited information. Comparing JILDA’s dialogues with MTurk’s ones, it is clear that JILDA’s dialogues are more complex and semantically diversified.

	MTurk	JILDA
# dialogues	220	525
avg turns/dialogue	8	17
# tokens	45972	217132
# sentences	5201	20644
# utterances	3380	14509
# types	1975	6519
# lemmas	1605	4913
type/token ratio	0.043	0.072*
lemma/token ratio	0.035	0.056*
avg length sentences	9.24	10.52
avg length utterances	13.58	14.94
# proactive/intent	1.97%	17.30%
# proactive/sentences	1.46%	12.70 %

Table 3: *Comparison between MTurk’s and JILDA’s dialogues. Values marked with an asterisk are computed considering the average value of three JILDA’s subsets, each of which includes the same number of tokens as MTurk*

A first analysis, for which we also used Profiling-UD (Brunato, 2020) and UDPipe (Straka, 2017), highlights differences of the new dataset with respect to the previous one ⁶ such as:

- **lexical variability.** As shown in Tab.3, JILDA has a greater lexical variability, which is extremely useful if the dataset is used to train new models. In fact, considering the whole dataset, JILDA has more tokens and types. Even more importantly, by selecting subsets of JILDA with the same number of tokens as MTurk, it is possible to verify that, on the average, JILDA’s lexical richness is higher (see the lemma and type/token ratio).
- **syntactic complexity.** With respect to the MTurk dataset, JILDA includes more subordinates and longer chains of dependencies, which is an indication of more complex sentences. In fact, the analysis conducted with Profiling-UD (Brunato, 2020) shows

⁶It is worth to highlight that the differences between the two resources are primarily related to the methods used for data collection and not to the platforms used.

for JILDA a higher percentage of subordinate propositions (51.46% against 39.87% in MTurk) and longer chains of embedded subordinate clauses (18.35% of the chains are long 2 or more in JILDA, 12.48% in MTurk).

- **dialogue naturalness.** The naturalness of JILDA’s dialogues partially emerged in the first evaluation conducted with the participants in the experiment (Table 1-2). In addition to this, Table 3 shows that JILDA contains a high number of proactiveness phenomena, which are significant in highlighting the complexity of a dialogue and its collaborative nature. In particular, JILDA contains a higher number of proactive intents, both in terms of percentage over the total number of intents and over the number of sentences. ⁷ This shows that our volunteers did not merely answer their interlocutor by providing the strictly required information, but rather on their own initiative provided additional information, which made the dialogues more natural and complex.

The annotation of the dialogues is now in progress in order to offer to the scientific community not only a new set of dialogues for the Italian language but also, and above all, a richly annotated dataset. The annotation will take as a basis the notation of Multiwoz, which is becoming a standard in dialogue datasets (Budzianowski, 2018). However, although in Multiwoz only user’s turns are annotated, we decided to annotate both applicant’s and navigator’s utterances, since we noticed that both utterances convey important and useful information. The preliminary analysis of the data presented here will be deepened once the annotation is complete. To support the annotation work of the JILDA dataset, we modified an open source dialogue annotation tool, LIDA, in collaboration with its developers (Collins, 2019). Specifically, we extended this tool to 1) allow support for multiple annotators working at the same project, 2) manage multiple annotation styles and metadata information, 3) manage different collections of dialogues and 4) simplify the annotation interface, improving the user experience. Both the new release of the LIDA Multi-user annotation tool and the JILDA annotated dataset will be made available to the scientific community.

⁷Proactive intents were explicitly annotated for this count.

5 Conclusion

In this paper we presented JILDA, a novel dataset of chat-based, mixed-initiative dialogues built for the Italian language and related to the job-offer domain. This new resource has been built adopting an experimental approach based on the Map Task experiment. This has allowed us to collect mixed-initiative data which represent effectively the naturalness which is typical in the human-human interaction. The JILDA dataset, which includes 525 dialogues, is in the process of being completely annotated with dialogue acts and entities related to this specific domain. For the annotation of those dialogues we are using our own extension of LIDA. The annotated dialogues will then be used to train a conversational agent. Thanks to this new resource, our goal is to allow an agent chat with the user in a natural and human-like way.

Acknowledgments

This work has been endorsed by AILC (Italian Association for Computational Linguistics). We thank Carla Congiu, Clara Casandra and Davide Cucurnia, students of Digital Humanities at the University of Pisa, for annotation work on JILDA and for contributing to the development of the annotation tool.

References

- Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R., 1991. The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351-366.
- Bentivogli, L., Magnini, B., 2014. An Italian Dataset of Textual Entailment Graphs for Text Exploration of Customer Interactions. In *Proceedings of the first Italian Computational Linguistics Conference*.
- Brown, G., Anderson, A., Yule, G., Shillcock, R., 1984. *Teaching talk: Strategies for production and assessment*. Cambridge University Press.
- Brunato D., Cimino A., Dell’Orletta F., Montemagni S., Venturi G., 2020. Profiling-UD: a Tool for Linguistic Profiling of Texts”. In *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*, pp. 11-16 May, 2020, Marseille, France.
- Budzianowski, P. Tsung-Hsien, W., Bo-Hsiang, T. Casanueva, I., Ultes, S., Ramadan, O., Gašić, M., 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5016-5026.
- Castellucci, G., Bellomaria, V. and Favalli, A., Romagnoli, R., 2019, Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model. In ArXiv abs/1907.02884.
- Collins, E., Rozanov, N., Zhang, B. 2019 LIDA: Lightweight Interactive Dialogue Annotator. In *Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations)*, pp. 121–126.
- Dell’Orletta, F., Montemagni, S., Venturi, G., 2011 READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, Association for Computational Linguistics, pp. 73-83.
- El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., Suleman, K. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. In arXiv:1704.00057
- Kelley, J.F. 1984. An iterative design methodology for user-friendly natural language office information applications. In *ACM Transactions on Information Systems (TOIS)*, 2(1), pp. 26-41.
- Li, R., Kahou, S.E., Schulz, H., Michalski, V., Charlin, L., Pal, C. 2018. Towards Deep Conversational Recommendations, In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, pp. 9748-9758.
- Lowe, R., Pow, N., Serban, I. and Pineau J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference*, pp. 285-294.
- Mana, N., Cattoni, R., Pianta, E., Rossi, F., Pianesi, F., and Burger, S. 2004. The Italian NESPOLE! Corpus: a Multilingual Database with Interlingua Annotation in Tourism and Medical Domains. In *Proceedings of 4th International Conference LREC*.
- Rieser, V., Lemon, O., 2008. Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation. In *Proceeding of ACL-08:HLT*, pp. 638-646.
- Shah, P., Hakkani-Tür, D., Liu, B., Tür, G., 2018. Bootstrapping a Neural Conversational Agent with Dialogue Self-Play, Crowdsourcing and On-Line Reinforcement Learning. In *Proceeding NAACL-HLT 2018*, pp. 41-45.
- Straka, M. and Straková, J., 2017 Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UD-Pipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88-99 Vancouver, Canada, August 2017.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L.M., Su, P.-H., Ultes, S., Young, S., 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1, pp. 438–449.

Yu, Z., Papangelis, A., Rudnicky, A.I., 2015. Tick-Tock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness. *AAAI Spring Symposium*.

Appendix A

Example of asynchronous message in JILDA

Navigator: Cercano persone che si occupino sia di gestire la comunicazione pubblicitaria del cliente attraverso il web, che di interagire direttamente con la clientela.

Applicant: Quanto tempo dura il periodo di formazione?

Navigator: Questo significa che abilità di comunicazione sono essenziali in questo lavoro

Applicant:

Navigator: L'annuncio non fornisce informazioni circa la durata del contratto, mi dispiace

Appendix B

Example of dialogue from Mturk

sys: Salve e benvenuto alla Recruiter Top, io sono Tony.

usr: Buongiorno Tony, mi chiamo Giorgio e sono alla ricerca di un lavoro come traduttore.

sys: Bene, mi dica qualcosa in più su di lei; attualmente lavora o studia? e quali sono le sue competenze?

usr: Mi sono appena laureato in lingue e letterature straniere, nello specifico con conoscenza di inglese, spagnolo e francese

sys: E per quanto riguarda esperienze lavorative?

usr: Ho lavorato 2 anni in una casa editrice

sys: in che senso ha lavorato per 2 anni in una casa editrice?

usr: Sono stato vicedirettore per una casa editrice locale.

sys: Ok, e per quello che riguarda le lingue straniere? essendo laureato in lingue...

usr: bhe sì..ovviamente ho una buona conoscenza di francese, inglese e spagnolo

sys: Che contratto e tipologia di azienda sta cercando o valterebbe?

usr: Propenderei per un tempo determinato in una azienda all'estero

sys: al momento si cerca per la sede di Gais figura professionale che si occupi della corrispondenza telefonica e scritta con i nostri clienti spagnoli e portoghesi e di organizzare eventi di marketing, potrebbe essere interessato?

usr: Sì, mi dia i dettagli e lo valuterò. Grazie e arrivederci

sys: Arrivederci e buona fortuna.

Example of dialogue from JILDA

sys: Ciao, sono il tuo Navigator di oggi, mi chiamo Mattia. Posso aiutarti in qualche modo?

usr: Buongiorno Mattia, mi chiamo Valentina e sto cercando un lavoro a tempo determinato.

sys: Ciao Valentina, puoi dirmi qualcosa in più sugli studi che hai fatto?

usr: Certamente! Mi sono laureata tre anni fa in Lingue e Letterature straniere.

sys: Ottimo, hai già avuto esperienza lavorativa in passato o sarebbe il tuo primo lavoro?

usr: Ho già avuto un'esperienza lavorativa, perché per due anni ho lavorato come guida museale.

sys: Ti è mai capitato di lavorare a progetti con bambini, durante questi due anni?

usr: Quando lavoravo per il museo non ho mai affrontato dei progetti specifici riguardanti i bambini. Ho però fatto da guida a delle scolaresche.

sys: Ho qui un annuncio riguardo la possibilità di fare assistenza scolastica a minori con disabilità, dalle scuole d'infanzia alle superiori. Pensi che ti piacerebbe provare qualcosa del genere?

usr: Sarebbe un'esperienza interessante, ma non credo di avere le competenze necessarie. Preferirei rimanere nel campo dei musei o, in generale, in quello dei luoghi turistici.

sys: Al momento non ho annunci per posti disponibili in campo turistico o museale, mi dispiace. Data la tua laurea in Lingue però, vorrei proporti un annuncio di CHANEL Cordination S.r.l., sono alla ricerca di una stagista da affinare alla Responsabile Qualità Prodotto referente per l'Italia.

usr:

sys: dovresti occuparti principalmente di Monitoraggio del database dei prodotti delle collezioni. Gestione dei contatti con i fornitori locali ed esteri. Archiviazione e consultazione dei Test di laboratorio e supporto della responsabile nella preparazione di presentazioni in PPT e nelle traduzioni della reportistica nelle lingue in inglese e francese

usr: Mi interesserebbe molto. Dove si trova l'azienda?

sys: La sede dell'azienda è a Milano, quindi probabilmente dovrai spostarti lì se non abiti già in zona,

usr: Non sarebbe un problema spostarmi. Il lavoro è a tempo pieno o a tempo parziale?

sys: Non è specificato nell'annuncio, so solo che si tratta di un tirocinio/stage. Probabilmente è un cosa da discutere in fase di colloquio direttamente con loro

usr: Ok grazie.

sys: Puoi contattare direttamente l'azienda a questo indirizzo e-mail info@azienda.com

usr: Perfetto, grazie mille! :)

sys: Figurati, buona fortuna per il lavoro!

usr: Grazie, buona giornata! :)

How “BERTology” Changed the State-of-the-Art also for Italian NLP

Fabio Tamburini

FICLIT - University of Bologna, Italy

fabio.tamburini@unibo.it

Abstract

The use of contextualised word embeddings allowed for a relevant performance increase for almost all Natural Language Processing (NLP) applications. Recently some new models especially developed for Italian became available to scholars. This work aims at evaluating the impact of these models in enhancing application performance for Italian establishing the new state-of-the-art for some fundamental NLP tasks.

1 Introduction

The introduction of contextualised word embeddings, starting with ELMo (Peters et al., 2018) and in particular with BERT (Devlin et al., 2019) and the subsequent BERT-inspired transformer models (Liu et al., 2019; Martin et al., 2020; Sanh et al., 2019), marked a strong revolution in Natural Language Processing, boosting the performance of almost all applications and especially those based on statistical analysis and Deep Neural Networks (DNN).

A recent study (He and Choi, 2019) tried to determine the new baselines for several NLP tasks for English fixing the new state-of-the-art for the examined tasks. This work aims at doing a similar process also for Italian. We considered a number of relevant tasks applying state-of-the-art neural models available to the community and fed them with all the contextualised word embeddings specifically developed for Italian.

2 Italian “BERTology”

The availability of various powerful computational solutions for the community allowed for

the development of some BERT-derived models trained specifically on big Italian corpora of various textual types. All these models have been taken into account for our evaluation. In particular we considered those models that, at the time of writing, are the only one available for Italian:

- Multilingual BERT¹: with the first BERT release, Google developed also a multilingual model (‘bert-base-multilingual-cased’ – *bertMC*) that can be applied also for processing Italian texts.
- ALBERTo²: last year, a research group from the University of Bari developed a brand new model for Italian especially devoted to Twitter texts and social media (‘m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0’ – *alUC*) trained by using 200 millions tweets from 2012 to 2015 (Polignano et al., 2019). Only the uncased model is available to the community. Due to the specific training of *alUC*, it requires a particular pre-processing step for replacing hashtags, urls, etc. that alter the official tokenisation, rendering it not really applicable to word-based classification tasks in general texts; thus, it will be used only for working on twitter or social media data. In any case we tested it in all considered tasks and, whenever results were reasonable, we reported them.
- GILBERTo³: it is a rather new CamemBERT Italian model (‘idb-ita/gilberto-uncased-from-camembert’ – *giUC*) trained by using the huge Italian Web corpus section of the OSCAR (Ortiz Suárez et al., 2019) Web-corpus project consisting of more than 11

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://github.com/google-research/bert>

²<https://github.com/marcopoli/ALBERTo-it>

³<https://github.com/idb-ita/GILBERTo>

billions of tokens. Also for GiBERTo it is available only the uncased model.

- UmBERTo⁴: the more recent model developed explicitly for Italian, as far as we know, is UmBERTo ('Musixmatch/umberto-commoncrawl-cased-v1' – *umC*). As well as GiBERTo, it has been trained by using OSCAR, but the produced model, differently from GiBERTo, is cased.

3 Evaluation Tasks

Following the work of He and Choi (2019), we selected some basic tasks both for word and sentence/text classification. We mainly concentrated our efforts on tasks for which evaluation procedures were well established in the Italian community and reliable evaluation benchmark were available. We choose (a) two very basic word-classification tasks, namely part-of-speech (PoS) tagging and Named Entity Recognition (NER), (b) the dependency parsing task and (c) two very important tasks for social-media text classification, namely Sentiment Analysis (Subjectivity/Polarity/Irony classification) and Hate Speech Detection (HSD).

We mainly relied on some benchmark proposed in one of the past EVALITA evaluation challenges⁵ or the Universal Dependencies (UD) project⁶.

After the influential paper from (Reimers and Gurevych, 2017) it is clear to the community that reporting a single score for each DNN training session could be heavily affected by the system initialisation point and we should instead report the mean and standard deviation of various runs with the same setting in order to get a more accurate picture of the real systems performance and make more reliable comparisons between them. Thus any new result proposed in this paper is presented as the mean and standard deviation of at least 5 runs.

With regard to the dataset splitting, if a specific dataset was already split in training/validation/test set, we adopted this subdivision, while, if the dataset was split only in development and test set, we split it and used the training/validation sets for training and tuning the stopping epoch and, once fixed that parameter, we retrained the system on

the entire development set maintaining the same epoch for the early stopping.

3.1 Part-of-Speech Tagging

The first task we worked on is the part-of-speech tagging. This is a very basic task in NLP and a lot of applications rely on precise PoS-tag assignments. There are various data sets available for this task taken from one of the EVALITA 2007 tasks (Tamburini, 2007) and from the UD annotated corpora.

System	EVALITA 2007
(Tamburini, 2016)	98.18
Fine-Tuning _{giUC}	98.75 \pm 0.04
Fine-Tuning _{bertMC}	98.80 \pm 0.05
Fine-Tuning _{umC}	99.10 \pm 0.04

Table 1: PoS-tagging Accuracy for the EVALITA 2007 benchmark.

System	UD-ISDT v2.5	
	UPOS	XPOS
Fine-Tuning _{giUC}	98.72 \pm 0.03	98.65 \pm 0.02
Fine-Tuning _{bertMC}	98.73 \pm 0.05	98.69 \pm 0.05
Fine-Tuning _{umC}	98.78 \pm 0.08	98.73 \pm 0.02

Table 2: PoS-tagging Accuracy for UD-ISDT v2.5 corpus both considering UPOS and XPOS.

System	UD-PoSTW v2.5	
	UPOS	XPOS
(Cimino and Dell'Orletta, 2016a)	93.19	-
(Basile et al., 2017)	93.34	-
Fine-Tuning _{giUC}	94.77 \pm 0.07	94.57 \pm 0.05
Fine-Tuning _{bertMC}	96.37 \pm 0.09	96.18 \pm 0.06
Fine-Tuning _{umC}	97.29 \pm 0.33	97.27 \pm 0.04

Table 3: PoS-tagging Accuracy for UD-PoSTWITA v2.5. N.B.: the baselines from the literature refer to the previous PoSTWITA version used in EVALITA 2016 campaign.

The best results for the EVALITA 2007 data set has been obtained by (Tamburini, 2016) using a BiLSTM-CRF system based on word2vec word embeddings enriched with morphological information. For UD corpora we considered the ISDT corpus v2.5 and PoSTWITA: there are no evaluation data in literature for the ISDT corpus while for PoSTWITA the best results were obtained by

⁴<https://github.com/musixmatchresearch/umberto>

⁵<http://www.evalita.it>

⁶<https://universaldependencies.org>

(Basile et al., 2017) using a BiLSTM-CRF system and by the best system at EVALITA 2016 (Cimino and Dell’Orletta, 2016a).

The PoS-tagging system used for our experiments is very simple and consist of a slight modification to the fine tuning script ‘run_ner.py’ available with the version 2.7.0 of the Huggingface/Transformers package⁷. We did not employ any hyperparameter tuning, the validation set has been used only for determining the stopping criterion.

Tables 1, 2 and 3 show the results obtained by fine tuning the considered BERT-derived models for this task. A very relevant increase in performance w.r.t. the literature is evident by looking at the results and UmBERTo is consistently the best system.

3.1.1 PoS-tagging on Speech Data

We participated to the EVALITA 2020 KIPOS challenge (Bosco et al., 2020) for evaluating PoS-taggers on speech data by using exactly the same tagger. In this case, we did not make any parameter tuning: we used the basic parameters and stopped the training phase after 10 epochs. After the challenge, we evaluated all the BERT-derived models in order to propose a complete overview of the available resources.

Tables 4 show the results obtained by fine tuning all the considered BERT-derived models for the Main Task. A very relevant increase in performance w.r.t. the other participants is evident looking at the results and UmBERTo is again the best system.

We did not participate at the official challenge for the two subtasks, but we included the results of our best system also for these tasks. Table 5 shows the results compared with the other two participating systems. Again, the simple fine tuning of a BERT-derived model, namely UnBERTo, exhibits the best performance on Sub-task B. The scarcity of data could probably affect the results on Sub-task A.

3.2 Named Entity Recognition

The second task we considered is Named Entity Recognition. For system evaluation we relied on the nice evaluation benchmark used in the EVALITA 2009 campaign (Bartalesi Lenzi et al., 2009). The best results gathered from literature are due to (Basile et al., 2017) that used a

⁷<https://huggingface.co/transformers/>

System	Main Task Accuracy		
	Form.	Inform.	Both
(Izzi and Ferilli, 2020)	81.58	79.37	80.43
(Proisl and Lapesa, 2020)	87.56	88.24	87.91
Fine-Tuning _{bertMC}	91.67	88.05	89.79
Fine-Tuning _{alUC}	90.02	89.82	89.92
Fine-Tuning _{giUC}	92.96	89.92	91.38
Fine-Tuning _{umC}	93.49	91.13	92.26

Table 4: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the Main Task.

System	Sub-Task A Accuracy		
	Form.	Inform.	Both
(Izzi and Ferilli, 2020)	78.73	75.79	77.20
Fine-Tuning _{umC}	86.47	83.16	84.75
(Proisl and Lapesa, 2020)	87.37	87.58	87.48
System	Sub-Task B Accuracy		
	Form.	Inform.	Both
(Izzi and Ferilli, 2020)	77.11	77.50	77.31
(Proisl and Lapesa, 2020)	87.81	88.10	87.96
Fine-Tuning _{umC}	89.74	89.52	89.63

Table 5: PoS-tagging Accuracy for the EVALITA KIPOS 2020 benchmark for the two Sub-Tasks A and B.

BiLSTM-CRF system and to the best system at the EVALITA 2009 campaign (Zanoli et al., 2009).

For this task we used exactly the same script of the previous task, being both tasks simple word-classification tasks, and did not apply any hyperparameter tuning at all, fixing a priori the number of epoch to 10.

Table 6 outlines the obtained results. Again a simple fine tuning of BERT-derived models is enough powerful to guarantee relevant increases of performance with respect to the previous literature and, again, UmBERTo resulted the model producing the best performance.

System	Macro F1
(Zanoli et al., 2009)	82.00
(Basile et al., 2017)	82.34
Fine-Tuning _{giUC}	82.37 \pm 0.31
Fine-Tuning _{bertMC}	85.07 \pm 0.29
Fine-Tuning _{umC}	87.66\pm0.44

Table 6: Macro-averaged F1-score for the various systems when evaluated with the EVALITA 2009 NER benchmark.

3.3 Parsing Universal Dependencies

Parsing is one of the most important tasks in NLP and the recent advances due to DNN and contextualised distributed representations allowed for large performance improvements.

Universal Dependencies project is the reference repository for standardised treebanks in various languages, thus it seemed natural to gather evaluation benchmarks from that project. As for PoS-tagging, we used two treebanks from UD v2.5, namely ISDT and PoSTWITA.

The recent work from Antonelli and Tamburini (2019) examined all the DNN parsers available at the time re-training them on some Italian dataset. In particular they showed that the neural parser from Dozat and Manning (2017) (version 1.0) was the parser exhibiting the best performance on UD-ISDT v2.1. Giving that experience, we included in our new experiments the last version (v3.0) of this parser⁸ considering it as a strong baseline for this task. The word embeddings we used for these experiments were the same used in (Antonelli and Tamburini, 2019) and are computed using the ItWaC corpus (Baroni et al., 2009) and word2vec (Mikolov et al., 2013a,b).

Very recently, a new work from Vacareanu et al. (2020) showed that we can efficiently compute dependency parsing structures by treating this task as a double fine tuning task over a BERT-derived model, the first for determining the attachments and the second the edge labels, getting state-of-the-art performance. Actually, the fine-tuning DNN is more complex than in the previous tasks, consisting of a bidirectional LSTM followed by some dense layers.

We applied their method and code (PaT) for our parsing experiments using the greedy cycle removal option. We changed text case depending on the BERT-derived model case used in a specific experiment. Tables 7 and 8 show the results for all the parsing experiments.

Considering the best results obtained by the Dozat and Manning (2017) parser and those presented in (Antonelli and Tamburini, 2019), we observe a relevant increase in performance due mainly to GiBERTo and UmBERTo.

3.4 Sentiment Analysis

Three main text-classification tasks are comprised in the ‘Sentiment Analysis’ umbrella: Subjectiv-

⁸<https://github.com/tDOZAT/Parser-v3>

System	UD-ISDT v2.5	
	UAS	LAS
(Antonelli and Tamburini, 2019)	94.00	92.48
PaT _{bertMC}	94.12 \pm 0.26	91.74 \pm 0.23
(Dozat and Manning, 2018)	94.53 \pm 0.14	93.35 \pm 0.18
PaT _{umC}	95.32 \pm 0.14	93.39 \pm 0.26
PaT _{giUC}	95.52 \pm 0.18	93.59 \pm 0.28

Table 7: Parsing Un/Labeled Attachment Score (UAS/LAS) for UD-ISDT v2.5.

System	UD-PoSTW v2.5	
	UAS	LAS
PaT _{bertMC}	87.97 \pm 0.20	82.03 \pm 0.24
(Dozat and Manning, 2018)	88.04 \pm 0.13	84.08 \pm 0.10
PaT _{alUC}	88.19 \pm 0.32	82.66 \pm 0.38
PaT _{umC}	89.16 \pm 0.17	83.25 \pm 0.23
PaT _{giUC}	89.29 \pm 0.27	83.66 \pm 0.22

Table 8: Parsing Un/Labeled Attachment Score (UAS/LAS) for UD-PoSTWITA v2.5.

ity, Polarity and Irony detection. Thanks to the EVALITA SENTIPOLC 2016 evaluation we could rely on a complete dataset annotated with respect to all the three tasks.

Given the specific nature of dataset texts, namely tweet texts, we adopted the particular pre-processing procedure introduced by AlBERTo and all the other parameters were kept as in (Polignano et al., 2019) for comparability; the only difference regards the training batch size that was 512 on TPU in the original paper and we had to use gradient accumulation on GPU (batch size = 32 and accumulation steps = 16) to avoid memory problems. Given the small size of the dataset and the high variability of the various results, for these tasks we decided to make 10 runs instead of 5.

System	Macro F1
TensorFlow+TPU _{alUC}	72.23*
Fine-Tuning _{bertMC}	72.92 \pm 0.86
(Castellucci et al., 2016)	74.44
Fine-Tuning _{alUC}	75.83 \pm 0.63
Fine-Tuning _{umC}	77.14 \pm 0.78
Fine-Tuning _{giUC}	77.58 \pm 1.20
(Polignano et al., 2019) (alUC)	79.06 *

Table 9: Subjectivity detection macro F1-score for EVALITA SENTIPOLC 2016. * results that we were not able to reproduce using the same code.

System	Macro F1
Fine-Tuning _{bertMC} (Cimino and Dell’Orletta, 2016b)	65.38±1.65 66.38
TensorFlow+TPU _{alUC} (Polignano et al., 2019) (alUC)	71.59* 72.23*
Fine-Tuning _{alUC}	72.60±1.38
Fine-Tuning _{umC}	72.74±0.88
Fine-Tuning _{giUC}	74.75±0.94

Table 10: Polarity detection macro F1-score over 4 classes for EVALITA SENTIPOLC 2016. * results that we were not able to reproduce using the same code.

System	Macro F1
Fine-Tuning _{bertMC} (Di Rosa and Durante, 2016)	52.17±1.55 54.12
Fine-Tuning _{umC}	55.65±3.09
Fine-Tuning _{alUC}	56.80±1.92
TensorFlow+TPU _{alUC}	57.21*
Fine-Tuning _{giUC} (Polignano et al., 2019) (alUC)	60.60±1.45 60.90*

Table 11: Irony detection Macro F1-score for EVALITA SENTIPOLC 2016 dataset. * results that we were not able to reproduce using the same code.

We slightly modified the script ‘run_glue.py’ from the version 2.7.0 of the Huggingface/Transformers package considering the three tasks as a BERT-derived model fine-tuning for text classification tasks respectively with 2, 4 and 2 classes.

Tables 9, 10 and 11 present the obtained results. We have to say that we had a lot of problems in reproducing the results in Polignano et al. (2019), both by using our script and also by using the original TPU-based script on Google Colab. In the cited tables, you can find the original results and the ones produced by us using the same script and setting marked by an asterisk (TensorFlow+TPU_{alUC}).

3.5 Hate Speech Detection

Hate Speech on social media has become a relevant problem in recent years and the automatic detection of such messages got a great importance in NLP.

Thanks to the dataset produced by Bosco et al. (2018) we had the possibility to test the same text

System	Macro F1	
	FB	TW
Fine-Tuning _{bertMC} (Cimino et al., 2018)	77.62±0.46 82.88	76.07±0.78 79.93
Fine-Tuning _{umC}	83.55±0.40	80.28±0.55
Fine-Tuning _{alUC}	84.23±0.37	79.00±0.84
Fine-Tuning _{giUC}	84.36±0.69	80.86±0.46

Table 12: Macro F1-score for the HaSpeDe EVALITA 2018 Facebook (FB) and Twitter (TW) datasets.

classification procedures we used for Sentiment Analysis also for this task both on Facebook and Twitter data. Table 12 shows the results we obtained comparing them with the best system at the EVALITA 2018 HaSpeDe campaign (Cimino et al., 2018). GilBERTo exhibit the best performance on both subtasks.

4 Discussion and Conclusions

The starting idea of this work was to derive the new state-of-the-art for some NLP tasks for Italian after the ‘BERT-revolution’ thanks to the recent availability of Italian BERT-derived models. Looking at the results presented in previous sections for some very important tasks, we can certainly conclude that BERT-derived models, specifically trained on Italian texts, allow for a large increase in performance also for some important Italian NLP tasks. On the contrary, the multilingual BERT model developed by Google was not able to produce good results and should not be used when are available specific models for the studied language.

A side, and sad, consideration that emerges from this study regards the complexity of the models. All the DNN models used in this work for the various tasks involved very simple fine-tuning processes of some BERT-derived model. Machine learning and Deep learning changed completely the approaches to NLP solutions, but never before we were in a situation in which a single methodological approach can solve different NLP problems always establishing the state-of-the-art for that problem. And we did not apply any parameter tuning at all! The only optimisation regards the early stopping definition on validation set. By tuning all the hyperparameters, it is reasonable we can further increase the overall performance.

For the future, it would be interesting to eval-

uate end-to-end systems, for example for solving PoS-tagging + Parsing and PoS-tagging + NER by using the BERT-derived model fine tuning code and PaT for both end-to-end tasks.

A lot of scholars are working in studying new transformer-based models or training the most promising ones on different languages; there are brand new Italian models that were made available very recently not yet included into our evaluations like the one produced by Stefan Schweter at CIS, LMU Munich⁹; it would be interesting to insert them into our tests.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

- O. Antonelli and F. Tamburini. 2019. State-of-the-art Italian dependency parsers based on neural and ensemble systems. *Italian Journal of Computational Linguistics*, 5(1):33–55.
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- V. Bartalesi Lenzi, M. Speranza, and R. Sprugnoli. 2009. EVALITA 2009 The Entity Recognition Task. In *Proceedings of the EVALITA 2009 Workshop*, Reggio Emilia, Italy.
- P. Basile, G. Semeraro, and P. Cassotti. 2017. Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 18–23, Roma, Italy.
- C. Bosco, S. Ballarè, M. Cerruti, E. Goria, and C. Mauri. 2020. KIPoS@EVALITA2020: Overview of the Task on KIParla Part of Speech tagging. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.
- C. Bosco, F. Dell’Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *In Proc. of the EVALITA 2018 Workshop*, Torino, Italy.
- G. Castellucci, D. Croce, and R. Basili. 2016. Context-aware Convolutional Neural Networks for Twitter Sentiment Analysis in Italian. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.
- A. Cimino, L. De Mattei, and F. Dell’Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In *In Proc. of the EVALITA 2018 Workshop*, Torino, Italy.
- A. Cimino and F. Dell’Orletta. 2016a. Building the state-of-the-art in POS tagging of Italian Tweets. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.
- A. Cimino and F. Dell’Orletta. 2016b. Tandem LSTM-SVM Approach for Sentiment Analysis. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.
- E. Di Rosa and A. Durante. 2016. Tweet2Check evaluation at Evalita Sentipolc 2016. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italy.
- T. Dozat and C.D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the 2017 International Conference on Learning Representations*.
- T. Dozat and C.D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 484–490, Melbourne, Australia.
- H. He and J.D. Choi. 2019. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. In *The Thirty-Third International Flairs*

⁹<https://github.com/stefan-it/fine-tuned-berts-seq>

- Conference, AAAI Publications, pages 228–233.
- G.L. Izzi and S. Ferilli. 2020. A hybrid approach for part-of-speech tagging. In *Proceedings of the Seventh International Workshop EVALITA 2020*.
- Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- L. Martin, B. Muller, P.J. Ortiz Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*.
- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS’13*, pages 3111–3119, USA. Curran Associates Inc.
- P.J. Ortiz Suárez, B. Sagot, and L. Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom.
- M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT 2018*, pages 2227–2237, New Orleans, Louisiana.
- M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. 2019. ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy.
- T. Proisl and G. Lapesa. 2020. Klumsy: Experiments on part-of-speech tagging of spoken italian. In *Proceedings of the Seventh International Workshop EVALITA 2020*.
- N. Reimers and I. Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. ACL.
- V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.
- F. Tamburini. 2007. EVALITA 2007: the Part-of-Speech Tagging Task. *Intelligenza Artificiale*, IV(2):4–7.
- F. Tamburini. 2016. (Better than) State-of-the-Art PoS-tagging for Italian Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 280–284, Napoli, Italy.
- R. Vacareanu, G.C. Gouveia Barbosa, M.A. Valenzuela-Escárcega, and M. Surdeanu. 2020. Parsing as tagging. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5225–5231, Marseille, France. ELRA.
- R. Zanolì, E. Pianta, and C. Giuliano. 2009. Named entity recognition through redundancy driven classifiers. In *Proceedings of the Workshop EVALITA 2009*, Reggio Emilia, Italy.

Valutazione umana di DeepL a livello di frase per le traduzioni di testi specialistici dall'inglese verso l'italiano

Sirio Papa - Mirko Tavosanis

Dipartimento di Filologia, letteratura e linguistica

Università di Pisa

Via Santa Maria 36 – 56126 Pisa PI

s.papa4@studenti.unipi.it -

mirko.tavosanis@unipi.it

Riassunto¹

Il contributo presenta una valutazione delle prestazioni di DeepL nella traduzione di testi specialistici dall'inglese all'italiano. La valutazione è stata condotta a livello di frase, su un campione di 108 frasi tratte da testi relativi ad ambiente, energia, biomedicina e discipline del farmaco, e le traduzioni prodotte sono state valutate da traduttori in formazione dotati di competenze disciplinari. La traduzione di DeepL ha ottenuto una valutazione statisticamente pari a quella della traduzione umana per quanto riguarda l'adeguatezza e leggermente inferiore per quanto riguarda la scorrevolezza. La traduzione automatica dei testi ha inoltre ricevuto un punteggio superiore a quello ottenuto, con modalità simili, dalla traduzione automatica di testi giornalistici.

Abstract

The paper presents an evaluation of the performance of DeepL in the translation of specialized texts from English to Italian. The evaluation was carried out at sentence level, on a sample of 108 sentences

taken from texts relating to the environment, energy, bio-medicine and drug science, and the translations produced were evaluated by translators in training, with disciplinary skills. The translation by DeepL was statistically rated at the same level of human translation in terms of adequacy and slightly lower in terms of fluency. Machine translation of the texts also received a higher score than that obtained in another analysis, carried out in a similar way, by machine translation of journalistic texts.

1 Introduzione

La valutazione delle effettive prestazioni dei sistemi di traduzione automatica continua a essere un problema complesso sia dal punto di vista teorico sia dal punto di vista pratico.

Dal punto di vista pratico, è oggi evidente che le metriche di valutazione più usate dopo il Duemila, e in particolare BLEU, non sono in realtà in grado di descrivere adeguatamente le differenze e i miglioramenti di prestazioni dei sistemi oggi in uso, e in particolare di quelli basati su reti neurali (Bentivogli e altri 2018a; Shterionov e altri 2018; Tavosanis 2019). Metriche proposte più di recente, come BERTScore, devono ancora essere valutate a fondo e sembrano comunque fornire risultati molto simili a quelli di BLEU (Zhang e altri 2020). Si è quindi ritenuto metodologicamente

¹ Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Il testo è stato concepito unitariamente dagli autori, ma ai fini della ripartizione del lavoro si dichiara che sono opera di Sirio Papa i paragrafi 4, 7 e 8 e di Mirko Tavosanis i restanti paragrafi.

Per la collaborazione generosamente prestata, si ringraziano le professoressa Silvia Barra e Isabella Blum e gli studenti del Master online in Traduzione specialistica inglese > italiano realizzato dalle università di Genova e Pisa ed erogato dal Consorzio ICoN.

opportuno non usarle per questa valutazione, nemmeno come termine di confronto.

L'unico tipo di valutazione soddisfacente, a oggi, resta quindi quello condotto da valutatori umani. Non tutti i tipi di valutazione umana sono ugualmente soddisfacenti e affidabili. Le valutazioni condotte attraverso crowdsourcing da individui di cui non sono note le competenze assegnano per esempio alla traduzione automatica, sistematicamente, punteggi più alti rispetto a quelli assegnati da persone con provata competenza nella valutazione di traduzioni (Castilho e altri 2017a; Läubli e altri 2020: 658).

In questo contesto non mancano dichiarazioni in cui si rivendica il raggiungimento della "parità" tra traduzione automatica e traduzione umana per alcuni sistemi commerciali (Hassan e altri 2018). Le verifiche indipendenti in alcuni casi hanno confermato questi risultati, ma in altri hanno evidenziato differenze significative (Toral e altri 2018). Questa incertezza è poi in buona parte riconducibile alle circostanze della valutazione, che possono essere molto diverse tra di loro.

Il presente contributo punta a fornire ulteriori risultati inserendosi nel percorso di ricerca descritto in Tivosanis (2019), rispetto al quale rinforza il metodo di indagine e varia la tipologia testuale di riferimento. In Tivosanis (2019) le valutazioni sono state eseguite su testi giornalistici; nel presente contributo sono stati invece scelti testi specialistici. La valutazione punta in primo luogo a valutare la qualità delle traduzioni specialistiche in sé e in secondo luogo a vedere se i punteggi assegnati alle traduzioni specialistiche sono superiori o inferiori a quelli assegnati alle traduzioni di testi giornalistici. La traduzione automatica viene infatti normalmente usata su testi appartenenti a generi molto diversi, e valutare un unico genere è senz'altro molto limitante. (Burchardt e altri 2017: 159-160).

In particolare, date le sensibili differenze linguistiche tra i testi specialistici e i testi non specialistici, sembra verosimile che la stessa tecnologia di traduzione possa produrre risultati molto diversi nei due casi. Assicurare la qualità di traduzioni di testi provenienti da domini diversi è stato quindi considerato un problema fin dalla prima diffusione dei sistemi basati su reti neurali. Koehn e Knowles (2017: 29), per esempio, notando che "in different domains, words have different translations and meaning is expressed in different styles", presentano il *domain mismatch* come prima "sfida" per questi sistemi: nelle loro valutazioni, in questi contesti la NMT otteneva risultati infe-

riori a quelli dei sistemi SMT. Sembra inoltre diffusa l'idea che un sistema di traduzione a reti neurali generalista produca risultati di bassa qualità se applicato a testi specialistici (Chu e Wang 2020: 413). La *domain adaptation* è quindi un importante settore di sviluppo per la NMT (Chu e Wang 2018). Sono però rare, e quasi assenti per l'italiano, le valutazioni condotte con valutatori umani.

2 Il sistema valutato

Le verifiche descritte di seguito sono state compiute usando le traduzioni generate dal sistema DeepL, che è frequentemente segnalato come uno dei migliori prodotti della sua categoria. In particolare, nelle valutazioni comparative DeepL ha ottenuto negli ultimi anni punteggi spesso superiori a quelli di Google Traduttore (Heiss e Soffritti 2018; Tivosanis 2018; Tivosanis 2019).

Realizzato dall'azienda tedesca DeepL GmbH, DeepL è stato reso disponibile al pubblico nell'agosto del 2017 (sito: <https://www.deepl.com/>). Nell'ottobre 2020, il sistema copre un numero relativamente ridotto di lingue: undici in tutto, principalmente indoeuropee (italiano, inglese, tedesco, francese, spagnolo, portoghese, olandese, polacco e russo), con l'aggiunta di giapponese e cinese. Dal punto di vista tecnico, l'azienda ha dichiarato che il sistema di traduzione si basa su reti neurali, ma non ha fornito informazioni più specifiche.

Per quanto riguarda il rapporto con i domini, l'azienda non fornisce nessuna indicazione specifica. Si può quindi ipotizzare che il sistema sia generalista e non specializzato.

3 Composizione del corpus

Per la valutazione del lavoro è stato usato un corpus di testi specialistici di vario genere, composto da testi selezionati casualmente da due docenti del Master online in Traduzione specialistica inglese > italiano erogato congiuntamente dalle Università di Genova e Pisa e gestito dal Consorzio ICoN (<http://www.traduzione.icon-master.it/>).

I testi sono stati scelti dalle docenti di due dei domini trattati dal Master: Ambiente ed energia (professoressa Silvia Barra) e Biomedicina e discipline del farmaco (professoressa Isabella Blum). In tutti i casi dovevano essere disponibili sia il testo originale sia una traduzione professionale in lingua italiana realizzata da esseri umani. Le tipologie testuali sono state selezionate in modo da renderle rappresentative dell'ampia gamma di testi specialistici effettivamente trattati

nel Master: manuali, articoli scientifici, brevetti, schede di sicurezza. La definizione di “testo specialistico” è naturalmente piuttosto arbitraria, e comprende diverse tipologie testuali e diversi generi testuali. Tuttavia, è sembrato perfettamente adeguato agli scopi della valutazione riprendere i tipi di testo usati nella formazione dei traduttori umani professionali, senza distinzioni ulteriori.

4 Formazione del campione

Il campione da esaminare è stato costruito innanzitutto sottoponendo a DeepL, nella loro interezza, i testi selezionati; la traduzione è stata eseguita nel giugno del 2020. Dagli stessi testi sono poi state selezionate casualmente 108 frasi, 40 provenienti dal dominio Ambiente ed energia e 68 dal dominio Biomedicina e discipline del farmaco; la distribuzione per dominio è proporzionale alla consistenza del rispettivo corpus. Si è ritenuto che non fosse possibile indicare a priori uno dei due domini come più difficile da tradurre rispetto all'altro e che quindi non fosse necessario bilanciare la composizione. Nella selezione sono state evitate le frasi ripetute e quelle nominali o disposte in tabella o in elenco.

La dimensione del campione è ridotta rispetto a quello di campagne di valutazione recenti come Intento, che ha preso in esame 500 “segmenti” per numerose coppie di lingue e numerosi domini (Intento 2020). Tuttavia, Intento ha valutato le frasi usando il sistema automatico BERTScore, menzionato al § 1, senza ricorrere a valutatori umani. La dimensione del campione usato qui è invece simile a quelle dei campioni usati in altre esperienze con valutatori umani, condotte per esempio con 150 frasi (Hassan e altri 2018), 299 frasi (Läubli e altri 2020: 657), 104 frasi (Läubli e altri 2020: 658-659), e così via.

Le dimensioni complessive del campione sono state di 2.826 token (1664 per Biomedicina e 1162 per Ambiente). La lunghezza media è quindi di poco superiore ai 26 token per frase e la mediana si attesta a 22 token. La frase più breve è lunga 9 token, mentre quella più lunga 91, ma rappresenta chiaramente un outlier dato che il 75% delle frasi ha una lunghezza entro i 33 token.

Per ogni frase selezionata sono stati raccolti:

1. La frase originale in inglese
2. La corrispondente traduzione in italiano realizzata da un traduttore umano

3. La corrispondente traduzione in italiano realizzata da DeepL

Le 108 frasi tradotte da DeepL sono state divise in due gruppi di 54, denominati gruppo A e gruppo B. All'interno di ogni gruppo sono state poi inserite le altre 54 frasi nella versione realizzata da traduttori umani. Ognuna delle 108 frasi tradotte da DeepL e da esseri umani è stata poi valutata sia per l'adeguatezza (*adequacy*) sia per la scorrevolezza (*fluency*) da ogni valutatore del gruppo a cui era stata associata. Questo ha permesso di ottenere una valutazione di tutte le 108 frasi tradotte da esseri umani e di tutte le 108 frasi tradotte da DeepL.

Durante la valutazione, le frasi state sottoposte ai valutatori umani in ordine casuale e senza indicazioni sulla loro origine: i valutatori non avevano quindi elementi esterni per decidere se l'origine di una singola frase era un traduttore umano o DeepL. Nella valutazione per adeguatezza le frasi erano accompagnate dal testo originale in lingua inglese, secondo l'orientamento *DA-src* (Bentivogli e altri 2018b: 62), mentre nella valutazione per scorrevolezza era disponibile solo il testo italiano. La valutazione è stata eseguita online, usando il sistema KantanLQR², per un tempo medio di un'ora per ogni campione.

5 Criteri di valutazione

Anche se i risultati delle verifiche sulla traduzione automatica condotte in rapporto ai convegni WMT hanno confermato la maggior rilevanza dell'adeguatezza rispetto alla fluenza (Bentivogli e altri 2018b: 62), le due diverse valutazioni sono state conservate per verificare l'esistenza di differenze tra di loro. Va comunque notato che, nonostante sia teoricamente possibile che una frase tradotta con sistemi a reti neurali si allontani molto dal senso testo di partenza, nella pratica non si è prodotto nessun caso di questo genere.

Per l'adeguatezza è stata usata una scala di valori basata su criteri relativi:

1. Il contenuto informativo dell'originale è stato completamente alterato
2. È stata trasmessa una parte del contenuto informativo, ma non la più importante
3. Circa metà del contenuto informativo è stata trasmessa
4. La parte più importante del contenuto informativo originale è stata trasmessa

² KantanLQR è un sistema che fornisce strumenti automatizzati per valutazione e l'analisi di segmenti linguistici. Il sistema è implementato sulla piattaforma di

KantanMT (<<https://kantanmt.com/>>), ma può essere utilizzato indipendentemente da essa, su qualsiasi corpus organizzato e diviso in singole frasi.

5. Il contenuto informativo è stato tradotto completamente

Per la scorrevolezza, sulla base del livello medio di traduzione visto in altre verifiche, la scala è invece stata basata su criteri in parte assoluti:

1. Impossibile da ricondurre alla norma
2. Con più di due errori morfosintattici
3. Con non più di due errori morfosintattici e/o molti usi insoliti di collocazioni
4. Con non più di un errore morfosintattico e/o un uso insolito di collocazioni
5. Del tutto corretta

6 Composizione del gruppo dei valutatori

Il gruppo dei valutatori è stato interamente composto da studenti del Master online in Traduzione specialistica inglese > italiano citato al § 3. La maggior parte dei valutatori, all'interno del Master, aveva approfondito l'uno o l'altro dei domini presi in esame, o entrambi. Tutti avevano comunque l'italiano come lingua madre e disponevano di una conoscenza della lingua inglese valutabile tra C1 e C2. Nessuno di loro è stato coinvolto nella fase di scelta e preparazione degli articoli.

La scelta di valutatori specializzati è conseguenza di due idee di base: innanzitutto, solo le persone dotate di conoscenze disciplinari sono i destinatari normali di testi specialistici; inoltre, solo le persone dotate di conoscenze disciplinari possono valutare con cognizione di causa un testo specialistico. Per esempio, valutare anche solo la correttezza grammaticale di frasi come questa sembra possibile solo a chi sa se nell'italiano specialistico sono o no accettabili sintagmi come *in aperto* e parole come *farmacocinetica*:

“È stato condotto uno studio a dose singola in aperto per valutare la farmacocinetica di una dose ridotta di sitagliptin (50 mg) in pazienti con vari gradi di compromissione renale cronica rispetto a soggetti sani di controllo.”

Per migliorare l'omogeneità del risultato, alcuni mesi prima della valutazione vera e propria è stata fatta una sessione di addestramento con i valutatori interessati. In questa sessione sono state valutate numerose frasi (diverse da quelle esaminate in seguito), e i punteggi assegnati sono stati di-

scussi collettivamente, cercando di arrivare a parametri di valutazione quanto più possibile condivisi

I valutatori sono stati complessivamente 15: 7 per il gruppo A, 8 per il gruppo B. Il numero è quindi superiore a quello usato in valutazioni umane simili, come quelle descritte in Hassan e altri (2018) e Läubli e altri (2020).

7 Esito generale della valutazione

I risultati della valutazione sono riportati in Tabella 1.

Traduttore e sottocorpus	Media adeguatezza	Media scorrevolezza	σ adeguatezza	σ scorrevolezza
Umano complessivo	4,29	4,17	0,43	0,60
Biomedicina	4,38	4,41	0,36	0,38
Ambiente	4,15	3,78	0,49	0,68
DeepL complessivo	4,31	4,09	0,45	0,56
Biomedicina	4,36	4,06	0,48	0,59
Ambiente	4,24	4,14	0,39	0,51

Tabella 1: Risultati della valutazione.

La variazione nei giudizi è, in generale, piuttosto limitata. Per quanto riguarda l'adeguatezza della traduzione umana, la deviazione standard è stata di 0,43 e 39 frasi su 108 hanno ottenuto un punteggio maggiore di 4,50. Solamente 2 frasi hanno ottenuto un punteggio minore o uguale a 3. Le traduzioni di DeepL hanno ottenuto una deviazione standard di 0,45; 40 frasi hanno ottenuto un punteggio maggiore di 4,50, e solo una un punteggio minore o uguale a 3.

La deviazione standard collegata alla scorrevolezza è stata più alta, ma comunque contenuta: 0,60 per la traduzione umana, 0,56 per la traduzione di DeepL. Per la scorrevolezza, va notato, inoltre, che il punteggio 5 è stato assegnato all'unanimità solo a pochissime frasi e il punteg-

gio minimo ottenuto (2,00 in entrambe le traduzioni) è più basso di quello dell'adeguatezza (2,75 per entrambe le traduzioni). Tuttavia, 26 traduzioni di DeepL hanno ottenuto un punteggio medio superiore a 4,5, contro 32 traduzioni umane.

Complessivamente, la traduzione automatica ha ricevuto un punteggio migliore della traduzione umana per quanto riguarda l'adeguatezza, e inferiore per quanto riguarda la scorrevolezza. I dati sono stati, inoltre, sottoposti ad un t-test per verificare la significatività delle differenze. I risultati presentano un p value di 0,762 per l'adeguatezza e un p value di 0,313 per la scorrevolezza. I valori dei p value fanno concludere che, con il 95% di confidenza statistica, non è possibile affermare che i risultati dell'adeguatezza ottenuti da DeepL siano effettivamente migliori dei risultati ottenuti dalla traduzione umana, o viceversa (p value > 0,05). Al contrario, i risultati della scorrevolezza ottenuti dalla traduzione umana possono dirsi significativamente migliori rispetto ai risultati ottenuti dalla traduzione automatica (p value < 0,05).

8 Valutazioni particolari

Per l'adeguatezza, solo una frase tradotta da DeepL ha ottenuto un risultato minore o uguale a 3:

Originale: "After discontinuation of short-term and long-term treatment with pregabalin withdrawal symptoms have been observed in some patients".

Traduzione: "Dopo l'interruzione del trattamento a breve e a lungo termine con sintomi di astinenza da pregabalin sono stati osservati in alcuni pazienti".

Lo stesso è avvenuto per due frasi tradotte da traduttori umani:

Originale: "Ampersand's leadership knew that to keep the product from being cost prohibitive, it'd have to create a model that was sustainable for the people who needed the electric mototaxis the most: the motars".

Traduzione: "I dirigenti di Ampersand sapevano che evitare che il prodotto avesse un costo proibitivo avrebbe creato un modello sostenibile per coloro che avevano bisogno più degli altri del mototaxi elettrico: i motars".

Originale: "This information is based on our current knowledge and is intended to describe the product for the purposes of health, safety and environmental requirements only."

"Queste informazioni sono basate sulle nostre conoscenze attuali e sono intese descrivere il prodotto per il solo scopo dei requisiti di salute, sicurezza e ambientali."

Il punteggio pieno è stato assegnato a 3 frasi tradotte da DeepL:

Originale: "This medicinal product does not require any special storage conditions".

Traduzione: "Questo medicinale non richiede particolari condizioni di conservazione".

Originale: "The other ingredients are: lactose monohydrate, maize starch, talc, gelatine, titanium dioxide (E171), sodium laurilsulphate, anhydrous colloidal silica, black ink, (which contains shellac, black iron oxide (E172), propylene glycol, potassium hydroxide) and water".

Traduzione: "Gli altri ingredienti sono: lattosio monoidrato, amido di mais, talco, gelatina, biossido di titanio (E171), laurilsolfato di sodio, silice colloidale anidra, inchiostro nero, (che contiene gommalacca, ossido di ferro nero (E172), glicole propilenico, idrossido di potassio) e acqua".

Originale: "Animal data do not suggest an effect of treatment with sitagliptin on male and female fertility".

Traduzione: "I dati relativi agli animali non suggeriscono un effetto del trattamento con sitagliptina sulla fertilità maschile e femminile".

Lo stesso punteggio è stato assegnato a una sola frase tradotta da un essere umano:

Originale: "Pregabalin should be discontinued immediately if symptoms of angioedema, such as facial, perioral, or upper airway swelling occur".

Traduzione: "Il trattamento con pregabalin deve essere immediatamente interrotto in presenza di sintomi di angioedema come gonfiore del viso, gonfiore periorale o gonfiore delle vie respiratorie superiori".

Per la scorrevolezza, nessuna frase ha ottenuto un punteggio pieno, né per la traduzione umana né per quella automatica. Sono state più frequenti, invece, le frasi che hanno ottenuto un punteggio minore o uguale a 3. Nel caso delle traduzioni di

DeepL sono state quattro, tra cui per esempio questa:

“L'analisi del ricovero in ospedale per insufficienza cardiaca è stata adattata per una storia di insufficienza cardiaca al basale”.

Le traduzioni umane ad avere ottenuto un punteggio di scorrevolezza minore o uguale a 3 sono state invece cinque, tra cui per esempio questa:

“Modulo di cella solare comprendente un insieme di pre-laminazione per cella solare, in cui l'insieme è come elencato in qualsiasi rivendicazione da 1 a 11.”

9 Confronto con il testo giornalistico

In Tavosanis (2019) la valutazione umana delle traduzioni di testi giornalistici, condotta con gli stessi criteri di valutazione e con un numero di valutatori comparabile, aveva fornito i risultati riportati nella Tabella 2.

Traduttore	N. frasi	Media adeguatezza	Media scorrevolezza
Google	37	4,15	3,90
DeepL	39	4,30	3,94
Umano	24	4,60	4,46

Tabella 2: Valutazione complessiva delle traduzioni di testi giornalistici in Tavosanis (2019).

Confrontando questi risultati con quelli presentati nella Tabella 1, la differenza principale consiste nel peggioramento del punteggio assegnato alla traduzione umana. Se si presuppone che la qualità della traduzione umana sia stabile da una rilevazione all'altra e da un tipo di testo all'altro, questo peggioramento potrebbe essere attribuito a una maggiore severità dei revisori in quanto esperti di dominio (possibilità anticipata nel § 5). Intuitivamente, esistono però numerose altre spiegazioni possibili, in isolamento o in combinazione: per esempio, che il testo specialistico sia più adatto a questo tipo di traduzione automatica rispetto al testo giornalistico, o che sia più difficile da gestire per i traduttori umani. Allo stato attuale delle conoscenze non ci sono fattori che spingano a preferire una spiegazione rispetto a un'altra.

10 Conclusioni e sviluppi futuri

I risultati ottenuti con questa prova supportano l'ipotesi che anche per l'italiano, perlomeno per alcune tipologie testuali e a livello di frase, la traduzione automatica abbia raggiunto un livello qualitativo statisticamente pari a quello della traduzione umana per quanto riguarda l'adeguatezza e leggermente inferiore per quanto riguarda la scorrevolezza. Sono quindi coerenti con diversi altri risultati recenti, presentati per altre lingue (Läubli e altri 2020: 660); va però ricordato che l'italiano non è stato incluso negli importanti task di WMT 2019 (Barrault e altri 2019).

Inoltre, i risultati ottenuti non supportano l'ipotesi che la NMT di sistemi di uso generale ottenga risultati inferiori quando viene applicata a testi specialistici rispetto a quando viene applicata a testi non specialistici.

L'analisi ha naturalmente diversi limiti: per esempio, il campione valutato è relativamente ristretto, le oscillazioni nel giudizio dei valutatori non possono essere confrontate con una media professionale sperimentata e i domini specialistici presi in considerazione sono solo due. Tuttavia, l'estensione e il miglioramento di queste pratiche sembrano a oggi l'unico modo per valutare correttamente le capacità della traduzione automatica in italiano.

Per quanto riguarda gli sviluppi futuri, la necessità di una valutazione realistica sembra rendere indispensabile il passaggio dalla valutazione di singole frasi a quella di testi interi. La qualità della traduzione automatica a livello di testo risulta infatti, in diversi casi, sensibilmente peggiore rispetto a quella a livello di frase (Läubli e altri 2020: 660). La mancanza di sistemi strutturali per garantire la coerenza a livello di testo nella traduzione a reti neurali fa pensare che il fenomeno sia strutturale; per verificare queste ipotesi sono però necessarie valutazioni dedicate.

Al tempo stesso, il confronto con la valutazione dei testi giornalistici suggerisce l'idea che i risultati possano variare in modo sensibile da un genere testuale all'altro, e che almeno in alcuni casi possano essere migliori rispetto a quelli che si ottengono con testi non specialistici. La variabilità collegata al genere non è contemplata nella peraltro dettagliatissima sintesi di Läubli e altri (2020), ma sembra indispensabile prenderla strutturalmente in considerazione per rendere più solide tutte le valutazioni future.

Bibliografia

- Barrault, Loïc, e altri (2019). *Findings of the 2019 Conference on machine translation (WMT 2019)*. In *Proceedings of the WMT*, Firenze, Association for computational linguistics, pp. 1-61.
- Bentivogli, Luisa, e altri (2018a). *Neural versus phrase-based MT quality: an in-depth analysis on English–German and English–French*. In *Computer speech & language*, 49, pp. 52-70.
- Bentivogli, Luisa, e altri (2018b). *Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment*. In *Proceedings of the 15th international workshop on spoken language translation, Iwslt*, pp. 62-69.
- Burchardt, Aljoscha, e altri (2017). *A linguistic evaluation of rule-based, phrase-based, and neural MT engines*. In *The Prague bulletin of mathematical linguistics*, 108, pp. 159-70.
- Castilho, Sheila, e altri (2017). *Crowdsourcing for NMT evaluation: professional translators versus the crowd*. In *Proceedings of translating and the computer*, 39, pp. 1-22.
- Chu, Chenhui e Rui Wang (2018). *A survey of domain adaptation for machine translation*. In *Proceedings of the 27th international conference on computational linguistics*, Association for computational linguistics, pp. 1304-1319.
- Chu, Chenhui e Rui Wang (2020). *A survey of domain adaptation for machine translation*. In *Journal of information processing*, 28, pp. 413-426.
- Hassan, Hany, e altri (2018). *Achieving human parity on automatic Chinese to English news translation*. arXiv preprint arXiv:1803.05567 (2018).
- Heiss, Christine e Marcello Soffritti (2018). *DeepL Traduttore e didattica della traduzione dall'italiano in tedesco—alcune valutazioni preliminari*. In *Translation and Interpreting for Language Learners (TAIL). Lessons in honour of Guy Aston, Anna Ciliberti, Daniela Zorzi*, a cura di Laurie Anderson, Laura Gavioli e Federico Zanettin, Milano, AltLA, pp. 241-258.
- Intento (2020). *Independent multi-domain evaluation of commercial machine translation engines*. Intento, Inc.
- Koehn, Philipp e Rebecca Knowles (2017). *Six Challenges for Neural Machine Translation*. In *First Workshop on Neural Machine Translation*, Association for Computational Linguistics, pp. 28-39.
- Läubli, Samuel, e altri. (2020). *A set of recommendations for assessing human-machine parity in language translation*. In *Journal of artificial intelligence research*, 67, pp. 653-672.
- Shterionov, Dimitar, e altri (2018). *Human versus automatic quality evaluation of NMT and PBSMT*. In *Machine Translation*, 32, 3, pp. 217-235.
- Tavosanis, Mirko (2018). *Lingue e intelligenza artificiale*. Roma: Carocci.
- Tavosanis, Mirko (2019). *Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano*. In *CLiC-it 2019 – Proceedings of the Sixth Italian Conference on Computational Linguistics*, a cura di Raffaella Bernardi, Roberto Navigli e Giovanni Semeraro, CEUR Workshop Proceedings, Aachen University, pp. 1-7.
- Toral, Antonio, e altri (2018). *Attaining the unattainable? Reassessing claims of human parity in neural machine translation*. arXiv preprint arXiv:1808.10432.
- Zhang, Tianyi, e altri (2020). *BERTScore: evaluating text generation with Bert*. arXiv preprint arXiv:1904.09675.

Overprotective Training Environments Fall Short at Testing Time: Let Models Contribute to Their Own Training

Alberto Testoni

DISI, University of Trento
alberto.testoni@unitn.it

Raffaella Bernardi

CIMeC, DISI, University of Trento
raffaella.bernardi@unitn.it

Abstract

Despite important progress, conversational systems often generate dialogues that sound unnatural to humans. We conjecture that the reason lies in their different training and testing conditions: agents are trained in a controlled “lab” setting but tested in the “wild”. During training, they learn to generate an utterance given the human dialogue history. On the other hand, during testing, they must interact with each other, and hence deal with noisy data. We propose to fill this gap by training the model with mixed batches containing both samples of human and machine-generated dialogues. We assess the validity of the proposed method on Guess-What?!, a visual referential game.

1 Introduction

Important progress has been made in the last years on developing conversational agents, thanks to the introduction of the encoder-decoder framework (Sutskever et al., 2014) that allows learning directly from raw data for both natural language understanding and generation. Promising results were obtained both for chit-chat (Vinyals and Le, 2015) and task-oriented dialogues (Lewis et al., 2017). The framework has been further extended to develop agents that can communicate about a visual content using natural language (de Vries et al., 2017; Mostafazadeh et al., 2017; Das et al., 2017a). It is not easy to evaluate the performance of dialogue systems, but one crucial aspect is the quality of the generated dialogue. These systems must in fact produce a dialogue that sounds natural to humans in order to be employed in real-world scenarios. Although there is not a general

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

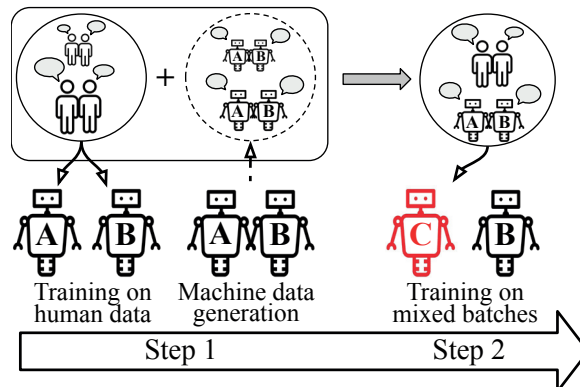
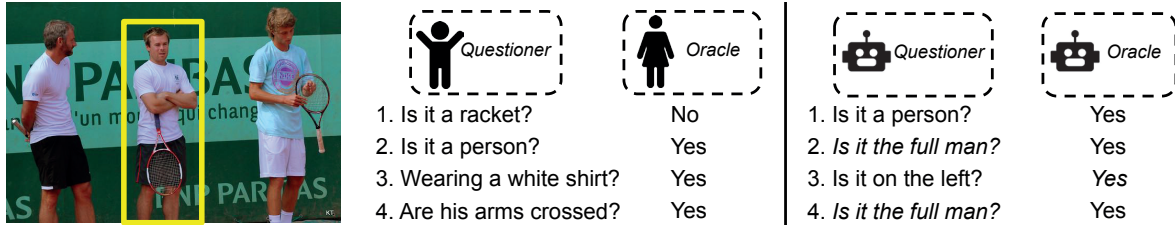


Figure 1: Two-steps training method of the C Bot: two Bots, A and B, are trained independently to reproduce human dialogues; then they play together to generate new dialogues (step 1). In step 2 the Bot C is trained on mixed batches of human and machine-generated data (by A and B in step 1).

agreement on what makes a machine-generated text sound natural, some features can be easily identified: for instance, natural language respects syntactic rules and semantic constraints, it is coherent, it contains words with different frequency distribution but that crucially are informative for the conveyed message, and it does not have repetitions, both at a token and a sentence level.

Unfortunately, even state-of-the-art dialogue systems often generate a language that sounds unnatural to humans, in particular with respect to the large number of repetitions contained in the generated output. We conjecture that part of the problem is due to the training paradigm adopted by most of the systems. In the Supervised Learning training paradigm, the utterances generated by the models during training are used only to compute a Log Likelihood loss function with the gold-standard human dialogues and they are then thrown away. In a multi-turn dialogue setting, for instance, the follow-up utterance is always generated starting from the human dialogue and not from the previ-



Human Annotators	Human Oracle	Machine-generated Questioner	Machine-generated Oracle
1. Is it a racket?	No	1. Is it a person?	Yes
2. Is it a person?	Yes	2. <i>Is it the full man?</i>	Yes
3. Wearing a white shirt?	Yes	3. Is it on the left?	Yes
4. Are his arms crossed?	Yes	4. <i>Is it the full man?</i>	Yes

Figure 2: GuessWhat sample dialogues between two human annotators (left) and two conversational agents (right, generated by GDSE-SL as in Shekhar et al. (2019b)). The yellow box highlights the target entity that the Questioner has to guess by asking binary questions to the Oracle. Both humans and conversational agents have to guess the target object only at the end of the dialogue. Note that the machine-generated dialogue on the right contains repetitions from the Questioner and wrong answers from the Oracle (both in *italic*).

ously generated output. In this way, conversational agents never really interact one with the other. This procedure resembles a controlled “laboratory setting”, where the agents are always exposed to “clean” human data at training time. Crucially, when tested, the agents are instead left alone “in the wild”, without any human supervision. They have to “survive” in a new environment by exploiting the skills learned in the controlled lab setting and by interacting one with the other.

Agents trained in a Reinforcement Learning fashion are instead trained “in the wild” by maximizing a reward function based on the task success of the agent, at the cost of a significant increase of computational complexity. Agents trained according to this paradigm generate many repetitions and the quality of the dialogue degrades. This issue is mildly solved by the Cooperative Learning training, but still, several repetitions occur in the dialogues, making them sound unnatural.

In this paper, we propose a simple but effective method to alter the training environment so that it becomes more similar to the testing one (see Figure 1). In particular, we propose to replace part of the human training data with dialogues generated by conversational agents talking to each other; these dialogues are “noisy”, since they may contain repetitions, a limited vocabulary etc. We then propose to train a new instance of the same conversational agent on this new training set. The model is now trained “out of the lab” since the data it is exposed to are less controlled and they get the model used to live in an environment more similar to the one it will encounter during testing.

We assessed the validity of the proposed method on a referential visual dialogue game, Guess-What?! (de Vries et al., 2017). We found that

the model trained according to our method outperforms the one trained only on human data with respect both to the accuracy in the guessing game and to the linguistic quality of the generated dialogues. In particular, the number of games with repeated questions drops significantly.

2 Related Work

The need of going beyond the task success metric has been highlighted in Shekhar et al. (2019b), where the authors compare the quality of the dialogues generated by their model and other state-of-the-art questioner models according to some linguistic metrics. One striking feature of the dialogues generated by these models is the large number of games containing repeated questions, while the dialogues used to train the model (collected with human annotators) do not contain repetitions. In Shekhar et al. (2019a) the authors enrich the model proposed in Shekhar et al. (2019b) with a module that decides when the agent has gathered enough information and is ready to guess the target object. This approach is effective in reducing repetitions but, crucially, the task accuracy of the game decreases.

Murahari et al. (2019) propose a Questioner model for the GuessWhich task (Das et al., 2017b) that specifically aims to improve the diversity of generated dialogues by adding a new loss function during training: the authors propose a simple auxiliary loss that penalizes similar dialogue state embeddings in consecutive turns. Although this technique reduces the number of repeated questions compared to the baseline model, there is still a large number of repetitions in the output. Compared to these methods, our method does not require to design ad-hoc loss functions or to plug

additional modules in the network.

The problem of generating repetitions not only affects dialogue systems, but instead it seems to be a general property of current decoding strategies. Holtzman et al. (2020) found that decoding strategies that optimize for an output with high probability, such as the widely used beam/greedy search, lead to a linguistic output that is incredibly degenerate. Although language models generally assign high probabilities to well-formed text, the highest scores for longer texts are often repetitive and incoherent. To address this issue, the authors propose a new decoding strategy (Nucleus Sampling) that shows promising results.

3 Task and Models

Task The GuessWhat?! game (de Vries et al. 2017) is a cooperative two-player game based on a referential communication task where two players collaborate to identify a referent. This setting has been extensively used in human-human collaborative dialogue (Clark, 1996; Yule, 2013). It is an asymmetric game involving two human participants who see a real-world image. One of the participants (the Oracle) is secretly assigned a target object within the image and the other participant (the Questioner) has to guess it by asking binary (Yes/No) questions to the Oracle.

Models We use the Visually-Grounded State Encoder (GDSE) model of Shekhar et al. (2019b), i.e. a Questioner agent for the GuessWhat?! game. We consider the version of GDSE trained in a supervised learning fashion (GDSE-SL). The model uses a visually grounded dialogue state that takes the visual features of the input image and each question-answer pair in the dialogue history to create a shared representation used both for generating a follow-up question (QGen module) and guessing the target object (Guesser module) in a multi-task learning scenario. More specifically, the visual features are extracted with a ResNet-152 network (He et al., 2016) and the dialogue history is encoded with an LSTM network. Since QGen faces a harder task and thus requires more training iterations, the authors made the learning schedule task-dependent. They called this setup *modulo-n* training, where n specifies after how many epochs of QGen training the Guesser component is updated together with QGen. The QGen component is optimized with the Log Likelihood of the training dialogues, and the Guesser computes a score

for each candidate object by performing the dot product between visually grounded dialogue state and each object representation. As standard practice, the dialogues generated by the QGen are used only to compute the loss function, and the Guesser is trained by receiving human dialogues. At test time, instead, the model generates a fixed number of questions (5 in our work) and the answers are obtained with the baseline Oracle agent presented in de Vries et al. (2017). Please refer to Shekhar et al. (2019b) for any additional detail on the model architecture and the training paradigm.

4 Metrics

The first metric we considered is the simple task accuracy (ACC) of the Questioner agent in guessing the target object among the candidates. We use four metrics to evaluate the quality of the generated dialogues. (1) Games with repeated questions (GRQ), which measures the percentage of games with at least one repeated question verbatim. (2) Mutual Overlap (MO), which represents the average of the BLEU-4 score obtained by comparing each question with the other questions within the same dialogue. (3) Novel questions (NQ), computed as the average number of questions in a generated dialogue that were not seen during training (compared via string matching). (4) Global Recall (GR), which measures the overall percentage of learnable words (i.e. words in the vocabulary) that the models recall (use) while generating new dialogues. MO and NQ metrics are taken from Murahari et al., (2019) while the GR metric is taken from van Miltenburg et al., (2019). We believe that, overall, these metrics represent a good proxy of the quality of the generated dialogues.

5 Datasets

We are interested in studying how modifying part of the human data in the training set affects the linguistic output and the model’s accuracy on the GuessWhat game. More specifically, we aim at building a training set in which part of the dialogues collected with human annotators are replaced with dialogues generated by the GDSE-SL questioner model while playing with the baseline Oracle model on the same games being replaced. In this way, we build a training set containing dialogues that are more similar to the ones the model will generate at test time while playing with the Oracle.

% Human Dialogues	% Generated Dialogues	Generated Dial. Length	Voc size	MO↓	GRQ↑
100	0	variable	10469	0.05	0
75	25	fixed	4642	0.07	2.9
75	25	variable	4646	0.07	2.6
50	50	fixed	4391	0.08	5.4
50	50	variable	4396	0.07	4.7
0	100	fixed	2586	0.10	10.4
0	100	variable	2680	0.10	10.6

Table 1: Statistics of training sets built with different proportions of human machine-generated dialogues. Human data (100-0) vs. Mixed Batches (75-25, 50-50) vs. Fully Generated data (0-100). Voc size: size of the vocabulary used. GRQ: % games with at least one repeated question verbatim. MO: Mutual Overlap. Refer to Section 4 for additional details on the metrics.

Human data The training set contains about 108K dialogues and the validation and test sets 23K each. Dialogues contain on average 5.2 turns. The GuessWhat?! dataset was collected via Amazon Mechanical Turk by de Vries et al. (2017). The images used in GuessWhat?! are taken from the MS-COCO dataset (Lin et al., 2014). Each image contains at least three and at most twenty objects. More than ten thousand people in total participated in the dataset collection procedure. Humans could stop asking questions at any time, so the length of the dialogues is not fixed. Humans used a vocabulary of 17657 words to play Guess-What?!. 10469 of these words appear at least three times, and thus make up the vocabulary given to the models. For our experiments, we considered only those games in which humans succeeded in identifying the target object and that contain less than 20 turns.

Mixed Batches We let the GDSE-SL model play with the baseline Oracle on the same games of the human training dataset. This produces automatically generated data for the whole training set. The model uses less than 3000 words out of a vocabulary of more than 10000 words. We built new training sets according to two criteria: the proportion of human and machine-generated data (50-50 or 75-25) and the length of the generated dialogue. Either we always keep a fixed dialogue length (5 turns, as the average length in the dataset) or we take the same number of turns that the human Questioner used while playing the game we are replacing.

Table 1 reports some statics of different training sets. Human dialogues have a very low mutual overlap and a much larger vocabulary than

both the generated (0-100) and mixed batches datasets (50-50, 75-25). Looking at the number of games with at least one repeated question in the training set (GRQ column in Table 1), it can be observed that human annotators never produce dialogues with repetitions. The 75/25 dataset configuration contains less than 3% of dialogues with repeated questions and this percentage rises to around 5% for the 50/50 configuration and to around 10% for generated dialogues. Looking at the vocabulary size, the human dataset (100-0) contains around ten thousand unique words, the mixed batches datasets (50-50, 75-25) around 4500 words, and the generated dialogues (0-100) approximately 2500 words.

6 Experiment and Results

6.1 Experiment

As a first step, we trained the GDSE-SL model for 100 epochs as described in Shekhar et al. (2019b). At the end of the training, we used GDSE to play the game with the Oracle on the whole training set, saving all the dialogues. We generate these dialogues with the model trained for all the 100 epochs since it generates fewer repetitions, although it is not the best-performing on the validation set. The dialogues generated by GDSE while playing with the Oracle are noisy: they may contain duplicated questions, wrong answers, etc. See Figure 2 for an example of human and machine-generated dialogues for the same game. We design different training sets as described in Section 5 and train the GDSE-SL model on these datasets. We scrutinize the effect of training on different sets using the metrics described in Section 4 by letting the model generate new dialogues on the test set

% Human Dialogues	% Generated Dialogues	Generated Dial. Length	ACC↑	GRQ↓	MO↓	NQ↑	GR↑
100	0	variable	46.3	36.8	0.27	0.53	20.6
75	25	fixed	47.9	24.0	0.20	0.43	20.2
75	25	variable	47.5	26.6	0.21	0.41	19.4
50	50	fixed	48.1	22.5	0.18	0.37	21.2
50	50	variable	47.0	21.0	0.18	0.42	21.1

Table 2: Test Set 5Q setting. GDSE-SL results on several training sets. At test time, the model generates 5 questions and then it guesses. Length “fixed”: 5-turns dialogues. Length “variable”: same turns human annotators used for that game. ACC: accuracy. GRQ: % games with at least one repeated question. MO: Mutual Overlap. NQ: Novel Questions. GR: Global Recall. ↑: higher is better. ↓: lower is better.

while playing with the Oracle.

6.2 Results

Table 2 reports the results of the GDSE model trained on different training sets. To sum up, there are five dataset configurations: apart from the original GuessWhat dataset composed of dialogues produced by human annotators (100% Human Dialogues), there are datasets composed of 75% human dialogues and 25% generated dialogues or 50% human dialogues and 50% generated dialogues. For each dataset configuration, the generated dialogues can be always 5-turns long (“fixed” length) or they can have the same number of turns human annotators used for that game (“variable” length). We do not report the results on the dataset composed of generated dialogues only since it leads to a huge drop in the accuracy of the guessing game.

By looking at the results on the test set, we can see how even a small number of machine-generated dialogues affects the generation phase at test time, when the model generates 5-turns dialogues and, at the end of the game, it guesses the target object. First of all, it can be noticed that the accuracy of GDSE-SL trained on the new datasets outperforms the one trained on the original training set: in particular, the accuracy of GDSE trained on 50% human dialogues and 50% 5-turns generated dialogues is almost 2% higher (in absolute terms) than the model trained only on human dialogues. The model seems to benefit from being exposed to noisy data at training time to better perform in the guessing game using the dialogues generated by the model itself while playing with the Oracle.

The linguistic analysis of the dialogues generated on the test set reveals that the models trained on “mixed” batches produce better dialogues ac-

cording to the metrics described in Section 4. In particular, considering the best-performing model on the test set, the percentage of games with repeated questions drops by 14.3% in absolute terms and the mutual overlap score by 0.09. The percentage of vocabulary used (global recall), on the other hand, remains stable. Interestingly, the only metric that seems to suffer from the model being trained on mixed datasets is the number of novel questions in the generated dialogue: being trained on noisy data does not seem to improve the “creativity” of the model, measured as the ability to generate new questions compared to ones seen at training time.

Overall, our results show an interesting phenomenon: replacing part of the GuessWhat?! training set with machine-generated noisy dialogues, and training the GDSE-SL questioner model on this new dataset, is found to improve both the accuracy of the guessing game and the linguistic quality of the generated dialogues, in particular with respect to the reduced number of repetitions in the output.

7 Conclusion

Despite impressive progress on developing proficient conversational agents, current state-of-the-art systems produce dialogues that do not sound as natural as they should. In particular, they contain a high number of repetitions. To address this issue, methods presented so far in the literature implement new loss functions, or modify the models’ architecture. When applied to referential guessing games, these techniques have the drawback of gaining little improvement, degrading the accuracy of the referential game, or producing incoherent dialogues. Our work presents a simple but effective method to improve the linguistic out-

put of conversational agents playing the Guess-What?! game. We modify the training set by replacing part of the dialogues produced by human annotators with machine-generated dialogues. We show that a state-of-the-art model benefits from being trained on this new mixed dataset: being exposed to a small number of “imperfect” dialogues at training time improves the quality of the output without deteriorating its accuracy on the task. Our results show an absolute improvement in the accuracy of +1.8% and a drop in the number of dialogues containing duplicated questions of around -14%. Further work is required to check the effectiveness of this approach on other tasks/datasets, and to explore other kinds of perturbations on the input of generative neural dialogue systems.

Acknowledgements

We kindly acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research at the University of Trento. We acknowledge SAP for sponsoring the work.

References

- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.
- Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *2017 IEEE International Conference on Computer Vision*, pages 2951–2960.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European Conference on Computer Vision)*, pages 740–755.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 462–472.
- Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1449–1454.
- Ravi Shekhar, Alberto Testoni, Raquel Fernández, and Raffaella Bernardi. 2019a. Jointly Learning to See, Ask, Decide when to Stop, and then GuessWhat. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019b. Beyond task success: A closer look at jointly learning to see, ask, and Guess-What. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.
- Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2019. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741.
- Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ICML Deep Learning Workshop*.
- George Yule. 2013. *Referential communication tasks*. Routledge.

Topic Modelling Games

Rocco Tripodi

Sapienza NLP Group

Department of Computer Science, Sapienza University of Rome

tripodi@di.uniroma1.it

Abstract

English. This paper presents a new topic modelling framework inspired by game theoretic principles. It is formulated as a *normal form* game in which words are represented as players and topics as strategies that the players select. The strategies of each player are modelled with a probability distribution guided by a utility function that the players try to maximize. This function induces players to select strategies similar to those selected by similar players and to choose strategies not shared with those selected by dissimilar players. The proposed framework is compared with state-of-the-art models demonstrating good performances on standard benchmarks.

Italiano. *Questo articolo presenta un approccio di modellazione dei topic ispirato alla teoria dei giochi. La modellazione dei topic è vista come un gioco in forma normale in cui le parole rappresentano i giocatori e i topic le strategie che i giocatori possono scegliere. Ogni giocatore sceglie le strategie da impiegare tramite una distribuzione di probabilità che viene influenzata da una funzione di utilità che i giocatori cercano di massimizzare. Questa funzione incentiva i giocatori a scegliere strategie simili a quelle impiegate da giocatori simili e disincentiva la scelta di strategie condivise con giocatori dissimili. Il confronto con modelli allo stato dell'arte dimostra buone prestazioni su diversi dataset di valutazione.*

1 Introduction

Topic modeling is a technique that discovers the underlying topics contained in a collection of doc-

uments (Blei, 2012; Griffiths and Steyvers, 2004). It can be used in different tasks of text classification, document retrieval, and sentiment analysis, providing together vector representations of words and documents. State-of-the-art systems are based on probabilistic (Blei et al., 2003; McCalliffe and Blei, 2008; Chong et al., 2009) and neural networks models (Bengio et al., 2003; Hinton and Salakhutdinov, 2009; Larochelle and Lauly, 2012; Cao et al., 2015). A different perspective based on game theory is proposed in this article.

The use of game-theoretic principles in machine learning (Goodfellow et al., 2014), pattern recognition (Pavan and Pelillo, 2007) and natural language processing (Tripodi et al., 2016; Tripodi and Navigli, 2019) problems is developing a promising field of research with the development of original models. The main difference between computational models based on optimization techniques and game-theoretic models is that the former tries to maximize (minimize) a function (that in many cases is non-convex) and the latter tries to find the equilibrium state of a dynamical system. The equilibrium concept is useful because it represents a state in which all the constraints of a given system are satisfied and no object of the system has an incentive to deviate from it, because a different configuration will immediately lead to a worse situation in terms of payoff and fitness, at object and system level. Furthermore, it is guaranteed that the system converges to a mixed strategy Nash equilibrium (Nash, 1951). So far, game-theoretic models have been used in classification and clustering tasks (Pavan and Pelillo, 2007; Tripodi and Pelillo, 2017). In this work, it is proposed a game-theoretic model for inferring a low dimensional representation of words that can capture their latent semantic representation.

In this work, topic modeling is interpreted as a symmetric non-cooperative game (Weibull, 1997) in which, the words are the players and the topics

are the strategies that the players can select. Two players are matched to play the games together according to the co-occurrence patterns found in the corpus under study. The players use a probability distribution over their strategies to play the games and obtain a payoff for each strategy. This reward helps them to adjust their strategy selection in future games, considering what strategy has been effective in previous games. It allows concentrating more mass on the strategies that get high reward. The underlying idea to model the payoff function is to create two influence dynamics, the first one forces similar players (words that appear in similar contexts) to select similar strategies; the second one forces dissimilar players (words that do not share any context) to select different strategies. The games are played repeatedly until the system converges, that is, the difference among the strategy distributions of the players at time t and at time $t - 1$ is under a small threshold. The convergence of the system corresponds to an equilibrium, a situation in which there is an optimal association of words and topics.

2 Related Work

Hofmann (1999) proposed one of the earliest topic models, probabilistic Latent Semantic Indexing (pLSI). It represents each word in a document as a sample from a mixture model, where topics are represented as multinomial random variables and documents as a mixture of topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the most widely used topic model, is a generalization of pLSI that introduces Dirichlet priors for both the word multinomial distributions over topics and topic multinomial distributions over documents. This line of research has been developed building on top of LDA different features to infer correlations among topics (Lafferty and Blei, 2006) or to model jointly words and labels in a supervised way (Mcauliffe and Blei, 2008).

Topic models based on neural network principles have been introduced with the neural network language model proposed in (Bengio et al., 2003). This paradigm is very popular in NLP and many topic models are based on it because with these techniques it is possible to obtain a low-dimensional representation of the data. In particular, auto-encoders (Ranzato and Szummer, 2008), Boltzmann machines (Hinton and Salakhutdinov, 2009) and autoregressive distributions (Larochelle

and Lauly, 2012) have been used to model documents with layer-wise neural network tools. Neural Topic Model (NTM; (Cao et al., 2015)) tries to overcome some limitations of classical topic models, such as the initialization problem and the generalization to n-grams. It exploits word embedding to represent n-grams and uses backpropagation to adjust the weights of the network between the embedding and the word-topic and document-topic layers. A general framework for topic modeling based also on neural networks is Sparse Contextual Hidden and Observed Language Autoencoder (SCHOLAR; (Card et al., 2018)). It allows using covariates to influence the topic distributions and labels to include supervision. As Sparse Additive Generative models (SAGE; (Eisenstein et al., 2011)) it can produce sparse topic representations but differently from it and Structural Topic Model (STM; (Roberts et al., 2014)) it can easily consider a larger set of metadata. A graphical topic model was proposed by Gerlach et al. (2018). In this framework, the task of finding topical structures is interpreted as the task of finding communities in complex networks. It is particularly interesting because it shows analogies with traditional topic models and overcomes some of their limitations such as the bound with a Bayesian prior and the need to specify the number of topics in advance.

3 Topic Modelling Games

Normal-form games consist of a finite set of players $N = (1, \dots, n)$, a finite set of pure strategies, $S_i = \{1, \dots, m_i\}$ for each player $i \in N$ and a payoff (utility) function $u_i : S \rightarrow \mathbb{R}$, that associates a payoff to each combination of strategies $S = S_1 \times S_2 \times \dots \times S_n$. The payoff function does not depend only on the strategy chosen by a single player but by the combination of strategies played at the same time by the players. Each player tries to maximize the value of u_i . Furthermore, in *non-cooperative games* the players choose their strategies independently, considering what other players can play and trying to find the best response to the strategy of the co-players. Nash equilibria (Nash, 1951) represent the key concept of game theory and can be defined as those strategy combinations in which each strategy is a best response to the strategy of the co-player and no player has the incentive to unilaterally deviate from them because there is no way to do better. In addition to play pure strategies, that correspond to select-

ing just one strategy from those available in S_i , a player i can also use *mixed strategies*, which are probability distributions over pure strategies. A mixed strategy over S_i is defined as a vector $\mathbf{x}_i = (x_1, \dots, x_{m_i})$, such that $x_j \geq 0$ and $\sum x_j = 1$. In a *two-player game*, a strategy profile can be defined as a pair $(\mathbf{x}_i, \mathbf{x}_j)$. The expected payoff for this strategy profile is computed as:

$$u(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot A_{ij} \mathbf{x}_j$$

where A_{ij} is the $m_i \times m_j$ payoff matrix between player i and j .

Evolutionary game theory (Weibull, 1997) has introduced two important modifications: 1. the games are played repeatedly, and 2. the players update their mixed strategy over time until it is not possible to improve the payoff. The players, with these two modifications, can develop an inductive learning process, that allows them to learn their strategy distribution according to what other players are selecting. The payoff corresponding to the h -th pure strategy is computed as:

$$u(x_i^h) = x_i^h \cdot \sum_{j=1}^{n_i} (A_{ij} \mathbf{x}_j)^h \quad (1)$$

The average payoff of player i is calculated as:

$$u(\mathbf{x}_i) = \sum_{h=1}^{m_i} u(x_i^h) \quad (2)$$

To find the Nash equilibrium of the game, it is common to use the *replicator dynamics* equation (Weibull, 1997). It allows better than average strategies to grow at each iteration. It can be considered as an *inductive learning* process, in which the players learn from past experiences how to play their best strategy. It is important to notice that each player optimizes its individual strategy space, but this operation is done according to what other players simultaneously are doing so the local optimization is the result of a global process.

Data Preparation The players of the topic modelling games are the words $v = (1, \dots, n)$ in the vocabulary V of the corpus under analysis and the strategies $S = (1, \dots, m)$ are the topics to extract from the same corpus. The strategy space \mathbf{x}_i of each player i is represented as a probability distribution that can be interpreted as the mixture of topics typically used in topic modeling. The interactions among the players are modeled using

the $n \times n$ adjacency matrix (W) of an undirected weighted graph. Each entry w_{ij} encodes the similarity between two words. The strategy space of the games can be represented as a $n \times m$ matrix X , where each row represents the probability distribution of a player over its m strategies (topics that have to be extracted from the corpus).

Payoff Function and System Dynamics The payoff function of the game is constructed exploiting the information stored in W . This matrix gives us the structural information of the corpus. It allows us to select the players with whom each player is playing the games, indicated with the presence of an edge between two nodes (players), and to quantify the level of influence that each player has on the other, indicated with the weight on each edge. The absence of an edge in this graph indicates that two words are distributional dissimilar. Using these three sources of information we model a payoff function that forces similar players to choose similar strategies (topics) and dissimilar players to choose different ones. The payoff of a player is calculated as,

$$u(x_i^h) = x_i^h \left(\sum_{j=1}^{n_i} (A_{ij} \mathbf{x}_j)^h - \sum_{g=1}^{neg_i} (\epsilon \mathbf{x}_g)^h \right) \quad (3)$$

where the first summation is over all the n_i direct neighbors of player i that are the players with whom i share some similarity and the second summation is over the neg_i negative players of player i , that are players with whom player i does not share any similarity. With the first summation player i will negotiate with its neighbors a correlated strategy (topic), with the second he will deviate from the strategies chosen by negative players, this is done by subtracting the payoff that i would have gained if these negative players would have been his neighbors. The negative players are sampled from V according to frequency, in the same way, negative samples are selected in word embeddings models (Mikolov et al., 2013; Tripodi and Pira, 2017). The equation that gives us the probability of selecting a word as negative is:

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^n f(w_j)^{3/4}}, \quad (4)$$

where $f(w_i)$ is the frequency of word w_i . Since the similarity with negative players is 0 we introduced the parameter ϵ to weight their influence and set it to ($A > 0$). The number of negative players,

neg_i , is set to n_i (number of neighbours of player i).

Once the players have played all the games with their neighbors and negative players, the average payoff of each player can be calculated with Equation (2). The payoff is higher when two words are highly correlated and have a similar mixed strategy. For this reason the *replicator dynamics* equation (Weibull, 1997) is used to compute the dynamics of the system. It pushes the players to be influenced by the mixed strategy of the co-players. This influence is proportional to the similarity between two players (A_{ij}). Once the influence dynamics do not affect the players the Nash equilibrium of the system is reached. The stopping criteria of the dynamics and are: 1. the maximum number of iterations (10^5); and 2. the minimum difference between two different iterations (10^{-3}) that is calculated as $\sum_{i=1}^n \mathbf{x}_i(t-1) - \mathbf{x}_i(t)$.

4 Experimental Results

In this section, we evaluate TMG and compare it with state-of-the-art systems.

4.1 Data and Setting

The datasets used to evaluate TMG are 20 Newsgroups¹ (20NG) and NIPS². 20NG is a collection of about 20,000 documents organized into 20 different classes. NIPS is composed of about 1,700 NIPS conference papers published between 1987 and 1999 with no class information. Each text was tokenized and lowercased. The stop-words were removed and the vocabulary was constructed considering the 1000 and 2000 most frequent words in 20NG and NIPS, respectively. This choice is in line with previous work (Card et al., 2018). To keep the model as simple as possible, the *tf-idf* weighting was used to construct the feature vectors of the words and the cosine similarity was employed to create the adjacency matrix A . It is important to notice here that other sources of information can be easily included at this stage, derived from pre-trained word embeddings, syntactic structures or document metadata. Then A is sparsified taking only the r nearest neighbours of each node. r is calculated as $r = \log(n)$ this operation reduces the computational cost of the algorithm and guarantees that the graph remains connected (Von Luxburg, 2007).

¹<http://qwone.com/~jason/20Newsgroups/>

²<http://www.cs.nyu.edu/~roweis/data.html>

Dataset	TMG	SCHOLAR	NVDM	LDA
20NG	824	819	927	791
NIPS	1311	1370	1564	1017

Table 1: Comparison of the models as perplexity.

The strategy space of the players was initialized using a normal distribution to reduce the parameters of the framework³. The last two parameters of the systems concern the stopping criteria of the dynamics and are: 1. the maximum number of iterations (10^5); and 2. the minimum difference between two different iterations (10^{-3}) that is calculated as $\sum_{i=1}^n \mathbf{x}_i(t-1) - \mathbf{x}_i(t)$.

TMG has been compared with SCHOLAR⁴, LDA⁵ and NVDM⁶. We configured the NVDM network with two encoder layers (500-dimensional) and ReLu non-linearities. SCHOLAR has been configured using a more complex setting that consists in a single layer encoder and a 4-layer generator. LDA has been run with the following parameters: $\alpha = 50$, $iterations = 1000$ and $topic_{threshold} = 0$.

4.2 Evaluation

In this section, we compared the generalization performances of TMG and compared them with the models presented in the previous section. For the evaluation we used *perplexity* (PPL), even if it has been shown to not correlate with human interpretation of topics (Chang et al., 2009). We computed perplexity on unobserved documents (C), as.

$$PPL(C) = \exp\left(-\frac{1}{N} \frac{\sum_{n=1}^N \log P(C_n)}{\sum_{n=1}^N D_n}\right) \quad (5)$$

where N is the number of documents in the collection C . Low perplexity suggests less uncertainties about the documents. Held out documents represent the 15% of each dataset. Perplexity is computed for 10 topics for the NIPS dataset and 20 topics for the 20 Newsgroups dataset. These numbers correspond to the real number of classes of each dataset.

Table 1 shows the comparison of perplexity. As reported in previous work (Card et al., 2018), it is

³Experimentally it was also observed that using a Dirichlet distribution to initialize the strategy space with different α parameters did not affect much the performances of the model.

⁴<https://github.com/dallascard/scholar>

⁵<http://mallet.cs.umass.edu>

⁶<https://github.com/ysmiao/nvdm>

difficult to achieve a lower perplexity than LDA. The results in these experiments follow the same pattern, with LDA that has the lowest perplexity, TMG, and SCHOLAR that have similar results, and NVDM that performs slightly worse on both datasets.

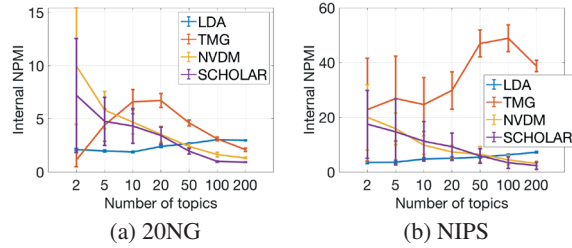


Figure 1: Internal PMI mean and std values.

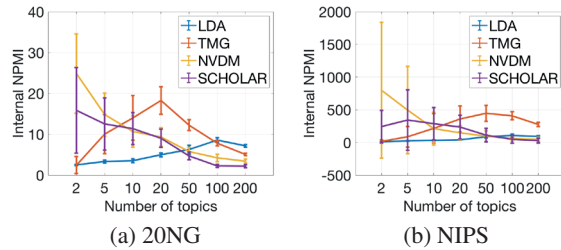


Figure 2: External PMI mean and std values.

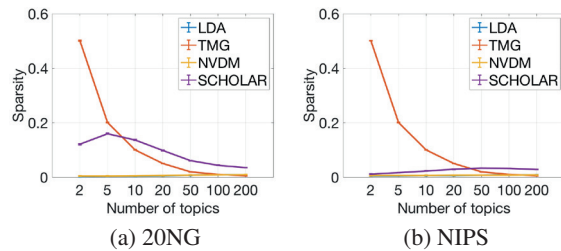


Figure 3: Sparsity mean and std values.

4.3 Topic Coherence and Interpretability

It has been shown that perplexity does not necessarily correlate well with topic coherence (Chang et al., 2009; Srivastava and Sutton, 2017). For this reason, we evaluated the performances of our system also on coherence (Chang et al., 2009; Das et al., 2015). The coherence is calculated by computing the relatedness between topic words using the pointwise mutual information (PMI). We used Wikipedia (2018.05.01 dump) as corpus to compute co-occurrence statistics using a sliding window of 5 words on the left and on the right of

each target word. For each topic, we selected the 10 words with the highest mass. Then we calculated the PMI among all the words pair and finally compute the coherence as the arithmetic mean of all these values. This metric has been shown to correlate well with human judgments (Lau et al., 2017). We used two different sources of information for the computation of the PMI: one is internal and corresponds to the dataset under analysis; the other one is external and is represented by the English Wikipedia corpus.

Internal PMI Figure 1 presents the PMI values of the different models computed on the two corpora. As it is possible to see from figure 1a, TMG has a low PMI compared to all other systems on the 20 Newsgroups dataset when there are few topics to extract (i.e.: 2 and 5). The situation changes drastically when the number of topics increases. In fact, it has the highest performances on this dataset when extracts 10, 20, 50, 100 topics. The performances of NDVM and SCHOLAR are similar and follow a decreasing pattern, with very high values at the beginning. On the contrary, the performances of LDA follow an opposite pattern this model seems to work better when the number of topics to extract is high. On NIPS (Figure 1b) the performances of the systems are similar to those on 20 Newsgroups. The only exception is that TMG has always the highest PMI and seems to behave better also when the number of topics to extract is high. This probably because the number of words in NIPS is higher and for this, it is reasonable to have also a higher number of topics. This is also confirmed from a qualitative analysis of the topics in Section 4.4, where it is demonstrated that with low values of k it is possible to extract general topics and increasing its value it is possible to extract more specific ones.

In general, we can find three different patterns in these experiments: 1. NDVM and SCHOLAR work well on extracting a low number of topics; 2. LDA works well when it has to extract a large number of topics; 3. TMG works well on extracting a number of topics that is close to the real number of classes in the datasets. Another aspect to take into account is the fact that even if TMG has the highest performances, its results have also a high standard deviation. This is due to the stochastic nature of negative sampling.

turks	schneider	drive	vms	god	intellect	bike	providing	fbi	gun	team	space	male	tim	amateur
soviet	allan	ide	disclaimer	jesus	banks	ride	encryption	compound	firearms	game	orbit	gay	israel	georgia
turkish	morality	scsi	vnews	christians	gordon	riding	clipper	batf	guns	play	shuttle	men	israeli	intelligence
armenian	keith	controller	vax	christ	surrender	dod	key	fire	criminals	season	launch	sexual	arab	ai
armenia	atheists	drives	necessarily	christianity	univ	bikes	escrow	waco	crime	hockey	earth	percentage	jews	programs
passes	moral	mb	represents	bible	pittsburgh	motorcycle	crypto	children	weapons	league	mission	study	arabs	michael
roads	political	disk	views	christian	significant	bmw	keys	koresh	criminal	nhl	flight	sex	policy	radio
armenians	pasadena	isa	expressed	faith	hospital	honda	chip	gas	violent	players	nasa	apparent	war	adams
argic	objective	bus	news	church	level	road	secure	branch	weapon	cup	moon	showing	land	ignore
proceeded	animals	floppy	poster	belief	blood	advice	wiretap	started	armed	stanley	solar	women	north	occur
29.71	15.27	12.7	11.72	10.79	10.18	8.94	8.93	8.55	7.52	7.45	7.14	6.92	6.21	6.13

Table 2: Best topics (each topic is represented on the columns) extracted from 20 Newsgroup using TMG (setting $k = 20$) ordered using external PMI (bottom row).

ocular	dendrites	oscillatory	crowdsourcing	kaiming	retina	auditory	graph	disturbances	lifted
eye	dendritic	oscillations	crowds	shaoqing	photoreceptor	sound	edges	plant	propositional
fovea	soma	oscillators	workers	xiangyu	retinal	sounds	graphs	controllers	predicate
dominance	dendrite	oscillator	worker	jian	vertebrate	cochlear	optimisation	controller	grounding
saccades	axonal	oscillation	labelers	yangqing	schulten	ear	edge	disturbance	predicates
saccadic	axons	synchronization	crowd	karen	photoreceptors	hearing	vertices	plants	domingos
fixation	nmda	decoding	turk	sergey	ganglion	ears	optimise	activate	clauses
foveal	pyramidal	locking	wisdom	trevor	kohonen	acoust	optimising	activated	compilation
eyes	somatic	synchronize	expertise	sergio	bipolar	tone	optimised	activating	formulas
saccade	axon	synchronized	dawid	jitendra	visualizing	cochlea	vertex	activates	logical
304.85	283.66	276.39	230.5	218.51	196.86	176.75	146.3	146.25	145.84

Table 3: Topics extracted from NIPS using TMG (setting $k = 10$) ordered using external PMI (bottom row).

Sparsity We compared the sparsity of the word-topics matrices, X , in Figure 3a and 3b, computed as $s = \frac{|X_{>10^{-3}}|}{|X|}$. From both figures, we can see that TMG can produce highly sparse representations especially when the number of topics to extract is low. This is a nice feature since it provides more interpretable results. Only SCHOLAR produces more sparse representations when the number of topics to extract is high. Experimentally we also noticed that we can control the sparsity of X , in TMG, increasing the number of iterations of the game dynamics.

4.4 Qualitative Evaluation

Examples of topics extracted from 20NG and NIPS are presented in Table 2 and 3, respectively⁷. The first difference that emerges from these results are the external PMI values. This is due to the fact that the texts in NIPS have a very specific language and for this reason the PMI values are very high. We can also see that TMG groups highly coherent set of words in each topic. We can easily identify in Table 2 the topics in which the dataset is organized and especially: talk.politics.midleast, alt.atheism, comp.graphics, soc.religion.christian, talk.politics.misc, rec.motorcycles, sci.crypt, talk.politics.guns, rec.sport.hockey, sci.space, talk.politics.misc.

⁷for space limitation we presented only 15 topics for 20NG

We can also easily identify from Table 3 highly coherent topics, related to optic, signal analysis, optimization, crowdsourcing, audio, graph theory and logics. We noticed from these topics that they are general and that it is possible to discover more specific topics increasing the number of topics to extract. For example, we discovered topics related to topic modelling and generative adversarial networks.

5 Conclusion and Future Work

In this paper, it is presented a new topic modeling framework based on game-theoretic principles. The results of its evaluation show that the model performs well compared to state-of-the-art systems and that it can extract topically and semantically related groups of words. In this work, the model was left as simple as possible to assess if a game-theoretic framework itself is suited for topic modeling. In future work, it will be interesting to introduce the topic-document distribution and to test it on classification tasks and covariates to extract topics using different dimensions, such as time, authorship, or opinion. The framework is open and flexible and in future work, it will be tested with different initializations of the strategy space, graph structures, and payoff functions. It will be particularly interesting to test it using word embedding and syntactic information.

References

- [Bengio et al.2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- [Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- [Blei2012] David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April.
- [Cao et al.2015] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *AAAI*, pages 2210–2216.
- [Card et al.2018] Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the ACL*, volume 1, pages 2031–2040.
- [Chang et al.2009] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296.
- [Chong et al.2009] Wang Chong, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *CVPR, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE.
- [Das et al.2015] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the ACL*, volume 1, pages 795–804.
- [Eisenstein et al.2011] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text.
- [Gerlach et al.2018] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. A network approach to topic models. *Science Advances*, 4(7).
- [Goodfellow et al.2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.
- [Griffiths and Steyvers2004] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.
- [Hinton and Salakhutdinov2009] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *NIPS*, pages 1607–1614.
- [Hofmann1999] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 50–57. ACM.
- [Lafferty and Blei2006] John D Lafferty and David M Blei. 2006. Correlated topic models. In *NIPS*, pages 147–154.
- [Larochelle and Lauly2012] Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *NIPS*, pages 2708–2716.
- [Lau et al.2017] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the ACL*, volume 1, pages 355–365.
- [Mcauliffe and Blei2008] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *NIPS*, pages 121–128.
- [Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- [Nash1951] John Nash. 1951. Non-cooperative games. *Annals of mathematics*, pages 286–295.
- [Pavan and Pelillo2007] Massimiliano Pavan and Marcello Pelillo. 2007. Dominant sets and pairwise clustering. *IEEE transactions on pattern analysis and machine intelligence*, 29(1).
- [Ranzato and Szummer2008] Marc’Aurelio Ranzato and Martin Szummer. 2008. Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning*, pages 792–799. ACM.
- [Roberts et al.2014] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.
- [Srivastava and Sutton2017] Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*.
- [Tripodi and Navigli2019] Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99, Hong Kong, China, November. Association for Computational Linguistics.

- [Tripodi and Pelillo2017] Rocco Tripodi and Marcello Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, 43(1):31–70.
- [Tripodi and Pira2017] Rocco Tripodi and Stefano Li Pira. 2017. Analysis of italian word embeddings. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*.
- [Tripodi et al.2016] Rocco Tripodi, Sebastiano Vascon, and Marcello Pelillo. 2016. Context aware nonnegative matrix factorization clustering. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 1719–1724.
- [Von Luxburg2007] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.
- [Weibull1997] J. W. Weibull. 1997. *Evolutionary game theory*. MIT press.

Dialog-based Help Desk through Automated Question Answering and Intent Detection

Antonio Uva[‡], Pierluigi Roberti[†], Alessandro Moschitti[‡]

[‡]DISI, University of Trento, Italy

[†]Im Service Lab Srl, Italy

Abstract

Modern personal assistants require to access unstructured information in order to successfully fulfill user requests. In this paper, we have studied the use of two machine learning components to design personal assistants: intent classification, to understand the user request, and answer sentence selection, to carry out question answering from unstructured text. The evaluation results derived on five different real-world datasets, associated with different companies, show high accuracy for both tasks. This suggests that modern QA and dialog technology is effective for real-world tasks.

I moderni personal assistant richiedono di accedere ad informazioni non strutturate per soddisfare con successo le richieste degli utenti. In questo articolo, abbiamo studiato l'uso dell'apprendimento automatico per progettare due componenti di un personal assistant: classificazione degli intenti, per comprendere la richiesta dell'utente, e la selezione della frase di risposta per rispondere alle domande con testo non strutturato. I risultati della valutazione derivati da cinque diversi datasets del mondo reale, associati a diverse società, mostrano un'elevata precisione per entrambi i modelli. Ciò suggerisce che la moderna tecnologia di question answering e dialogo è efficace per attività reali.

1 Introduction

Help-desk applications use Machine Learning to classify user's request into intents. The informa-

tion owned by companies generally is in free text form, from company's documents or websites. For example, corporate knowledge is typically encoded within documents in an unstructured way. This poses limitations on the effectiveness of standard information access. For example, searching documents by keywords is not a viable solution for the users, as they seldom can find an answer to their questions. The possibility of using QA systems to search for information on a corpus of documents, also through a dialogue system, offers an attractive solution for extracting the best information from the company knowledge bases.

IMSL company offers virtual agents that can be retrained based on the customer needs. The agent is composed of many Natural Language Understanding components, such as classifiers that map each user utterance in input to their corresponding intent. However, since it is not possible to forecast all the intents corresponding to the questions that the user are going to ask – which are potentially infinite – it is of paramount importance to have an automated QA system able to automatically provide the best answer (paragraph) extracted from a company owned knowledge base.

Information access is becoming an increasingly critical issue. Traditional Information Retrieval systems, used in industry, help the user in accessing information, but are often imprecise and impractical. Current search engines are an example of this. Searching for information on the web often requires a double effort for the user: first it is necessary to understand how to formulate a query in the most effective manner, and then filter out the proposed results in order to find the most relevant information.

In this paper, we described our QA system based on answer sentence selection and intent detection, and how we integrate them in a Conversational agent.

⁰Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

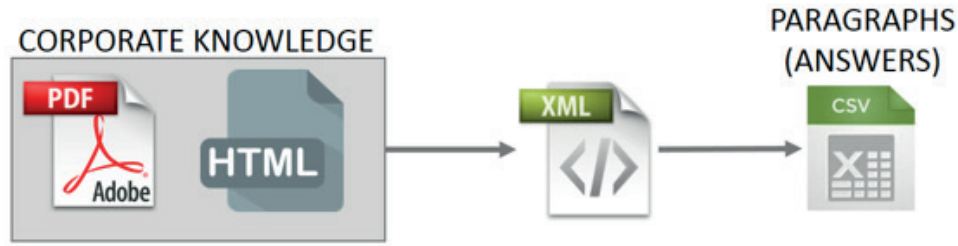


Figure 1: Paragraphs extraction.



Figure 2: Annotation data for training.

2 Related Work

As today, the largest part of general-purpose QA services are provided by big tech companies such as Amazon Alexa, Google Home, Ask Yahoo!, Quora and many others. Unfortunately, these types of applications are not easily accessible for smaller companies, as the offered QA service cannot be easily adapted to handle corporate knowledge, which is in form of unstructured text. To build their own solutions SMEs can exploit QA components such as Answer Sentence Selection.

In recent years, deep learning approaches have been successfully applied for automatically modeling text pairs, e.g., (Lu and Li, 2013; Yu et al., 2014). Additionally, a number of deep learning models have been recently applied to QA, e.g., Yih et al. (2013) applied CNNs to open-domain QA; Bordes et al. (2014) propose a neural embedding model combined with the knowledge base for open-domain QA. Iyyer et al. (2014) applied recursive neural networks to factoid QA over paragraphs. Miao et al. (2016) proposed a neural variational inference model and a Long-short Term Memory network for the same task. Yin et al. (2016) proposed a siamese convolutional network for matching sentences that employ an attentive average pooling mechanism, obtaining state-of-the-art results in various tasks and datasets.

The work closest to this paper is by Yu et al. (2014) and Severyn and Moschitti (2015). The

former presented a CNN architecture for answer sentence selection that uses bigram convolution and average pooling, whereas the latter use convolution with k-max pooling.

Nowadays, supporting customers in their activities across applications and websites is becoming always more demanding, due a large number of customers and the variety of topics that have to be covered.

New tools, such as chatbots, able to answer frequently asked questions, i.e., FAQs, are rising in response to this needs. Classifying the user need expressed in a natural question, into a predefined set of categories, allow conversational agents to recognize which users are asking which types of questions and to react accordingly.

Traditional approaches to this problem include the use supervised approaches such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Boosting (Iyer et al., 2000; Schapire and Singer, 2000), Kernel machines operating on input structured objects (Moschitti, 2006; Lodhi et al., 2002) and Maximum Entropy models (Yaman et al., 2008).

In the latest years, new models such as Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) (Cortes and Vapnik, 1995), Gated Recurrent Units (GRU) (Chung et al., 2014) and Convolutional Neural Networks (CNN) (Lecun et al., 1998; Kim, 2014) were established as state-of-the-art ap-

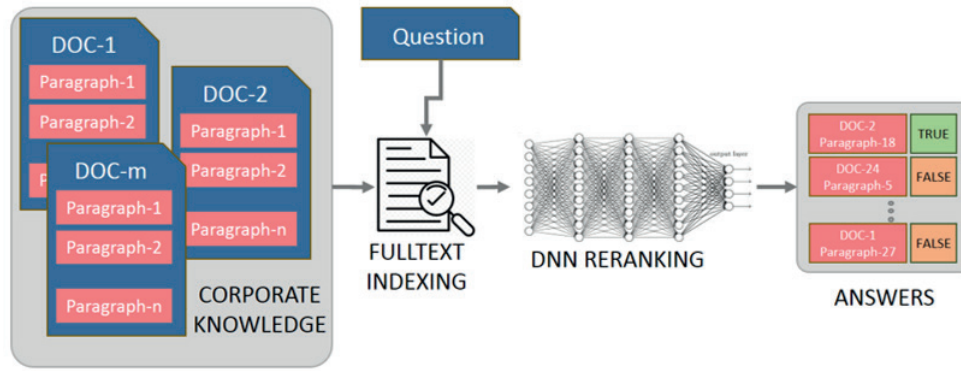


Figure 3: system architecture.

proaches for text classification.

3 System Description

Our QA system allows for extracting portions of text from company documents or from websites. This information is then organized into paragraphs, which are then used to provide an answer to the user’s questions.

One practical problem is the fact that not all PDF files encode text, and many fail to preserve the logical order of the text. Thus, in order to extract paragraphs, we used pdf2text.

Another practical problem we need to solve was to keep portions of text separated by punctuation together: such as bullet lists or very structured paragraphs. Our designed tool automatically assigns a reference index or summary to each paragraph to improve subsequent searches (see Figure 1).

Subsequently, each question and answer pair must be annotated with correctness (label TRUE/FALSE). This allows us to create a training set to train the re-ranking network (see Figure 2).

The final system, shown in Figure 3, therefore allows for using the target company data, appropriately reorganized into paragraphs, to provide answers to the user’s request. On average we provide from 3 to 5 answers for each question. However, we also provide the reference to the document and the summary which the paragraph refers to.

4 Answer Sentence Selection (AS2)

The AS2 goal is to rank a list of answer candidates by their similarity with respect to an input question q_i . We design a network that includes relational information between questions and answers. Our results show that CNNs reach better performance

than traditional IR models based on bag of words.

4.1 Model

The architecture of the network used for mapping sentences in embedding vectors is shown in Figure 4 and is inspired to the CNNs employed by Severyn and Moschitti (2015) to perform many classification activities over sentences. It includes two main components:

(i) an encoder that map an input document s_i into a vector x_{s_i} and (ii) a feed-forward network that computes the similarity between input sentences.

Our network takes two sentences in input, i.e., a question and a text paragraph that may contain an answer, and it represents each of them into vectors of fixed-size dimension $x_s \in \mathbb{R}^m$.

The sentence model is composed of a sequence of convolutional maps followed by some pooling operations. Such model achieves the state of the art in many NLP tasks (Kalchbrenner et al., 2014; Kim, 2014).

Then, the sentence vectors, x_{s_i} corresponding to the questions and answers, are concatenated together and passed to the following neural network layers. These are composed of a non-linear hidden layer and an output layer with a sigmoid activation unit. At the end, the network returns a value between 0 and 1 corresponding to the relevancy of the answer with respect to the question.

Finally, we included word overlap embeddings encoding relational information between words in questions and answers (Severyn and Moschitti, 2016).

5 Intent Classification

We adopted advanced techniques, such as deep learning models, to classify the user need, which is

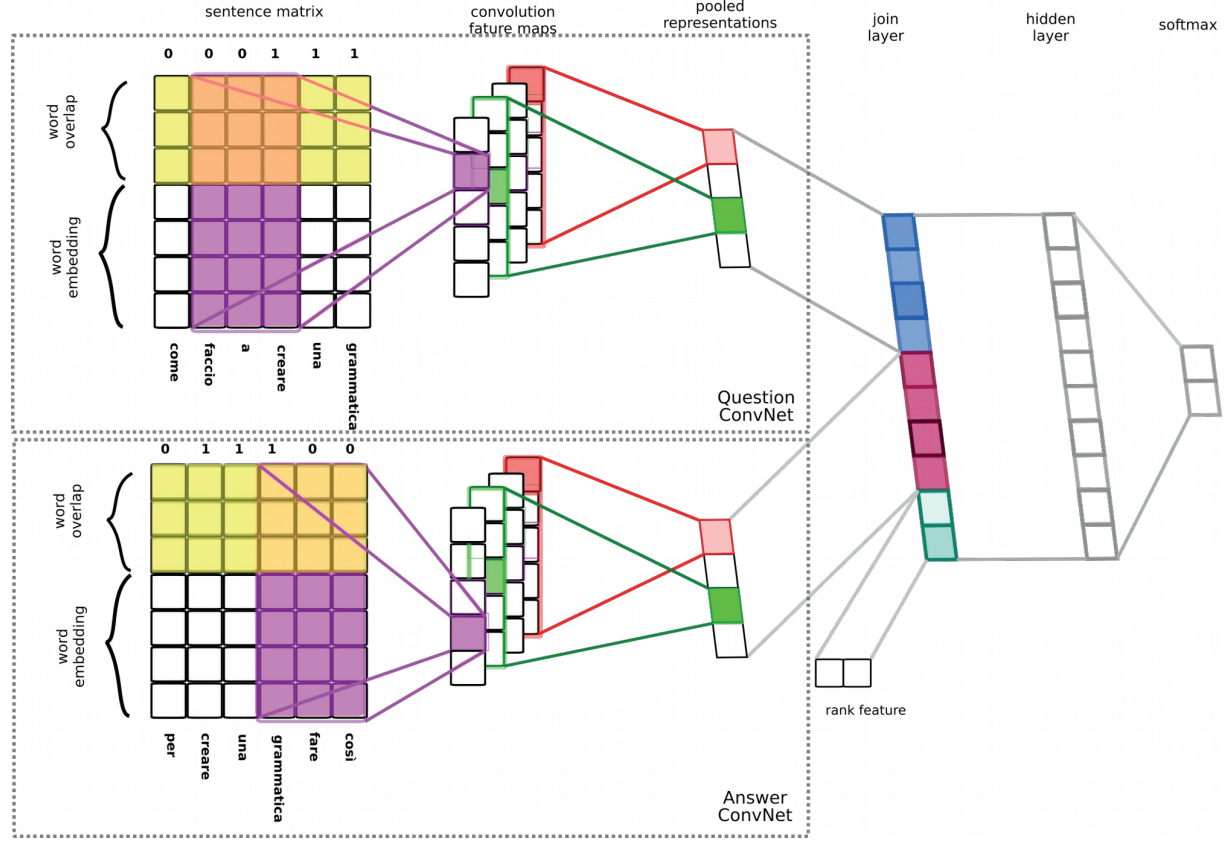


Figure 4: Architecture of the network computing relevancy of answers with respect to the questions. The network is composed by two subnetworks: (1) the Question ConvNet that encodes input questions into a fixed-size vector and (2) the Answer ConvNet that encodes the answer into a fixed-size vector. The vectors of questions and answers are concatenated into a new vector (join layer), where a new embedding is added, which embeds rank information. Then, a Multilayer perceptron (MLP) composed of an hidden layer and a softmax classifier, returns a value between 0 e 1. This indicates the relevancy of an answer with respect to a question.

semantically expressed by the user question, into a predefined set of categories, i.e., intents.

We used some common deep learning models for solving the intent detection task. The main point of our study is to test those models and observe how they perform on datasets containing real user questions addressed to a virtual agent, operating in the banking/financial sector.

At this stage, we do not consider novel methods based on transformer architecture such as BERT (Devlin et al., 2019), which require a large amount of resources, typically not available to SMEs. Instead, we focused on lighter approaches that can run on small GPUs. We report our experiments and discuss the obtained results using such lighter models.

5.1 Models

SVM (baseline) fed with word features, derived from the text of the utterances.

LSTM using recurrent units that take in input the embedding x_t of the current word at time step t and the hidden vector encoding the sub-phrase at previous step, i.e., h_{t-1} , and return the vector representation of the phrase at step h_t .

CNN uses a set of convolutional filters of different size and max pooling operations to extract the most important features, e.g., bigrams, trigrams, etc., which represent the sentence meaning.

LSTM + CNN based on an architecture composed of two layers: an LSTM layer that builds a fixed-size vector representation of the sentence at each word, and a convolutional layer. The latter applies a set of convolutional operations on the

Models	DEV. SET			TEST SET		
	MAP	MRR	P@1	MAP	MRR	P@1
BM25	64.20 \pm 0.00	70.20 \pm 0.00	57.60 \pm 0.00	55.40 \pm 0.00	62.40 \pm 0.00	46.70 \pm 0.00
CNN	65.04 \pm 1.10	69.34 \pm 1.36	53.34 \pm 2.66	68.38 \pm 1.08	72.21 \pm 1.33	57.42 \pm 2.16

Table 1: The results of the QA model on the dev. and test set of IMSL-WIKI corpus

representations returned by the first layer.

CNN + CNN composed of two CNN layers, where the second layer takes the previous layer representation as input, and applies a set of convolutional filters and pooling operation to compute the final vector representation of the sentence.

6 Experiments

In this section, we first describe the datasets we used in our experiments, then we provide the results on the answer sentence selection and the intent classification tasks. Finally, we report an end-to-end evaluation of our system.

6.1 Data Description

We built our datasets by collecting samples of questions asked by users to conversational agent for either **Credit Institution** or **Bank** websites. We collected two intent corpora from each data provider, resulting in a total of four datasets.

Istituto Credito - synthetic (IC_s): This corpus was created by expert dialog engineers. It contains a set of utterances annotated with their corresponding intents. The subject of questions are diverse and spans over many topics. For example, some questions seek information over the bank branch locations, problems regarding how to cash checks, and requests of availability of finance products. It contains 2,305 training examples, and 593 test examples, for a total of 2,898 examples.

Istituto Credito - full (IC_f): This dataset is composed of synthetic questions, generated by language engineers. Subsequently, it has been augmented to take into account also real sentences, retrieved from website chat-bot of a well known Credit Institution operating in Italy. It contains 2,898 training examples and 770 test examples, for a total of 3,668 examples.

Banca - Area Informativa ($Banca_{AI}$): This dataset contains real questions asked by users about the Area Informativa of a bank. It includes 3,947 training examples, and 987 test examples, for a total of 4,934 examples divided in 282 intents.

Banca - Internet Banking ($Banca_{IB}$): This

dataset includes real questions asked by users about the iBanking service offered from a well known Italian bank. It includes 4,380 training instances and 1,906 test instances divided in 251 intents.

Answer Sentence Selection data: We used an in-house dataset called IMSL-WIKI, which contains a list of question and answer regarding some of the products and services sold by IM Service Lab. For each question, a paragraph list was collected using an off-the-shelf search engine, i.e., Lucene, and manually annotated as either relevant or irrelevant. The dataset is divided into two parts, i.e., a training and test sets, which contain a total of 5,190 and 1,240 QA pairs, respectively. For each question, we retrieved a list of 10 candidate answers.

6.2 Model results

In this section we report the performance of our two main machine learning components of our system: Answer Sentence Selection and Intent Classification.

6.2.1 Answer Sentence Reranking

Table 1 reports the performance of the neural network and the baseline system. The first row, i.e., BM25, shows the baseline system, while the second row shows the performance of the CNN. The systems are evaluated according to the Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision at 1 (P@1). The final results reported at the bottom is obtained as the average of 5 different models trained and evaluated on the test set. For each measure in the table, we report both mean and standard deviation computed on dev. and test sets.

We used a small fraction of the training set, i.e., 15% of the data, for early stopping. As it can be seen from the table, CNN performs about 1 point more than the baseline algorithm (BM25) in terms of MAP on the dev. set, and almost 10 absolute points more of MAP on the test set.

In addition, we observe an increase of 9.8 absolute points in terms of MRR, and 10.65 absolute points of P@1 on the test set. The difference between results on dev. and test sets can be explained

Models	Test Sets (Accuracy)			
	IC_s	IC_f	$Banca_{AI}$	$Banca_{IB}$
Baseline (SVM)	0.7622	0.8065	0.8197	0.7235
CNN	0.7718	0.8058	0.8241	0.7633
CNN + CNN	0.7577	0.8094	0.8328	0.7663
LSTM	0.7698	0.8131	0.8529	0.7843
LSTM + CNN	0.7737	0.8231	0.8224	0.7479

Table 2: Accuracy over the datasets.

Models	Test Sets (F_1 score)			
	IC_s	IC_f	$Banca_{AI}$	$Banca_{IB}$
Baseline (SVM)	0.7595	0.8151	0.8009	0.7108
CNN	0.7722	0.8078	0.8064	0.7476
CNN + CNN	0.7606	0.8117	0.8158	0.7499
LSTM	0.7689	0.8163	0.8386	0.7691
LSTM + CNN	0.7742	0.8252	0.8065	0.7344

Table 3: F_1 score over the datasets.

by the fact that the used dev. set is very small: only 124 list of questions and 1,239 Q/A pairs, which made it difficult to optimize the three ranking metrics at the same time, so we focused on MAP.

6.2.2 Intent Classification

We ran state-of-the-art neural classifiers described in Section 6.2.1 on Credit Institute and Bank datasets. To choose the best performance, we used 30% of training data as validation set and select the best hyperparameters. We compare the performance of neural models with respect to strong baseline classifiers, i.e., SVMs, and report the results in terms of Accuracy (Table 2) and F_1 (Table 3). The tables show that the final performance heavily depends on the used dataset and models.

Istituto Credito (IC) datasets. Regarding the IC synthetic dataset, the best model, i.e., LSTM+CNN, obtains Accuracy of 77.37 and a micro-avg F_1 of 0.7742. This is about one absolute point of Accuracy higher than the base SVM model (77.37 vs. 76.22) and 1.47 absolute points of F_1 more than the base model (77.42 vs. 75.95). Similarly, on the IC full dataset, the performance of the best model, i.e., LSTM+CNN, achieved an accuracy of 82.31%, which is 0.66% absolute points better than the base model (82.31 vs. 80.65) and a micro-avg F_1 of 82.52, which is about one point better than the base SVM model (82.52 vs 81.51).

Banca datasets. Regarding Banca AI dataset, the

best model, i.e., LSTM obtained accuracy of 85.29, which is about 4 absolute points better than the base SVM model (85.29 vs. 81.97). Also, in terms of F_1 , the best model obtained 3.77 absolute points more than the baseline (82.86 vs 80.09). Regarding the Banca IB dataset, the best model, i.e., LSTM, obtained around 6 points more both in terms of Accuracy (78.43 vs 72.35) and F_1 (76.91 vs 71.08).

6.3 End-to-End system evaluation

We trained and evaluated our system using samples of data collecting from IMSL customers.

We noted that the accuracy of the system improved because more answers are generally provided (from 3 to 5) to the user’s question, thus allowing to almost certainly provide the correct answer.

The only point of attention is the fact that there is not always a valid answer to the user’s request in company knowledge. Indeed, the questions related to the user’s personal profile or data cannot be precisely answered by the company documentation.

Furthermore, it often happens that the company policy prevents to provide explicit answers to specific user problems. In all these cases, it is therefore necessary to support the QA system with operators, who can provide personal answers or those not coded in the corporate knowledge.

7 Conclusions

In this paper, we have presented a modern dialog system for real-world applications. We have tested advanced technology for QA and intent classification on several datasets derived from company data, such as Banks and Credit Institutions. The results show a promising direction for SMEs to build their own effective access to unstructured data.

References

- Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620, Doha, Qatar, October. Association for Computational Linguistics.
- Junyoung Chung, Çağlar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. In *Machine Learning*, pages 273–297.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Raj D Iyer, David D Lewis, Robert E Schapire, Yoram Singer, and Amit Singhal. 2000. Boosting for document routing. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 70–77.
- Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 633–644.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov.
- Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.
- Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In *Advances in neural information processing systems*, pages 1367–1375.
- Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.
- Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *11th conference of the European Chapter of the Association for Computational Linguistics*.
- Robert E Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2-3):135–168.
- Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382.
- Aliaksei Severyn and Alessandro Moschitti. 2016. Modeling relational information in question-answer pairs with convolutional neural networks. *arXiv preprint arXiv:1604.01178*.
- Sibel Yaman, Li Deng, Dong Yu, Ye-Yi Wang, and Alex Acero. 2008. An integrative and discriminative technique for spoken utterance classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1207–1214.
- Scott Wen-tau Yih, Ming-Wei Chang, Chris Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models.
- Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.
- Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

#andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline

Pierluigi Vitale, Serena Pelosi, Mariacristina Falco

Department of Political and Communication Sciences

University of Salerno

[pvitale,spelosi,mfalco]@unisa.it

Abstract

This research investigates Instagram users' sentiment narrated during the lockdown period in Italy, caused by the COVID-19 pandemic. The study is based on the analysis of all the posts published on Instagram under the hashtag #andràtuttobene on May 4, May 18 and June 3, 2020. Our research carried out a view on a national, regional and provincial scale. We analyzed all the different languages and forms (i.e. captions, hashtags, emojis and images) that constitute the posts. The aim of this research is to provide a set of procedures revealing the different polarity trends for each kind of expression and to propose a single comprehensive measure.

Introduction

This paper investigates the case of the Italian most used hashtag about the lockdown period for the COVID-19 pandemic on Instagram: #andràtuttobene¹.

The research team collected 7,482 posts, the entire amount published in three specific dates: May 4, May 18 and June 3, corresponding with three different steps of the reopening phase of the country, led by the government.

Instagram posts are composed by several kind of languages: captions (texts), hashtags, emojis and images. The aim of this work is to design a set of procedures revealing the different polarity trends for each one and to propose a unique measure. This measure can show the sentiment expressed by the texts, in their semiotic broad meaning.

The methodology proposed is based on a fully automatic natural language processing pipeline, including the images' analysis phase. Its output is an interactive dashboard (Figure 1) that is able to explore the sentiment analysis values about every single kind of text, and the synthesis of all of them. Thanks to a system of interactions and filters, the observation is leaded by the images' features, such as different kind of spaces (indoor or outdoor) and different kind of the photos' subject (human or not human).

The collected geographical data enabled the analysis of several dimensions, with an overview observation based on the regional scale. Hence, it gave us an opportunity to focus on the deeper level of the Italian provinces. This choice is motivated by the Italian DPCM (*Decreto Presidenza del Consiglio dei Ministri*) published on 24 March 2020², in which it is stated the partial autonomy of the regions.

1. State of the Art

In Natural Language Processing (NLP) studies, the automatic treatment of opinionated expressions and documents is known as Sentiment Analysis.

Lexical resources for sentiment analysis created for the Italian Language are *Sentix* (Basile, Nissim 2013); *SentIta* (Pelosi 2015a); the lexicon of the *FICLIT+CS@UniBO* System (Di Gennaro 2014); the *CELI Sentiment Lexicon* (Bolioli 2013); the *Distributional Polarity Lexicons* (Castellucci 2016).

For the Italian language, significant contributions on sentiment analysis of social media come from Bosco et al. (2013, 2014), Castellucci (2014, 2016) and Stranisci (2016), among others.

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹ The choice of Instagram is due to its success. It is in fact one of the most popular social networks, with 1

billion monthly active users, according to a study by *Hootsuite* and *WeAreSocial*.

² <https://www.gazzettaufficiale.it/eli/id/2020/06/17/20G00071/sg>

Hashtag processing in Sentiment Analysis is particularly challenging in terms of word segmentation. Obviously, the absence of white spaces between words poses several problems that concerns ambiguity. Among the most relevant contribution in this area, we cite Zangerle (2018), Reuter (2016), Simeon (2016); Bansal 2015, Srinivasan (2012) and Celebi (2018). The solution proposed in literature concerns mostly the use of n-grams, syntactic complexity, pattern length, or pos-tagging.

In the last years, the way to communicate online involves many kinds of languages, connected to verbal and non-verbal features. This complexity makes classical textual analysis less adequate to have a real and representative perspective on people's interests and opinions. In particular way, the conventional approach seems to be not suitable for visual social media, such as Instagram, where all the languages are involved and the images seem to be dominant.

The analysis of these social media tends to underline the issues of *textocentricity* (Singhal & Rattine-Flaherty, 2006) and *textocentrism* (Balomenou & Garrod, 2019), making necessary a different way to approach the participant generated images (PGI) or user generated contents (UGC) in general.

Opinions, emotions, and contents are expressed in a mixed way, that is the combination of several languages, visual and textual, and the related metadata, such as: geographical position and hashtag which they are labeled with.

The automatic treatment of emojis is faced through two main approaches: the processing of the textual descriptions of emojis³ (Fernández-Gavilanes 2018, Singh 2019) and the analysis of the emojis (co-)occurrences (Guibon 2018, Rakhmetullina 2018, Barbieri 2016, Novak et al., 2015)⁴.

The function of emojis is not limited to a predictable labeling of the emotional content (Felbo

2017, Tian 2017), but, it is possible to improve the score of sentiment analysis tools by knowing the meaning of emojis (LeCompte 2017, Felbo 2017, Guibon 2016, Novak 2015).

The content analysis of the images has been addressed from several perspectives and techniques.

Several studies are moving from a fully qualitative and manual approach (Tifentale & Manovich, 2015, Vitale et al., 2019, Palazzo et al., 2020, Esposito et al., 2020) to mixed methods involving algorithms and computer vision techniques combined with qualitative observations (Hochman 2015, Indaco & Manovich, 2016).

This work, starting from a well experimented innovative approach on previous studies (Vitale et al., 2020, Giordano et al., 2020), makes the choice to analyze the images in their textual translation, with a fully automatic analytical pipeline, designed in a semiotic point of view. Besides the semiotic interest to digital media date back to the early 2000s and continues to the present days, considering digital media a specific semiotic field (Cosenza 2014, Bianchi e Cosenza 2020).

Lastly, considering design, the visual representation of the social media data is increasing widespread as vehicle for knowledge of several fields (Ciuccarelli et al., 2014).

The research team doesn't provide an algorithm to analyze the images but adopt the automatic translation from the social media algorithm, designed to the visual impaired users by parsing the html code of the Instagram web interface. The metadata involved is the "accessibility_caption".

These are lists of words, hierarchically distributed, that let us to define and observe subject and attributes of the images, in addition to allowing the analysis of the entities.

³ Among the emoji resources, we mention; *Emojipedia* (emojipedia.org), *iEmoji* (www.iemoji.com); the annotated resources, such as *The Emoji Dictionary* (emojictionary.emoji.foundation.com); *EmojiNet* (emojinet.knoesis.org); the ones which are specific for Sentiment Analysis and Emotion Detection purposes, such as *Emoji Sentiment Ranking* (kt.ijs.si/data/Emoji_sentiment_ranking), *EmoTag* (abushoeb.github.io/emotag) and *The SentiStrength emoticon sentiment lexicon* (sentistrength.wlv.ac.uk); and the corpora, such as *ITAMoji* (Ronazano 2018) and *the Emojibot corpus* (Monti 2016).

⁴ Actually, the original meaning of emojis, specified in their descriptions, could be very different with the ones attributed by people into specific text occurrences (Fernández-Gavilanes 2018, Wood & Ruder 2016). Therefore, the manual annotation of emoji dictionaries could ignore important details that concerns usage dynamics over time (Ahanin 2020, Felbo 2017). Nevertheless, the representation of an emoji can vary widely across different communication platforms (Wagner, 2020) and their semantics can present culture or language specific usage patterns (Barbieri 2016). Thus, the results produced by the analyses of the emoji in large corpora could present some drawbacks as well.

2. Methodology

In this work, we propose the automatic treatment of the sentiment expressed into 7,482 Instagram posts.

All the information composing the dataset (i.e. captions, hashtags, emojis and images) are automatically put into relation with one another and visualized into an interactive dashboard. The phenomenon, can be observed through a system of filters, zooms and interdependent interactions. The result captures the topography of feelings, moods and needs expressed on the Instagram platform during the lockdown.

The NLP activities are performed in this research through the software NooJ⁵, which allows both the formalization of linguistic resources and the parsing of corpora. The dictionaries and grammars, which have been built *ad hoc* for this work, complement the open-source resources of the basic Italian and English modules of NooJ (Vietri 2014).

All the pictures published on May 4, May 18 and June 3, 2020 with the hashtag #andratut-tobene have been collected with a custom python script that simulate the human navigation. For each picture, we collected the entire source code of the web page in a JSON (JavaScript Object Notation) format.

This one has been parsed to a tabular one, in order to plan a format suitable for the adopted tools. The files have been refined selecting the endpoint useful for the analysis: captions (including hashtags); images hyperlink; accessibility captions; geographical coordinates and timestamp.

Some data required a data refinement phase. For the captions, it has been necessary to do a cleaning phase in which all the texts that were not

written Italian have been detected automatically by adopting the google translate API (Application Programming Interface) and removed. Moreover, from this field all the hashtags have been extracted, to allow their standalone analysis.

Accessibility captions have been clustered on two dimensions: “human or not human” and “indoor or outdoor”, previously defined thanks to a list of coherent words, subsequently matched by a pattern matching phase⁶. Geographical coordinates set the images on a specific point on the map, so it has been necessary to make a reverse geocoding procedure to find out region and province levels.⁷

Furthermore, Timestamp have been converted in a conventional date and time format.

After these steps, images and texts became ready to be analyzed through NLP procedures and mapped with geographical visualization techniques, observing them on the desired timeframe.

For the analysis of verbal features, we used *SentIta*, a semi-automatically built lexicon task (Pelosi 2015a), containing more than 15,000 lemmas, simple words and multiword units. Each entry is annotated with polarity and intensity scores, into a scale that ranges from -3 to +3. It must be applied to texts in conjunction with a network of almost 130 embedded local grammars, formalized in the shape of Finite State Automata (Pelosi 2015b), which systematically modulate the *prior polarity* of words according to their syntactic local context⁸. These resources can be directly applied to the Instagram captions, while hashtags need to be initially segmented. In this phase, they are analytically decomposed into their constituents through 10 morpho-syntactic grammars applied simultaneously, but with different priorities. In this way, the selection of the most probable sequences is decided for the upstream⁹.

⁵ <http://www.nooj-association.org/>

⁶ For instance, in the “human” cluster we have grouped all the accessibility caption containing words such as “people, man, woman, person” etc.

At the same time, in the “outdoor” cluster we have grouped all the pictures with words such as “sea, skyline, lawn, beach” and so on.

⁷ This phase has been possible in an automatic way adopting the python library reverse-geocoder (<https://github.com/thampiman/reverse-geocoder>)

⁸ The performances of our method produced satisfactory results in the sentence-level analysis of the textual part of the corpus: 0,85 Recall; 0,96 Precision and 0,9 F-score.

⁹ Basically, if the system produces more than one interpretation, the preferred one is the one in which the constituents have a longer length and the smallest number of constituents. In other words, the system firstly compares the whole normalized string with the word forms from *SentIta*, then continues the comparison with English and Italian word forms from the basic module. Hence, the dictionaries receive the higher priority and are applied before morphological grammars.

If the system does not match any word in the lexicon, it starts the structural analysis of the string, which consist of a systematic comparison of substrings with the all the words contained in the dictionaries, according to part of speech specific syntactic structures. Such structure, ordered here by priority assignments, can be

For the analysis of the non-verbal features, emojis are treated by using an electronic dictionary, which has been semi-automatically annotated with the same information used to analyze verbal features. We created this database with recognizable decimal codes in UTF-8 encoding from *Emojipedia*, then we carried out the automatic analysis of the textual descriptions of each emoji.

This dictionary has been used to locate and interpret the emojis occurring in the posts¹⁰.

After the clustering phase (human and not human; indoor and outdoor), all the findings of the sentiment on all the languages can be associated to the pictures' features, combined or not.¹¹

3. Visualization and Results

For a complete observation of the analysis' process and of its results, we developed a data visualization dashboard. In the following dashboard it is possible to observe the sentiment analysis on each language processed, with the chance of investigating the different trends during the days and the single hours day by day.

Adopting the clusters detected in the images, a system of filters let to focus the results basing on the subjects depicted.

On the left side of the dashboard, a map shows the geographical situation, merging the 4 sentiment values in a single one (weighted average) and coloring the regional shape on chromatic scale from the minimum value (-3) in orange, to the maximum value (+3) in blue. The same scale is applied to the line chart on the right, in which each line is related to the vertical axes and colored as described before.

$$\frac{(Sent_{Emoji} * P_{Emoji} + Sent_{Hashtag} * P_{Hashtag} + Sent_{Texts} * P_{Texts})}{P_{Emoji} + P_{Hashtag} + P_{Texts}}$$

Each score reached by the three languages are taken into account, namely texts, hashtags and emojis, are weighted according to the assumption that the euphoric level of emojis' sentiment is higher than hashtags' one, and both are higher than written texts' one in general. According to these results, we propose this weighted average formula, in which emojis, hashtags and texts have different weights (P), respectively 33, 50, and 100.

multiword expressions; free nominal, prepositional, adjectival and adverbial phrases; elementary sentences; and verbless sentences.

¹⁰ While the oriented words located into captions and hashtags respectively cover the 6% and the 9% of the

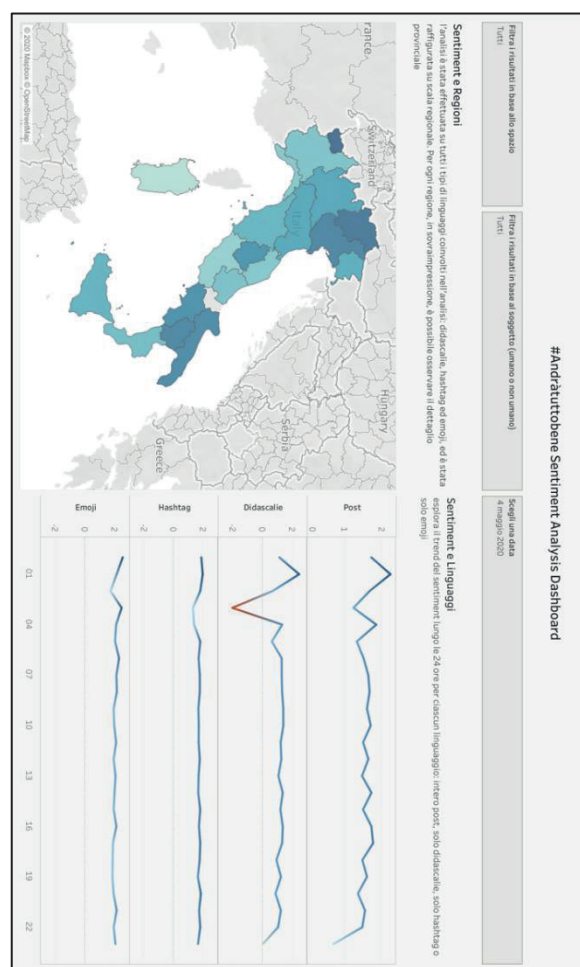


Figure 1 The sentiment analysis values

As a matter of fact, Novak (2015) underlined that it is more common the use of positive emojis with respect to the negative ones. Moreover, Boia et al. (2013) observed a poor correlation between the perceived emotional polarity of emojis and the accompanying linguistic text alone. Although it is actually challenging to predict the interaction between emoji and texts, there are cases in which the emojis express or reinforce the sentiment of the text with which they occur and cases in which they modify it or even express an opposite emotional state (Guibon 2018, Shoeb 2019).

Hashtags are conventionally used in two ways: on one hand, to describe the contents in a list of words, and on the other hand for strategic purposes, in order to place the images in useful thematic spaces. This is also the reason why we have removed from the analysis of all the Instagram-

full words contained in the posts, the sentiment labelled emojis cover the 19% of the total number of emojis in the corpus.

specific hashtags, such as: #likeforlike, #followforfollow etc., that are not suitable or even could be misleading or biased for our investigation. At the same time, the hashtags are also used as part of the messages, in substitution of words, so they deserve to be included in the final measure, but not with the same relevance of the captions.

The performances of emoji, hashtag and texts as indicators for sentiment analysis purposes, alone and combined with one another, have been tested on our corpus. We verified a significant improvement in terms of document-level precision when the indicators are considered together (0.98), if compared with the precision of texts (0.91), hashtag (0.81) and emoji (0.65) considered alone. The different precisions reached by the three languages considered alone empirically confirm the diversification of weights we proposed in our formula. This weighted measure has, then, been compared by three different judges¹² with the arithmetic mean on a sample of 100 Instagram posts from our corpus and performed better in the 92% of the cases.

Nevertheless, the geographical dimension is very important to observe the different kind of languages in the online community (Arnaboldi et al., 2017). Through an overlay function, moving the cursor on the map (figure 2), we show the geographical data in the deeper level of the single province, focusing on each region. The result represents the possible different polarity value between different provinces. For instance, on May 4 in the provinces of Oristano (Sardegna), Genova (Liguria) and Viterbo (Lazio), the sentiment value is negative, despite the positive average value of the region. However, the average sentiment value over the three days analyzed is found always positive, with different evidences on regional and provincial scale. Lastly, users can explore the results focusing on one or more region though a filter function (by clicking or selecting). All the filters are interdependent, so it is possible to select all the functions available investigating the phenomenon from all the possible perspectives.

Conclusion

Throughout the quantitative and qualitative analysis of the different expressive forms used on Instagram, this work proposes a general view of COVID-19 in Italy.

¹² For the evaluation of the three judges, we have calculated the intercoder reliability adopting the Krippen-

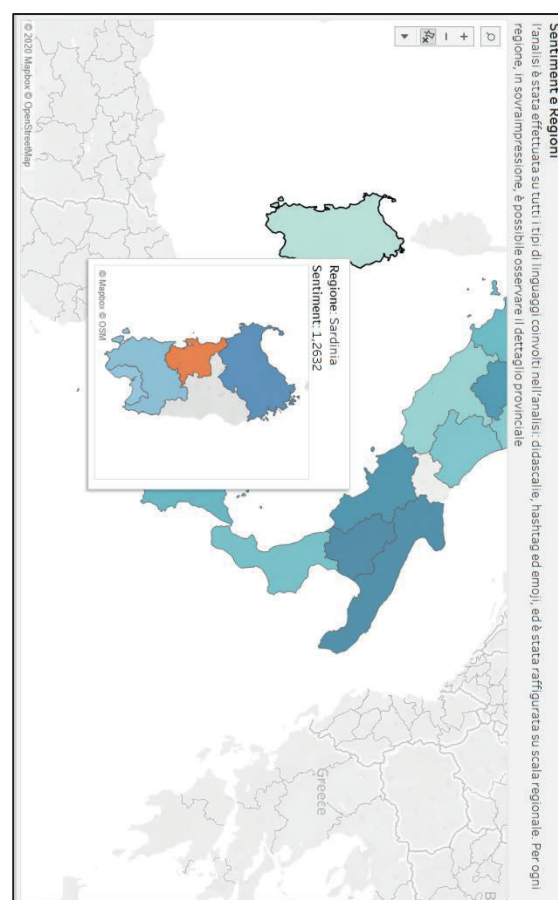


Figure 2 Overlay function: provincial scale

The research brought together linguistic analysis and design into a more general semiotic framework. The aim was, in fact, to put in shape the pandemic phenomenon through a selection of linguistic relevance.

The virus caused a series of unpredictable changes narrated on Instagram through the hashtag *#andràtuttobene*. A mantra for the Igers and an isotopy for the analysts (Greimas & Courtès 1979). Working on multiple levels, the research has offered a general and a local view of the emotions told during the lockdown period. Starting from a lexical base, made up by a list of words, and using electronic dictionaries also for the images, the analysis organized a large amount of data, developing a real map of emotions and needs expressed during the first wave of pandemic. The map can be visualized through a dashboard letting users observe general and local reactions, down to the single province. The emotional effects of sense have been evaluated thanks to a polar and unique measure.

dorff's Alpha formula (kalpha). The three coders selected have a kalpha of 0.9.

In the end, did everything really go well for Instagram's Italy? In general, it seems so. The average sentiment value over the three days analyzed is always positive, with variations on regional and provincial scale. Going down the single province, we can find differences, as the Sardinia, Lazio and Liguria cases.

References

- Agerri, R. and García-Serrano, A. (2010). Q-WordNet: Extracting polarity from WordNet senses. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2300–2305.
- Ahanin, Z., & Ismail, M. A. (2020). Feature extraction based on fuzzy clustering and emoji embeddings for emotion classification. *International Journal of Technology Management and Information System*, 2(1), 102-112.
- Arnaboldi, M., Brambilla, M., Cassottana, B., Ciuccarelli, P., & Vantini, S. (2017). Urbanscope: A lens to observe language mix in cities. *American Behavioral Scientist*, 61(7), 774-793.
- Balomenou, N., & Garrod, B. (2019). Photographs in tourism research: Prejudice, power, performance and participant-generated images. *Tourism Management*, 70, 201-217.
- Bansal, P., Bansal, R., & Varma, V. (2015). Towards deep semantic analysis of hashtags. In *European conference on information retrieval* (pp. 453-464). Springer, Cham.
- Barbieri, F., Ronzano, F., & Saggion, H. (2016). What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*; 2016 May 23-28; Portorož, Slovenia.
- Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 100-107).
- Bianchi, C., & Cosenza, G. (2020). LexiaRivista di semiotica, 33-34. *Semiotica e Digital marketing*, Roma, Aracne.
- Bolioli, A., Salamino, F., & Porzionato, V. (2013). Social Media Monitoring in Real Life with Blogmeter Platform. *ESSEM@ AI* IA*, 1096, 156-163.
- Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G. F., Sanguinetti, M., & Sulis, E. (2014). Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicità. In *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Es³Lod 2014* (pp. 56-63). European Language Resources Association.
- Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE intelligent systems*, 28(2), 55-63.
- Boia M., Faltings B., Musat C. C., & Pu P. (2013). A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets. In *“2013 International Conference on Social Computing”*, pp. 345-350. IEEE.
- Castellucci, G., Croce, D., & Basili, R. (2016). A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 38-45).
- Celebi, A., & Özgür, A. (2016). Segmenting hashtags using automatically created training data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2981-2985).
- Celebi, A., & Özgür, A. (2018). Segmenting hashtags and analyzing their grammatical structure. *Journal of the Association for Information Science and Technology*, 69(5), 675-686.
- Cosenza, G., (2014). *Introduzione alla semiotica dei nuovi media*, Laterza, Milano.
- Di Gennaro, P., Rossi, A., & Tamburini, F. (2014). The FICLIT+ CS@ UniBO System at the EVALITA 2014 Sentiment Polarity Classification Task. In *Proceedings of the Fourth International Workshop EVALITA 2014*.
- Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.
- Esposito, F., Falco, M., & Vitale, P. (2020). Experiencing Museums A Qualitative and Quantitative Description About Igers' Narration of an Exhibit Space. In *Workshops of the International Conference on Advanced Information Networking and Applications* (pp. 1011-1018). Springer, Cham.
- Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.
- Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castaño, F. J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, 103, 74-91.
- Giordano, G., Primerano, I., & Vitale, P. (2020). A Network-Based Indicator of Travelers Performativity on Instagram. *Social Indicators Research*, 1-19.

- Greimas, A. J., Courtés, J. (1979). *Sémiotique: dictionnaire raisonné de la théorie du langage*, Paris, Hachette.
- Guibon, G., Ochs, M., & Bellot, P. (2018). From emoji usage to categorical emoji prediction. In 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2018). Springer Lecture Notes in Computer Science, Switzerland.
- Hochman, N. (2015). *The social media image: Modes of visual ordering on social media* (Doctoral dissertation, University of Pittsburgh).
- Indaco, A., & Manovich, L. (2016). Urban social media inequality: definition, measurements, and application. arXiv preprint arXiv:1607.01845.
- LeCompte, T., & Chen, J. (2017). Sentiment analysis of tweets including emoji data. In 2017 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 793-798). IEEE.
- Monti J., Sangati F., Chiusaroli F., Benjamin M., and Mansour S.. (2016). *Emojitalianobot and emojiworldbot - new online tools and digital environments for translation into emoji*. In Proc. CLiC-it 2016, volume 1749 of CEUR Workshop Proceedings.
- Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS one*, 10(12).
- Palazzo, M., Vollero, A., Vitale, P., & Siano, A. Urban and rural destinations on Instagram: Exploring the influencers' role in# sustainable tourism. *Land Use Policy*, 100, 104915.
- Pelosi S., (2015a): SentIta and Doxa: Italian Databases and Tools for Sentiment Analysis Purposes. The second Italian Conference on Computational Linguistics (CLiC-it 2015). Trento, December 3-4 2015. Book of Proceedings. Accademia University Press srl, Torino.
- Pelosi, S. (2015b). A Lexicon-based Approach to Sentiment Analysis: The Italian Module for Nooj. *Formalising Natural Languages with Nooj 2014*, 37.
- Rathnayake, C., & Ntalla, I. (2020). "Visual Affluence" in Social Photography: Applicability of Image Segmentation as a Visually Oriented Approach to Study Instagram Hashtags. *Social Media+ Society*, 6(2), 2056305120924758.
- Reuter, J., Pereira-Martins, J., & Kalita, J. (2016). Segmenting twitter hashtags. *Intl. J. on Natural Lang. Computing*, 5(4).
- Ronzano, F., Barbieri, F., Wahyu Pamungkas, E., Patti, V., & Chiusaroli, F. (2018). Overview of the evalita 2018 italian emoji prediction (itamoji) task. In 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018 (Vol. 2263, pp. 1-9). CEUR-WS.
- Simeon, C., Hamilton, H. J., & Hilderman, R. J. (2016). Word segmentation algorithms with lexical resources for hashtag classification. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 743-751). IEEE.
- Singh, A., Blanco, E., & Jin, W. (2019). Incorporating emoji descriptions improves tweet classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 2096-2101).
- Singhal, A., & Rattine-Flaherty, E. (2006). Pencils and photos as tools of communicative research and praxis: Analyzing Minga Perú's quest for social justice in the Amazon. *International Communication Gazette*, 68(4), 313-330.
- Srinivasan, S., Bhattacharya, S., & Chakraborty, R. (2012). Segmenting web-domains and hashtags using length specific models. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1113-1122).
- Stranisci, M., Bosco, C., Patti, V., & HERNANDEZ FARIAS, D. I. (2015). Analyzing and annotating for sentiment analysis the socio-political debate on# labuonascuola. In second Italian Conference on Computational Linguistics (pp. 274-279). aAccademia University Press.
- Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook sentiment: Reactions and emojis. In Proceedings of the Fifth International
- Tifentale, A., & Manovich, L. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In *Postdigital aesthetics* (pp. 109-122). Palgrave Macmillan, London.
- Vitale, P., Mancuso, A., & Falco, M. (2019). Museums' tales: visualizing instagram users' experience. In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 234-245). Springer, Cham.
- Vitale, P., Palazzo, M., Vollero, A., Siano, A., & Foroudi, P. (2020). The Role of Igers in the Territorial Dynamics of Sustainable Tourism-Oriented Destinations. In International Symposium: New Metropolitan Perspectives (pp. 759-767). Springer, Cham.

Polarity Imbalance in Lexicon-based Sentiment Analysis

Marco Vassallo¹, Giuliano Gabrieli¹, Valerio Basile², Cristina Bosco²

1. CREA Research Centre for Agricultural Policies and Bio-economy, Italy

2. Dipartimento di Informatica, Università degli Studi di Torino, Italy

{marco.vassallo|giuliano.gabrieli}@crea.gov.it, {valerio.basile|cristina.bosco}@unito.it

Abstract

Polarity imbalance is an asymmetric situation that occurs while using parametric threshold values in lexicon-based Sentiment-Analysis (SA). The variation across the thresholds may have an opposite impact on the prediction of negative and positive polarity. We hypothesize that this may be due to asymmetries in the data or in the lexicon, or both. We carry out therefore experiments for evaluating the effect of lexicon and of the topics addressed in the data. Our experiments are based on a weighted version of the Italian linguistic resource MAL (Morphologically-inflected Affective Lexicon) by using as weighting corpus TWITA, a large-scale corpus of messages from Twitter in Italian. The novel Weighted-MAL (W-MAL), presented for the first time in this paper, achieved better polarity classification results especially for negative tweets, along with alleviating the aforementioned polarity imbalance.

Italiano. *Lo sbilanciamento della polarità è una situazione di asimmetria che si viene a creare quando si impiegano valori soglia parametrici nella Sentiment Analysis (SA) basata su dizionario. La variazione dei valori soglia può avere un impatto opposto rispetto alla predizione di polarità negativa e positiva. Si ipotizza che questo effetto sia dovuto ad asimmetrie nei dati o nel dizionario, o in entrambi. Abbiamo condotto esperimenti per misurare l'effetto del lessico e degli argomenti trattati nel nostro dataset. I nostri esperimenti sono basati su una versione ponderata della risorsa per l'italiano MAL (Morphologically-inflected Affective Lexi-*

con), usando come corpus per la ponderazione TWITA, un corpus di larga scala di messaggi da Twitter in italiano. La nuova risorsa Weighted-MAL (W-MAL), presentata per la prima volta in questo articolo, ottiene migliori risultati nella classificazione della polarità specialmente, per i messaggi negativi, oltre ad alleviare il problema sopracitato di sbilanciamento della polarità.

1 Introduction and Motivation

Sentiment Analysis (SA) is the task of Natural Language Processing that aims at extracting opinions from natural language expressions, e.g., reviews or social media posts. The basic approaches to SA typically fall into one of two categories: dictionary-based and supervised machine learning. Methods based on a dictionary make use of *affective* lexicons, language resources where each word or lemma is associated to a score indicating its affective valence (e.g., *polarity*). In SA they are faster than supervised statistical approaches and require minimal adaptation, unless the resource is domain-specific, also when applied to multiple environments with minimal adaptation overhead. However, they only achieve good performance for identifying coarse opinion tendencies in large datasets, since they cannot take into account the impact of the context on the polarity value associated to a word.

Supervised statistical methods, on the other hand, tend to provide better quality predictions across benchmarks, due to their better ability to generalize over individual words and expressions, and learning higher level features. These models also show a better ability to adapt to specific domains,

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

provided the availability of data suitable for training.

In order to access the lexical entries in an affective dictionary, lemmatization must be performed on each single word. Unfortunately, lemmatization is an error-prone process, with potentially negative impact on the performance of downstream tasks such as SA. Vassallo et al. (2019) introduced a novel computational linguistic resource, namely the Morphologically-inflected Affective Lexicon (henceforth MAL) in order to address this issue by avoiding the lemmatization step in favor of a morphologically rich affective resource.

In the experiments we carried out on a specific text genre, namely social media, we have observed that using a threshold to assign polarity classes is beneficial, and using the MAL instead of a lemmatization step improves the SA performance overall, in particular due to a better prediction of the negative polarity. However, the variation in threshold has opposite impact on the prediction of negative and positive tweets.

In this paper, we investigate the motivation beyond this polarity imbalance. In particular, we speculate that this may be due to asymmetries in the data (e.g., different internal topics), in the lexicon (e.g., different amounts of negative and positive terms), or both, and we provide experiments to better understand this result and validate these hypotheses. We can therefore summarize as follows our research questions:

- Is the polarity imbalance due to the topic addressed?
- Is the polarity imbalance due to the lexicon (i.e., the resources we used, Sentix and MAL)?
- Is the polarity imbalance due to both?

A further contribution of the paper consists in providing a statistical method for finding the threshold for using the lexicon in SA tasks.

The paper is organized as follows. In the next section, affective lexicons and the resource MAL are discussed. In section 3, we describe the issues related to polarity imbalance in lexicon-based approaches for SA. The fourth section is instead devoted to discuss the impact on SA of lexicon and to introduce W-MAL. Section 5 discusses how the topics addressed in the text may impact on SA.

The final section provides conclusive remarks and some hints about future work.

2 Affective Lexicons

SA is typically cast as a text classification task, very often approached by supervised statistical models among the NLP research community (Barbieri et al., 2016). However, there are several scenarios where dictionary-based methods are preferred, including large-scale industry-ready systems, and domain-specific applications. While generally less accurate than supervised classification, dictionary-based methods tend to be robust to the classification of sentiment across different domains, faster and with a higher level of scalability.

For the Italian language, several sentiment dictionaries, or, using a more general term, *affective lexicons*, were published with different levels of granularity of the annotation and availability to the public, as summarized on the website of the Italian Association of Computational Linguistics¹.

Sentix (Basile and Nissim, 2013) is one of the first affective lexicons created for Italian language, with a first release described in (Basile and Nissim, 2013), and a second release called Sentix 2.0². It provides an automatic alignment between SentiWordNet, an automatically-built polarity lexicon for English by Baccianella et al. (2010), and the Italian portion of MultiWordNet (Pianta et al., 2002). While the first version of Sentix associated two independent positive and negative polarity scores to each word, in Sentix 2.0³ all the senses of each lemma have been collapsed into one entry by means of a weighted average, where the weights are proportional to sense frequencies computed on the sense-annotated corpus SemCor (Langone et al., 2004). Moreover, the positive and negative polarity scores have been combined to form a single polarity score ranging from -1 (totally negative) to 1 (totally positive). Sentix 2.0 includes 41,800 different lemmas.

In order to use a lemma-based affective lexicon such as Sentix, lemmatization is a necessary step to undertake. In our previous work, we found that such intermediate step causes a considerable amount of noise, in the form of lemmatization er-

¹<http://www.ai-lc.it/en/affective-lexica-and-other-resources-for-italian/>

²<https://github.com/valeribasile/sentixR>

³<https://github.com/valeribasile/sentixR>

Table 1: A tweet with the output of the three lemmatization models where the lemmas are alphabetically ordered and the errors marked in bold.

Original	@ANBINazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgricolturaIT
ISDT	acqua adigire allarme crisi d dopo idrico po - Sentix score: 0.080
POSTWITA	acqua adigere allarme crisi di dopo idrico po - Sentix score: 0.080
PARTUT	acquare adigere allarme crisi d dopo idrico po - Sentix score: -0.078

rors such as the ones shown in Table 1 (Vassallo et al., 2019). We therefore built a new resource on top of Sentix, described in the next section.

2.1 MAL

We proposed the Morphologically-inflected Affective Lexicon in Vassallo et al. (2019, MAL). It is an extension of Sentix where the entries associated to polarity scores rather than lemmas are the inflected forms related to each lemma, and the polarity scores to be associated to each form are drawn from the original lemmas in Sentix. The approach consists in linking the lexical items found in tweets with the entries of Sentix 2.0, without the application of an explicit lemmatization step. The lexicon is indeed expanded by considering all the acceptable forms of its lemmas extracted from the Morph-It collection of Italian forms (Zanchetta and Baroni, 2005). Each form takes the same polarity score of the original lemma, but when different lemmas can assume the same form, the arithmetic mean of their polarity scores is assigned. The MAL comprises 148,867 forms and all the items linked to the lemmas of Sentix 2.0.

Using the MAL we performed a series of experiments on the impact of lemmatization on dictionary-based SA, which showed how the reduction in lemmatization errors leads to a better polarity classification performance.

3 Polarity Imbalance in Lexicon-based Sentiment Analysis

When using an affective lexicon to predict the polarity of natural language sentences, a threshold must be fixed to translate the numerical scores into discrete classes, e.g., positive, neutral, and negative. In Vassallo et al. (2019), we showed how the variation of such threshold has different, opposite impacts on the accuracy of the classification, using as a benchmark the corpus annotated with sentiment polarity made available by the SENTiment POLarity Classification (SENTIPOLC) shared task at EVALITA 2016. More precisely, the red dotted lines with label ALL in

Figure 1 show that the F1 score of the classification of positive polarity instances increases with stricter thresholds, while the F1 score of negative polarity instances decreases.

We postulate two non-mutually exclusive hypotheses on the origin of the polarity imbalance, namely the effect of lexicon and topic. The affective scores in the lexicon may be biased towards one end of the polarity spectrum due to a number of causes, resulting in skewed classification results. On the other hand, some topics tend to attract opinions more polarized towards one end of the spectrum than the other (e.g., “war” is an inherently negative topic), therefore the classification might be influenced by this intrinsic polarization.

4 The Effect of Lexicon on SA

In order to shed some light on the polarity imbalance due to lexicon we applied a weighted approach to MAL by developing the Weighted Morphologically-inflected Affective Lexicon (WMAL). It originates from the intuition that less frequent terms should have a higher impact on the computation of the polarity of the sentence where they occur. This principle stems from the observation that more sought-after terms are often used to convey stronger opinions and feelings.

We therefore computed the relative frequency of every item in MAL by using TWITA, a large-scale corpus of messages from Twitter in the Italian language (Basile et al., 2018). TWITA is indeed large (covering over 500 million tweets from 2012 to 2018, and the collection is currently ongoing) and domain-agnostic enough to provide a sufficiently representative sample of the distribution of the Italian language words, although specific to one social media platform.

Despite its size, not all the terms from the MAL occur in TWITA: 57.9% of the 148,867 terms occurring in MAL were found in TWITA, due to the sparseness of particular inflected forms, and to the presence of multi-word expressions in the lexicon (18,661, about 12%) that were not considered for

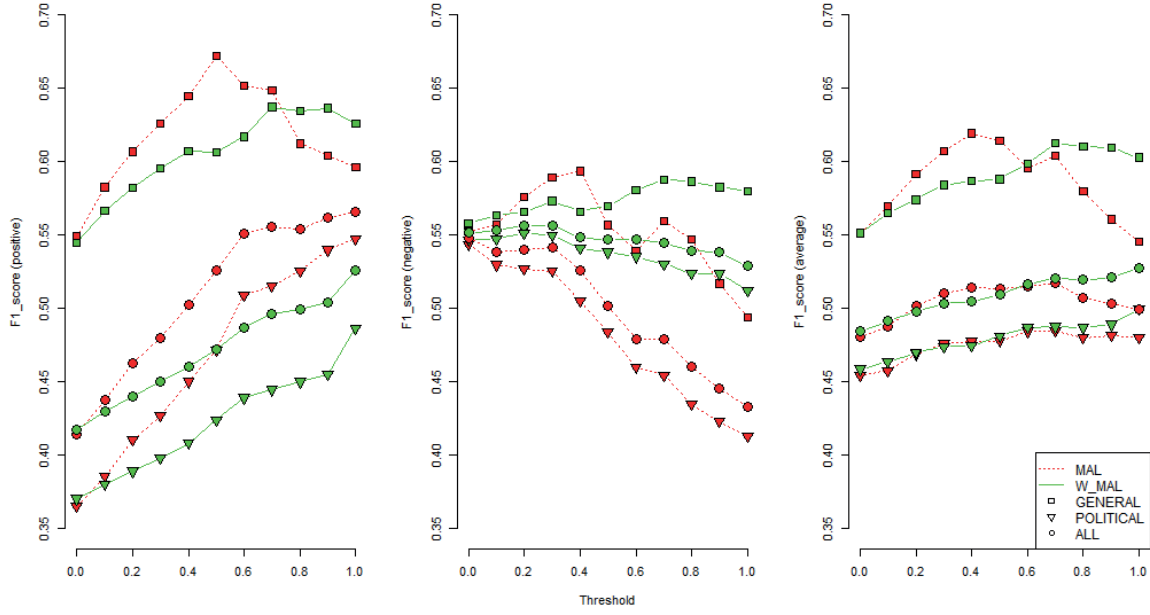


Figure 1: Results of the polarity classification on SENTIPOLC. The threshold value on the X-axis is applied to transform the sum of the scores from the lexicon into a positive or negative label.

matching the resources. For comparison, 73,36% of Sentix lemmas were found in TWITA.

Accordingly, the scores of MAL were recalculated by weighting them with the associated words frequency in TWITA, using the Zipf scale measure (van Heuven et al., 2014). We decided to use this measure because of its easy understanding and the short computation timing. Actually, the Zipf scale measure is a logarithmic scale based on the well-known Zipf law of word frequency distribution (Zipf, 1949). The computation of Zipf values of terms frequencies from TWITA is straightforward and essentially equals to the logarithm of the absolute frequency scaled down by a multiplicative factor:

$$Zipf(i) = \log_{10} \left(\frac{f(i)}{\frac{\sum_{i=1}^N f(i)}{10^6} + \frac{N}{10^6}} \right) + 3$$

where N is the number of tokens in TWITA (6,644,867), $f(i)$ is the absolute frequency of the i -th token in TWITA, and the sum of the token frequencies $\sum_{i=1}^N f(i) = 6,906,070,053$, therefore:

$$Zipf(i) = \log_{10} \left(\frac{f(i)}{6,906.07 + 6.644} \right) + 3$$

The original Zipf scale is a continuous scale and it ranges from 1 (very low frequency) to 6 (very high frequency) or even 7 (e.g., for very frequent words like auxiliar verbs). By computing the Zipf score of the MAL terms on TWITA, we found some terms with very low frequencies, resulting in negative values because of the logarithmic function. These were re-coded with the minimum Zipf value. The resulting weights in the W-MAL range from a minimum of -5.16 to a maximum of 5.95 (the original MAL ranged from -1 to 1). Eventually, we decided to keep the terms that were not found in TWITA in the W-MAL with their MAL original score.

We initially applied the Zipf scale to MAL polarity scores by simply multiplying the two found scores and thus giving more weight to high frequent terms. However, using the affective lexicon with such weighting scheme resulted in a decrease in its polarity classification performance. We therefore simply reversed the Zipf scale by weighting the original scores inversely with respect to their words frequency. By doing so, we tested for our speculation of giving more weight to low frequent terms. We replicated the polarity detection experiment on SENTIPOLC. The results, shown in the green solid lines in Figure 1 labeled ALL, indicate a better performance over-

all, and a reduced imbalance between the positive (F1-scores standard deviation across the thresholds of 0.035 with W-MAL vs 0.054 with MAL) and (especially) the negative polarity class (F1-scores standard deviation across the thresholds of 0.008 with W-MAL vs 0.042 with MAL).

To further clarify the effect found on the polarity scores, we show two example tweets in Figure 2⁴. In the figure, the MAL and W-MAL scores are included for the highlighted words, along with the total polarity scores computed with both dictionaries, showing how the final judgment can change from neutral to polarized (bottom example) or switch polarity entirely (top example). In particular in the top example the scores are associated with "confondesse" (to confuse in subjunctive mood) and to "diritto" (right), while in the bottom example the scores are associated with "Istituto" (school) and to the periphrastic verbal form "viene taciuto" (is silenced). This result confirms our speculation that negative polarity is expressed with more specific words than positive polarity. Psychology studies also show that more complex forms of language were used for expressing criticisms rather than positive evaluations (Stewart, 2015).

We also notice how the F1-score on the negative polarity is generally higher than the one on the positive polarity class. This means that the negative polarity of tweets is better predicted than the positive polarity by means of the weighted process with the inverse coding. This outcome seems to be substantially supported also by the W-MAL directly proportional performance that worked worse than the inverse version in terms of prediction. This trend was also observed across most of the results of the SENTIPOLC shared task, mostly based on supervised models with lexical features, further indicating that the vocabulary of negative sentiments is richer than that of positive sentiment.

⁴The translation of the examples is as follows. For the top example: *They would be #thegoodschool if meritocracy were not confused with "doormatocracy": the one whereby even a right becomes a concession.* For the bottom example: *@steGiannini #thegoodschool In the rankings of the School there are also TFA qualified teachers with 48 months of service. Why is it silenced?* where steGiannini refers to the Italian minister for school

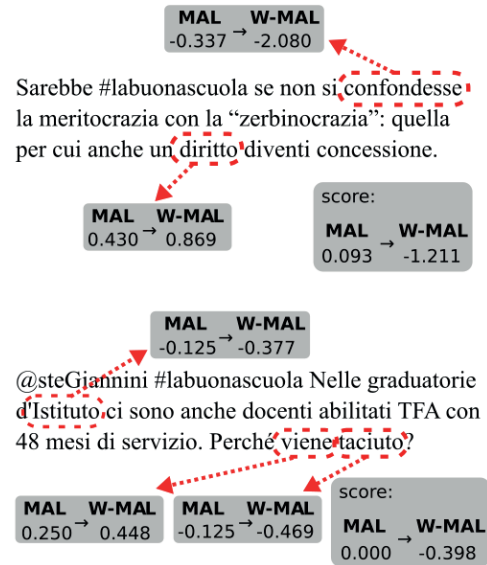


Figure 2: A comparison between the scores calculated for polarized words of a tweet according to MAL and W-MAL in two tweets from the test set.

5 The Effect of Topic on Sentiment Analysis

In order to investigate the interaction between the imbalance of dictionary-based polarity classification and a possible asymmetry in the data (i.e. different internal topics), we performed such classification with MAL and W-MAL with the reversed Zipf scale on a benchmark with explicitly stated topics. As a matter of fact, the test set of SENTIPOLC is composed of 1,982 Italian tweets, organized in 496 *general* i.e. domain-independent tweets, and 1,486 *political* tweets, obtained by filtering data with specific keywords related to political Italian figures. The results of our experiment are also included in Figure 1 with the GENERAL and POLITICAL labels.

The first observation we draw from this experiment is that the polarity imbalance is a phenomenon restricted to the topic-specific section of the dataset. This confirms the hypothesis that dictionary-based polarity classification is affected by the imbalance issue with the extent to which its topic is specific. In particular, we hypothesize that some topics (such as politics) tend to attract opinions more polarized towards one end of the spectrum (the negative one in this case), therefore inducing the observed imbalance.

The second observation is that weighting the polarity scores in the dictionary based on word frequency (W-MAL) provides better overall results.

In particular, the F1 scores are better in the topic-specific case, specifically due to a better prediction of the negative polarity. This result reinforces the idea that a polarized topic induces polarity imbalance, and therefore a method to alleviate such imbalance (i.e., a weighting scheme) leads to better performance. In our view, a reason for this effect is that topic-specific messages make use of less frequent words on average.

6 Conclusion and Future Work

The weighting scheme proposed in this work is a promising solution to the polarity imbalance in dictionary-based SA. The experiments show that weighting the polarity scores with word frequencies yielded a more precise prediction of the polarized tweets, with lessened bias in the thresholds for neutral scores. The novel resource here presented, W-MAL, is an attempt to better characterize the most sought-after words, which have an impact on the interaction between sentiment and topic. We believe it also represents a promising attempt to control for context-dependency while using lexicon-based methods for SA.

In particular, with this resource we try to give voice to the linguistic intuition that the exploitation of a specific form within a message might meaningfully impact on the sentiment expressed in the message. For instance, referring to the top example in figure 2, by exploiting the subjunctive mood "confondesse" of the verb "confondere" (to confuse), the author joins together with the meaning of the verb also a sense of doubtfulness and of unreality. This is also improved by the fact that this form introduces a clause which is coordinated with the clause headed by a verb in conditional mood, i.e. "sarebbe" (form of to be). This form of the verb "confondere" seems especially adequate for contexts where a negative polarity is expressed and less appropriate for other cases. The use of this specific mood for the verb has therefore a meaningful impact on the sentiment expressed. The MAL properly encodes this information, which may be lost when a lemmatization step is applied on text and all forms are subsequently considered as bearing the same meaning without further nuances. But the W-MAL does also better: it encodes the probabilistic information about how suitable a form is for expressing a particular sentiment with respect to other available forms in a given context.

For all the aforementioned reasons, this work has drawn our attention to the necessity of weighting the dictionary-based affective lexicons to SA with corpora-based word frequencies. The resource is freely available at <https://github.com/valeribasile/sentixR/blob/master/sentix/inst/extdata/W-MAL.tsv>

In future work, we plan on working on more refined weighting strategies, e.g., leveraging the frequency information of word forms in addition to lemmas, and taking the topic distribution into consideration. Reducing the computation load is a challenging goal as well (see Prakash et al. (2015)). On the other hand, modern transformer-based models have reached state-of-the-art results on the task of polarity detection (Polignano et al., 2019), although they are far more expensive and time-consuming to run. We plan therefore to compare the predictions of these systems, and study ways to integrate their respective strengths (i.e., speed and transparency of the dictionary-based approach vs. the superior prediction capability of the deep neural models) in order to boost the overall performance.

The present work was originally conceived in the framework of the AGRItrend project led by the CREA Research Centre for Agricultural Policies and Bio-economy, aiming at collecting and analyzing social media data for opinions in the domain of public policies and agriculture. As such, we plan on studying the impact of the techniques presented in this paper on that particular domain, and observe if the same, or different, patterns emerge. On a similar line, so far we conducted experiments on data from Twitter, which facilitates access to large quantity of data but restricts the range of text style and genre found in them.

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).
- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTiment

- POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR-WS.org.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.
- Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. CEUR-WS.org.
- Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 63–69. Association for Computational Linguistics (ACL).
- Emanuele Pianta, Luisa Bentivogli, and Christian Giardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR-WS.org.
- Saurabh Prakash, T. Chakravarthy, and E. Kaveri. 2015. Statistically weighted reviews to enhance sentiment classification. *Karbala International Journal of Modern Science*, 1:26–31.
- Martyn Stewart. 2015. The language of praise and criticism in a student evaluation survey. *Studies In Educational Evaluation*, 45:1–9.
- Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67:6:1176–1190.
- Marco Vassallo, Giuliano Gabrieli, Valerio Basile, and Cristina Bosco. 2019. The tenuousness of lemmatization in lexicon-based sentiment analysis. In *Proceedings of the Sixth Italian Conference on Computational Linguistics - CLiC-it 2019*. Academia University Press.
- Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).
- George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley.

Suoidne-varra-bleahkka-mála-bihkka-senet-dielku **‘hay-blood-ink-paint-tar-mustard-stain’ – Should compounds be lexicalized in NLP?**

Linda Wiechetek

linda.wiechetek@uit.no

Chiara Argese

chiara.argese@uit.no

Tommi A Pirinen

tommi.pirinen@uit.no

Trond Trosterud

trond.trosterud@uit.no

Divvun & Giellatekno, UiT Norgga árkálaš universitehta

Abstract

English.

Lexicalizing compounds, in addition to treating them dynamically, is a key element in giving us idiomatic translations and detecting compound errors. We present and evaluate an e-dictionary (*NDS*) and a grammar checker (*GramDivvun*) for North Sámi. We achieve a coverage of 98% for *NDS*-queries and of 96% for compound error detection in *GramDivvun*.

Italiano.

*La lessicalizzazione delle parole composte, in aggiunta a trattarle in maniera dinamica, è un elemento chiave per ottenere traduzioni idiomatiche e rilevare errori nelle stesse. Presentiamo e valutiamo un e-dizionario (*NDS*) e un correttore grammaticale (*GramDivvun*) per il Sami del Nord. Otteniamo una copertura del 98% per le ricerche in *NDS* e del 96% per il rilevamento di errori nelle parole composte in *GramDivvun*.*

1 Introduction

In this paper¹, we discuss the use and necessity of the lexicalization of compounds – in addition to the dynamic approach to compounding – in two rule-based Natural Language Processing (NLP) applications, a grammar checker *GramDivvun* and an electronic dictionary *NDS* (short for *Neahtadigisáni*). We argue for a dual approach and support this view with an evaluation of these tools. For comparison, we also look at a third application, a corpus tool (*Korp*) for the North Sámi corpus *SIKOR*. *SIKOR*, the Sámi International KORpus, is the collection of texts in different Sámi languages compiled by UiT The Arctic University of Norway and the Norwegian Sámi Parliament.

In the past, we have mostly focussed on the dynamic approach to morphological analysis. This means that we have a lexicon with lemmata and stems, which in a finite-state manner are combined

with inflectional and derivational affixes and other stems and modified when morpho-phonological processes apply. In this way the linguistic processes inflection, derivation and compounding are modelled in a dynamic way, i.e. by means of concatenation and composition as opposed to listing of all forms. Lexicalization, i.e. listing compounds or inflected word forms as such, is the alternative approach to the dynamic one. In addition to these two approaches we also use guessers for certain tasks, i.e. proper name guessing in morpho-syntactic parsing. Our approach is entirely rule-based and open source. Within our 20 year experience with language tools for the Sámi languages and other languages with complex morphology, we have achieved good results and produced reliable tools.

There are a number of approaches to error detection of a few errortypes for morphologically complex - although less complex than North Sámi - languages like Latvian (Deksne, 2019) and Russian (Rozovskaya and Roth, 2019). The Latvian neural network grammar checker focusses on preposition-postposition confusion, adjective-noun agreement, mood errors in verb forms, number and case in noun forms, definiteness of adjectives and missing commata. All of these error types have a good performance with precisions between 78% and 98.5%. Judging from their regular expressions to insert artificial errors, most of their error types seem to be fairly local errors that can be resolved based on bigrams.

The Russian system focusses on more advanced error types - case, number agreement, gender agreement, preposition and aspect. However, the results show that the system is still in its initial phase with low precision and recall for most error types (precision is between 22% and 56%, only gender agreement reaches 68%, and recall is significantly lower, between 9% and 36%). None of these approaches deals with compound error de-

¹Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tection.

For neural network approaches, large corpora with error mark-up are necessary, which are not available for North Sámi. The error marked-up corpus contains 120 459 words, and when looking at specific error types – as in this case compound errors – the corpus is even smaller. The Russian system is based on an error-marked corpus of 200k words (deemed too small by its authors), the Latvian system works with artificial errors, an approach that can be problematic as it does not reflect real text errors.

In compounding, two or several words are combined to form a new word. In Sámi, Finnic and Germanic languages, compounding is a productive process and new compounds like in (1) can be made on the fly.² In Romance languages, these compounds typically correspond to prepositional constructions (ital. ‘la federa del cuscino del divano’).³

- (1) soffá|guoddá|olgoža (North Sámi)
sofa|pute|trekk (Norwegian)
‘sofa pillow cover (English)’

The initial motivation for extensive lexicalization of compounds of North Sámi goes back to adapting the spellchecker to users’ needs, i.e. avoiding false alarms in *Ávvir* newspaper’s texts.

North Sámi is a Uralic language spoken in Norway, Sweden and Finland by approximately 25 700 speakers (Simons and Fennig, 2018). It is a synthetic language, where the open parts of speech (PoS) – nouns, adjectives, etc. – inflect for case, person and number. The grammatical categories are expressed by a combination of suffixes and stem-internal processes affecting root vowels and consonants alike, making it perhaps the most fusional of all Uralic languages. In addition to compounding, inflection and derivation are common morphological processes in North Sámi.

North Sámi has seven morpho-syntactic cases, i.e. nominative (Nom.), genitive (Gen.), accusative (Acc.), illative (Ill.), locative (Loc.), comitative (Com.), and essive (Ess.). Case plays a more central role in Sámi than in preposition-based case languages, since here syntactic functions are identified based on case only. In addition, nouns can bear possessive suffixes. Verbs are inflected

²To avoid confusion with hyphenated compounds, “[|]” is used to mark word boundaries in compounds

³Although there are a number of real compounds in Italian, such as *fruttivendolo*, as well.

for person, number (singular, dual, plural), tense (present and past tense) and mood (indicative, conditional, and potential). Derivational processes (passive, causative, inchoative, diminutive, reflexive, to name only some of them) enhance the combinatory possibilities of each verb.

Table 1 illustrates that compounding in North Sámi is by no means restricted to noun noun combinations, but includes a number of other *parts-of-speech* (PoS) as well, also as heads.⁴

Type	Example	Gloss and translation
N N	láhka rievdadusat	law change.PL ‘law changes’
A.Attr N	boahhte áigi	coming time ‘future’
Adv N	dáppe olmmoš	here person ‘person from here’
Pron A	iešgudet lágan	each alike ‘different kinds of’
Pron N	eanet lohku	more number ‘majority’
Adv	dušše fal	only really ‘just’
Pcle		
Adv V	vuostái váldojuvvo	against take.PASS.3SG ‘received’
PrfPrc N	mearridan fápmu	decide.PRFPRC power ‘authority’
Num	okta nuppe lohkái	one second ten.ILL ‘eleven’
Num		
Num N	1978 -láhka	1978 -law ‘1978 law’
Num A	3 -ivnnat	3 -colored ‘3-colored’
Num A	golmma ivnnat	three colored ‘three colored’

Table 1: Compound types according to PoS; ‘|’ is used to mark word boundaries

In North Sámi, compounds are formed without a hyphen, except for those involving a proper noun, a digit, or an acronym like *Davvi-Norgii* ‘Northern Norway (Ill.)’, *3-juvllatsykkell* ‘tricycle’, and *ILO-álgoálbmotsoahpamuš* ‘ILO-indigenous people agreement’ (Riektačállinrávvagat, 2015, p.46). There are a number of multiwords where a space is obligatory (*albma ládje* ‘properly’ and *duollet dálle* ‘sometimes’). Also genitive first compounds have an alternative interpretation when written apart, which makes error detection more difficult.

2 Background

The North Sámi tools described in this article – *NDS*, *Korp* for *SIKOR* and *Gram-Divvun* (Wiechetek, 2012) – all rely on the *Giel-*

⁴The following abbreviations are used: N=noun, V=verb, A=adjective, Attr=attributive, Adv=adverb, Pron=pronoun, Pcle=particle, PrfPrc=past participle, Num=numeral, Prop=propernoun.

laLT infrastructure (Moshagen et al., 2013), a technological framework for managing lexical data and building it into language technology applications including e-dictionaries and grammar checkers. All of them make use of a morphological analyzer, an *FST* (Finite-State Transducer) described in Pirinen (2014), where word formation processes are moduled. Additionally, *SIKOR* and *Gram-Divvun* include a Constraint Grammar-based syntactic analysis. The full modular structure of the latter is described in Wiecheteck (2019b).

The computational modeling of the language is done using finite-state morphology (Beesley and Karttunen, 2003). The method of recognizing grammatical words as well as querying their grammatical information is based on looking up the words in an *FST* that contains the morphological dictionary of the language. There are two types of compounds in the language model: the ones that are stored in the lexicon as lexicalized units and the ones generated dynamically using a compounding model. Table 2 gives the statistics over the length of lexicalized compounds.⁵

Lexicalized four-element compounds are quite common in the noun lexicon, e.g. *davvisámegiel-terminologiija* ‘North Sámi language terminology’. Even six-element compounds (*sáivačáhce-guolleuostáiváldindilli* ‘fresh water fish receive situation’) can be found.

The different types of North Sámi compounds in Table 1 are not treated equally in the morphological analyzer. Only the compounds in the first two lines can be derived dynamically. All others need to be lexicalized, i.e. listed in the lexicon, to receive a compound analysis. Numeral compounding is not treated dynamically in the *FST*. The dynamic compounds are generated from the dictionary by concatenating word forms (such as a genitive or nominative noun followed by other noun) and adding a compound tag +*Cmp*. The main dynamic compounds are (derived and non-derived) noun + noun pairs. One feature of the underlying technology is that the compounding mechanism is capable of modeling infinitely long compounds: for example nouns of any magnitude are compounds and modeled by the finite-state automaton. Since the compounding mechanism of an *FST* is very powerful, it also leads to ambiguity. When we allow arbitrary lexemes to combine to form compounds,

⁵The table is based on the dictionary size at the time of the writing (September 2020); it is actively developed daily. Further abbreviations are *Adp*=adposition, *Conj*=conjunction.

some will overlap other existing lexemes, cf. ex. (2).

- (2) Davvi **regiuvdna**
North region;direction.oven
‘The northern region’

Here, *regiuvdna* ‘region’ has a typical spelling error, o>u. The *FST* analyzes it as a misspelling of *regiovdna* ‘region’, but also as a compound with the elements *regi*, a common wrong form of *regiija* ‘direction’, and *uvdna* ‘oven’. While this example has only two possible analyses, twenty or more different analyses are not uncommon.

Roots PoS	2	3	4	5	6+
N	16 603	1 048	1 665	86	15
Num	408	1 048	42	0	4
Prop	11 680	3 005	115	9	1
A	3 854	333	13	0	0
V	478	4	0	0	9
Adv	896	109	1	0	0
Adp	152	49	0	0	0
Conj	3	0	0	0	0

Table 2: Lexical compounds in the lexicon by the PoS of their head and the number of their roots

3 Compounds in three NLP applications

We present three applications, an e-dictionary, a corpus tool, and a grammar checker tool.

3.1 An e-dictionary (NDS)

The North Sámi – Norwegian dictionary contains 25 000 lemmata and uses an *FST*. The e-dictionary was first implemented in 2013 with no use of relational databases (all linguistic resources are contained within static files and external command-line tools) (Ryan Johnson, 2013). It is an intelligent dictionary in the sense that is able to look up North Sámi word forms and find lemmas via the *FST*. It also allows a tolerant mode, which accepts the letters *acdntz* for *áčđņšťž* in addition to their usual values. The e-dictionary can split compounds to provide the user with its elements as well as the whole compound if a translation is available. The lexicalization of compounds is important since the translation of the compound cannot necessarily be derived from the translation of its parts (Antonsen, 2018, p.54).

In the FST 90% of the 100 000 nouns, and in the dictionary 75% of the 25 000 nouns are compounds.

3.2 A corpus tool

The web application and corpus search tool *Korp* (Borin et al., 2012) does not show the internal structure of compounds in *SIKOR*. Neither lexicalized, nor dynamic compounds are searchable as either the lexicalized analysis is picked instead of the dynamic one or – in the case of compounds that are not listed in the lexicon – a lexicalized compound is made by the preprocessor. This is a problem inherent in the implementation of the tool. However, when searching for the compound tag used in the FST (+Cmp), there are 94 658 results. The reason for that is that the first element in split compounds in coordination receives a specific compound tag (+Cmp/SplitR) as well.

Table 3 shows the statistics for compounds in *SIKOR*.⁶ The results are obtained using the scripts that can be found in *GiellaLT*.⁷ According to our analyses 8.6% of the tokens in corpus are compounds, and 86% are lexicalized. The rest is mainly composed of 2-elements compounds (13.4%) and a very small part of 4-7 elements (0.5%).

Many of the longer compounds in *SIKOR* are quite creative and are hyphenated as the one in ex. (3).

- (3) **suoidne-varra-bleahkka-mála-bihkka-senet-dielku**
 hay-blood-ink-paint-tar-mustard-stain
 mu báiddis lei dušše lihkohisvuohta.
 my shirt.LOC was only mishap
 ‘The hay-blood-ink-paint-tar-mustard-stain on my
 shirt was only a mishap.’

Parts PoS	2	3	4	5	6/7
N	96.2	98.9	89.2	80	66.7
Prop	3.8	1.1	10.8	20	33.3

Table 3: Compound types in *SIKOR* by the PoS of their head and the number of their root (amounts given in percentage)

The current public version of the Sámi corpus *SIKOR* (SIKOR, 2018) (in *Korp*) consists of 32.2 million words. It was analyzed with a preprocessor

⁶The search was done on 2020-09-07.

⁷<https://github.com/giellalt/conf-clitic2021>

that does not distinguish between lexicalized and dynamic compounds. The (non-public) version of *SIKOR* used in this article makes this distinction, though, as will future versions in *Korp*.

A search for compound tags only returns split compounds, i.e. the first coordinated hyphenated nominal element, cf. in ex. (4), i.e. *riddo-* ‘coast-’.

- (4) **riddo-** ja vuotnaguovlluin
 coast- and fjordregion.LOC.PL
 ‘in coastal and fjord regions’

GiellaLT has already produced a solution, i.e. a tag for cohorts with a dynamic compound (<with-dynamic-compound>) added by a Constraint Grammar module. However, this tag does not provide any information about the number of elements and the beginning and ending of each element.

3.3 A grammar checker (GramDivvun)

GramDivvun, the North Sámi grammar checker (Wiechete et al., 2019b) takes input from the FST to a number of other modules, the core of which are several Constraint Grammar modules. Constraint Grammar is a rule-based formalism for writing disambiguation and syntactic annotation grammars (Karlsson, 1990; Karlsson et al., 1995). In our work, we use the free open source implementation VISLCG-3 (Bick and Didriksen, 2015). All components are compiled and built using the *GiellaLT* infrastructure (Moshagen et al., 2013).

Lexicalization of compounds is relevant for grammar checking within compound error detection. One common error that cannot be resolved by a spellchecker is the spelling of compounds as two or more words. *GramDivvun* performs this type of error detection as part of the tokenization. The tokenization is done in two steps. In the first step potential compounds are tokenized ambiguously (either as one or as two words, the first of which is accompanied by an errortag). In the second step, a Constraint Grammar module⁸ selects or removes the error reading. Two conditions need to be met to find the compound error: 1. the compound needs to be lexicalized, and 2. the syntactic context needs to support the compound reading.

The syntactic context is specified in handwritten Constraint Grammar rules. The

⁸<https://github.com/giellalt/lang-sme/blob/3a43911929458fd39da309ed23178bf5dbd04bcd/tools/tokenisers/mwe-dis.cg3>

REMOVE-rule below removes the compound error reading (identified by the tag Err/SpaceCmp) if the head is a 3rd person singular verb (cf. 1.2) and the first element of the potential compound is a noun in nominative case (cf. 1.3). The context condition further specifies that there should be a finite verb (VFIN) somewhere in the sentence (cf. 1.4) for the rule to apply.

```

1 REMOVE (Err/SpaceCmp)
2 (0/0 (V Sg3))
3 (0/1 (N Sg Nom))
4 (*0 VFIN);

```

All possible compounds written apart are considered to be errors by default, unless the lexicon specifies a two or several word compound or a syntactic rule removes the error reading. There are numerous syntactic contexts where the potential parts of compounds make perfectly sense. In the case of noun-noun compounds, the second element can for example be a simple adverbial, as in ex. (5). The second element can be homonymous with another PoS, it can be a finite verb or an infinitive.

- (5) son lea boarráseamus **mánná joavkkus**.
s/he is oldest child group.LOC
's/he is the oldest child in the group.'

4 Evaluation

We evaluate the e-dictionary (coverage) and the grammar checker (precision, recall) for compounding (errors). The corpus search tool does not exhibit compounding information and is therefore not evaluated.

4.1 An e-dictionary (NDS)

We analyzed the logs for NDS (*Neahtdigisánit*) for 2019, and found that 12.6% of the types in the user queries are compounds. The results are obtained using the scripts that can be found in *Giel-laLT*⁷. The amount of lexicalized compounds in the logs (72.1%) is approximately the same as in the dictionary, where it is 75% (cf. Section 3.1 above). As much as 98% of the compound queries get a translation, either a lexicalized one or of its parts. Thus dynamic compounding contributes with a substantial improvement to dictionary coverage. If the alternatives are “getting no help from the dictionary” and “getting help to translate the parts” then the latter is to be preferred, even though the correct translation would be different from just

joining the parts. For example, the compound word *ruhtahearrá* ‘rich man’ is not lexicalized in NDS but it does get a translation of its parts *ruhta* ‘money’ and *hearrá* ‘man’, which can help the user to understand the meaning of the compound word itself.

Most of the non lexicalized compounds are composed of 2 elements (96% in the logs and 93% in the entries). When analyzing the entries in the dictionary, we found that 24.8% are compounds and of those 97.6% are lexicalized. Table 4 shows PoS for compounds in NDS logs and entries.

Parts PoS	Logs				Entries			
	L	2	3	4	L	2	3	4
N	90	87	85	100	86	87	82	0
A	3	0	0	0	2	0	0	0
Prop	3	0	0	0	12	4	0	0
V	2	13	14	0	0	8	18	0
Adv	1	0	0	0	0	0	0	0

Table 4: Compounds according to the number of their parts and PoS in NDS logs and entries (L=lexicalized)

4.2 A grammar checker (GramDivvun)

We evaluate error detection for syntactic compound errors (i.e. words that are written apart and should be a compound) in *GramDivvun* in two ways. Firstly, we compare last year’s results in Wiechetek (2019a) with a newer version of *GramDivvun*, from now on referred to as the *Nodalida*-corpus. Last year’s results are based on version *r183544* (Wiechetek et al., 2019a)⁹. The new results are based on version *r28510*¹⁰ of *GramDivvun*.

However, as the focus in the last analysis was a different one, i.e. we evaluated other error types as well, we ran a second evaluation on a 2 363 word-corpus¹¹ specifically made to test compound error detection, i.e. every sentence contains a potential compound. These sentences are hand-selected from *SIKOR*.

The results of the evaluation are presented in Table 5. We can see that precision has gone significantly up, i.e. the average precision is 95.5%.

⁹<https://github.com/giellalt/lang-sme/releases/tag/nodalida-2018> on 2019-09-26

¹⁰<https://github.com/giellalt/lang-sme/releases/tag/clcit> on 2020-09-07

¹¹http://gtsvn.uit.no/freecorpus/orig/sme/odda_mahppa/compounds.correct.txt

However, the recall has gone down to average 46%. We are investigating the reasons for that. But in general, a high precision is desirable in grammar checking, even at the cost of a lower recall.

The results of the evaluation of *GramDivvun* compound grammar checking are shown in Table 5.

Measure	(2019)	(2020)	
	Nodalida corpus		Compound corpus
Precision	75.0%	93.1%	98.0%
Recall	72.9%	43.2%	48.5%
F1-Score	73.9	59.0	64.9
TP	51	54	50
FP	17	4	1
FN	19	67	53

Table 5: Measures for GramDivvun (TP/FP= true/false positives, FN=false negatives)

False negatives are typically due to the lack of lexicalization. Many of those are proper noun combinations which are very productive, e.g. *Murmánska-aviisa* ‘Murmansk newspaper’, *Várggát-festiválas* ‘at the Várggát festival’, *km-galba* ‘km sign’ and *Divttasvuotna-regiovnna* ‘Divttasvuotna region’.

Other reasons are certain (unlikely) analyses of especially the first element, e.g. that generally suggest a syntactic construction rather than a compound as in ex. (6). Here the first element *duorastat* ‘Thursday’ has a finite verb reading as well.

- (6) dán **duorastat** **veaiggi**.
 this.GEN Thursday twilight.GEN
 ‘this Thursday evening’

The false positive is due to an error in the recognition of the span of the target. In ex. (7), *lulli sámí guvlui* is concatenated, but it should only be *lulli sámí*.

- (7) dohko **lulli** **sámí** guvlui.
 thither South Sámi area.ILL
 ‘thither towards the South Sámi area.’

5 Conclusion

We have shown that the lexicalization of compounds – in addition to their dynamic treatment – is useful and necessary for two language applications for North Sámi, an e-dictionary (*NDS*) and a grammar checker (*GramDivvun*). The evaluation of *NDS* shows that we get a good coverage: 98%

of the compounds logged do get a translation and 72% are lexicalized in the FST. The evaluation of *GramDivvun* has shown that we manage to identify compound errors with a precision of 98% and a recall of 49% utilising a combination of information from the lexicon and syntax.

We conclude that there are perfectly good reasons for lexicalizing compounds, i.e. providing idiomatic translations for when it cannot be derived from the parts, and to support compound grammar checking. At the same time, lexicalization can dissimulate word formation information in corpus tools. This can be resolved and we have already implemented a solution in Constraint Grammar to make the information available in a future version of the corpus tool. As dynamic compounding is limited to few PoS at the moment, in the future we want to investigate and model compounding of other PoS (in the FST). Also experiments with neural network approaches and a comparison of the results to our rule-based grammar checker could be an interesting future project.

Acknowledgments

Thank you to Thomas Omma for doing the error corpus mark-up and for fun linguistic discussions, and to Lene Antonsen for digging in our corpus and helping to find just the right example.

References

- Lene Antonsen. 2018. *Sámegielaide modelleren – huk-sen ja heiveheapmi duohta giellamáilbmái*. [Modeling Saami languages. Construction and adaptation to real-world linguistic issues]. Ph.D. thesis, UiT The Arctic University of Norway, Tromsø.
- Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.
- Eckhard Bick and Tino Didriksen. 2015. CG-3 – beyond classical Constraint Grammar. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)*, pages 31–39. Linköping University Electronic Press, Linköpings universitet.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA).

- Daiga Dekšne. 2019. Bidirectional lstm tagger for latvian grammatical error detection. In *Ekšteins K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, vol 11697*. Springer.
- Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
- Fred Karlsson. 1990. Constraint Grammar as a Framework for Parsing Running Text. In Hans Karlgren, editor, *Proceedings of the 13th Conference on Computational Linguistics (COLING 1990)*, volume 3, pages 168–173, Helsinki, Finland. Association for Computational Linguistics.
- Sjur N. Moshagen, Tommi A. Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *NODALIDA*.
- Tommi A. Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404, CICLing 2014*, pages 519–532, Berlin, Heidelberg. Springer-Verlag.
- Riektačállinrávvagat. 2015. Riektačállinrávvagat. Sámedikki giellaossodat/Sámedikki oahpahusossodat, Guovdageaidnu.
- Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of russian. In *Transactions of the Association for Computational Linguistics, vol. 7, pp. 1–17, 2019*.
- Trond Trosterud Ryan Johnson, Lene Antonsen. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)*, Proceedings Series 16: 59–71.
- SIKOR. 2018. SIKOR uit norgga árkálaš universitehta ja norgga sámedikki sámi teakstačoakkáldat, veršuvdna 06.11.2018. <http://gtweb.uit.no/korp>. Accessed: 2018-11-06.
- Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-first edition.
- Linda Wiecheteck, Kevin Brubeck Unhammer, and Sjur Nørstebø Moshagen. 2019a. Seeing more than whitespace – Tokenisation and disambiguation in a North Sámi grammar checker. In *Proceedings of the third Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 46–55.
- Linda Wiecheteck, Sjur Nørstebø Moshagen, Børre Gaup, and Thomas Omma. 2019b. Many shades of grammar checking – launching a constraint grammar tool for north sámi. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar - Methods, Tools and Applications*, NEALT Proceedings Series 33:8, pages 35–44.
- Linda Wiecheteck. 2012. Constraint Grammar based correction of grammatical errors for North Sámi. In G. De Pauw, G-M de Schryver, M.L. Forcada, K. Sarasola, F.M. Tyers, and P.W. Wagacha, editors, *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)*, pages 35–40, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams

Mehmet Can Yavuz

Faculty of Engineering and Natural Science, Sabancı University, Tuzla
İstanbul, Türkiye
mehmetyavuz@sabanciuniv.edu

Abstract

In theatrical pieces, written language is the primary medium for establishing antagonisms. As one of the most important figures of renaissance, Shakespeare wrote characters which express themselves clearly. Thus, the emotional landscape of the plays can be revealed from the texts. It is important to analyze such landscapes for further demonstrating these structures.

We use word-emotion association lexicon with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). By using this lexicon, the emotional state of each character is represented in 10 dimensional space and mapped onto a plane. This principle axes planes position each character relatively. Additionally, tempora-emotional evaluation of each play is graphed.

We conclude that the protagonist and the antagonist have different emotional states from the rest and these two emotionally oppose each other. Temporal-Emotional timeline of the plays are meaningful to have a better insight into the tragedies.

1 Introduction

Shakespeare's plays are one of the most important works of early modernity with their dramaturgy, strong and in-depth characters and poems, that are all still contemporary. Antagonistically most powerful works of Shakespeare, who wrote in three genres, are tragedies. Tragedies have strong antagonisms and written language is the primary

medium for establishing antagonisms. Therefore, the characters in these plays express themselves through dialogues or monologues clearly. By using the emotional association of each word, from this perspective, it is possible to reveal the emotional landscape of the plays. Emotionally positioning the characters relative to each other is important for further understanding of the structure of the plays. It is also important to extract overall emotional variations throughout the play to have an insight about the tragedies. In this study, it is aimed to answer these two questions by using word-emotion association lexicon with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

These tasks are not only important for artificial literature (Lebrun, 2017; Yavuz, 2020), but also for the purpose of increasing artistic creativity through computerized analysis. The algorithmic study of literary works has given rise to objective criticism in literary theory, (Moretti, 2013). Accordingly, the discovery of new or previously theoretically laid out features of literary texts through mathematical experiments or statistics brought along conjunctions on literature, (Moretti, 2000; Yavuz, 2020). Art and criticism develop in parallel today as of yesterday. Computational linguistics enriched with fields such as advanced chatbots, conversational AI, or text style transfers, the fields give clues that artificial literature will also develop rapidly in the upcoming period. Whether it is a computer-assisted writing process or fully automated *writing machines*, these generative models need evaluation metrics. For this purpose, objective metrics should be developed for outstanding literary pieces such as Shakespeare, and for what makes these plays so great, if possible... Aim of this paper is in accordance with the previous works, (Yavuz, 2019; Yavuz, 2020), we would like to discuss objective

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and quantifiable reasons behind the literariness of a drama.

In order to convey such analyses, each character in any play represented in 10 dimensional emotional space and mapped onto a plane. The emotional timeline of the plays are also revealed. We would like to overview the word-emotion association lexicon and dimension reduction algorithm. Then, the last section is left to discussions on the emotional states of the characters and temporal analyses of the emotions.

1.1 Related Works

Distant reading is a established approach/methodology in digital humanities (DH) and mostly deals with the quantitative analyses of literary and cultural studies (Clement, 2008; Crane, 2006). Inside DH, "Drametrics" is a sub-research field, specialized on quantitative analysis of the literary genre of drama (Romanska, 2015). Digital Shakespeare research and projects have gotten attention since the 2000s (Hirsch, 2014; Mueller, 2008). The dramatic structures in the form of antagonisms are revealed by topic modeling algorithms, (Yavuz, 2019). Exceptional characters, such as Deus-ex-Machina, is also detected in semantic space, (Yavuz, 2020). The graphs are also used to extract secondary antagonisms, (Yavuz, 2020). Machine learning based text analyses are also carried out for genre classifications (Yavuz, 2019; Ardanuy, 2014; Hope, 2010; Schöch, 2016; Underwood, 2013; Yu, 2008). In literature, structural elements such as *dramatis persone* are also analyzed and applications are developed for further analyses (Dennerlein, 2015; Krautter, 2018; Schmidt, 2019, Trilcke, 2015; Wilhelm, 2013). In addition to dramatic structure works, there is literature that apply sentiment analyses for dramatic works, (Nalisnick, 2013; Schmidt, 2018; Schmidt, 2018a; Schmidt, 2018b).

2 Methodology

Our approach is lexicon based. Lines uttered by each character treated as a document and represented with a tf-idf. Emotional weights are multiplied with the vectors and the summed up to have the final 10 dimensional emotion space and then reduced to a plane by linear dimension reduction. Overall emotional state of any character, thus, is represented in relative point to each other. Lexicon is also used to extract temporal emotional varia-

tions along the texts.

2.1 EmoLex Unigrams

Psychologists proposed many theories for classifying human emotions, (Dalglish, 2000). Some emotions are considered basic, while others are considered as complex. The distinction can be between emotions that we can sense and perceive (instinctual), and emotions that can arrive after some thinking and reasoning (cognitive), (Zajonc, 1984). There are also oppositions, (Lazarus, 1984). According to Plutchik, (Plutchik, 1985), the discussion may not be resolvable because there is no empirical basis. There is a high correlation between basic and instinctive emotions, similarly between complex and cognitive emotions. Many of the basic emotions are also instinctual.

There are theoretical studies on what the basic emotions are. Ekman argues about the existence of 6 basic emotions: joy, sadness, anger, fear, disgust, and surprise, (Ekman, 1992). Plutchik includes two new emotions: trust and anticipation, (Plutchik, 1994). Plutchik displayed the emotions on a wheel. The distance from the center in the circle indicates intensity. Plutchik also state the basic emotions as opposing pairs: joy-sadness, anger-fear, trust-disgust, and anticipation-surprise. The opposing places and neighborhoods on the circle are formed accordingly.

Emolex is the dictionary of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and

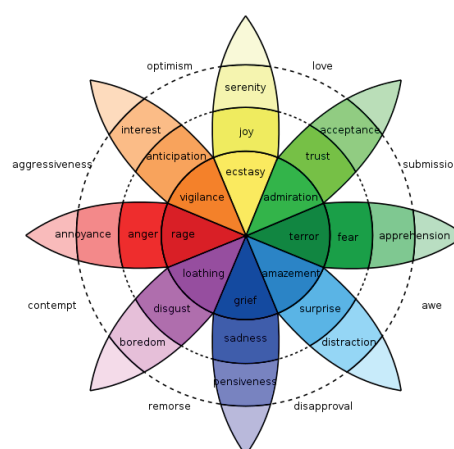


Figure 1: Eight basic emotions (*anger, fear, anticipation, trust, surprise, sadness, joy, and disgust*) on Plutchik's Wheel of Emotions

positive), (Mohammad, 2013). Thus, representing each word in 10 dimensional emotion space is effective as well as theoretically relevant. The dataset is unigram. Each single word represented in 10 dimensional feature space: 8 basic emotions and 2 sentiment labeling. Each emotion can be either 0 or 1. No intensity information is given. The words are crowd-sourced and manually labeled with Mechanical Turk. There are 14182 unigrams in the dictionary.

2.2 SVD (Golub, 1970)

Linear dimension reduction methods are for getting a n-dimensional plane over the hyperspace. For instance, if you have data cloud in 10 dimension, by mapping onto a plane, one can visualize such points. SVD is one of the relevant methods, it breaks any A matrix into three,

$$A = USV' \text{ which} \quad (1)$$

$$UU' = I \text{ and } VV' = I \quad (2)$$

S is a diagonal matrix that consists of r singular values. r is the rank of A. Truncated SVD is a reduced rank approximation. Only the most relevant dimensions are selected, these are the largest singular values. The dimensions of truncated SVD are $[uxk] * [kxk] * [kxv]$ Therefore A matrix is approximated by k dimensions, this is the dimension reduction. A descriptive subset of the data is called T, which is a dense summary of the matrix A,

$$T = US_k \quad (3)$$

S_k denotes k largest singular values, which is the number of reduced features. Each feature is represented with a percentage of variance. Higher variance means more information gain.

3 Discussion

There are two analyzes we would like to discuss. The first is about whether we can get ideas about the plays based on the emotional state of the characters. For this purpose, each character is treated as a text and represented with Tf-Idf features. The weights of each word for a specific character multiplied with its 10-dimensional emotion vector. Right afterwards, SVD lets you extract principle axes and mapped onto the plane. This is the first analysis to be discussed. Secondly, the temporal dimensions of the plays are considered.

The opposing emotional pairs (given by Plutchik) were represented as a time series as positive / negative. The states are represented as a cumulative temporal sum and the emotional landscape of the tragedies are revealed.

3.1 Analyses of Character Emotions

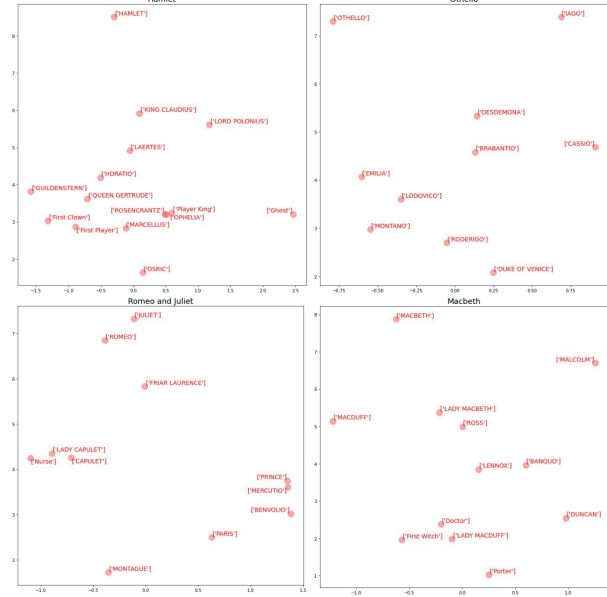


Figure 2: Emotional position of each character is represented, Hamlet, Othello, Romeo and Juliet, Macbeth, respectively.

The weights of Tf-Idf features help to position character emotions. Each word contributes to the resulting emotion with their weights. Weights of more frequent terms affect the resulting emotional state more, while less frequent terms affect less. The 10-dimensional emotional space mapped onto an abstraction plane by linear dimension reduction. The dimensions of these planes correspond to abstract emotions or a mixture of the other 10 dimensions with certain proportions. The important thing in these graphics is the position of the characters relative to each other. The emotional positioning of Hamlet in the upper left, Othello in the upper right, Romeo and Juliet in the lower left, and Macbeth in the lower right.

In these graphics, the basic characteristics of the plays can be observed. The main characters or pairs of characters are emotionally different from the rest. The protagonist and the antagonist always have emotional contrast. For example, the Hamlet play is basically determined by the tension between the two people, Hamlet and King

Claudius. Although Hamlet is emotionally very different from other characters, King Claudius is emotionally close to the main character cluster. In all tragedies, there is a cluster of emotionally indifferent characters, we can call *the main cluster*. Characters like Lord Polonius and Laertes are also located around King Claudius with the main cluster. The Ghost character, like Hamlet, is different from all other characters and is in an opposite position to Hamlet. These observations follow the readings of the play. In Othello, Iago sets traps to harm Desdemona. Desdemona is also compatible with the main cluster. But Iago and Othello are positioned far apart and apart from the main cluster. In the Macbeth play, the two enemies, Macduff and Malcolm, are opposite and separately positioned. Lady Macbeth is emotionally compatible with the main cluster. In an interesting observation on Romeo and Juliet, the positioning of the clusters is placed in symmetry in accordance with Renaissance thought. It is known that the play is written symmetrically. There are three family positions in symmetry: Ruling house of Verona, House of Capulet, House of Montague.

The graphs show that the emotional positioning of tragedies is compatible with the readings of the play. What we mean is the protagonists and antagonists are clearly observable. Distances or orientations, or rather relative positions, are significant. The main characters that experience basic tensions could be demonstrated. In the play of Romeo and Juliet, the affinities are observed and there is symmetrical positioning of the families.

3.2 Temporal-Emotional Evaluation of the Tragedies

Temporal-emotional evaluation of the tragedies are drawn. We can assume that each play is a temporal series, and we summed up the emotional state in each timestep, the cumulative emotional curves are calculated. Therefore, the emotional directions are determined. Emotions are positioned in contrast as (Negative-Positive), (Fear-Anger), (Anticipation-Surprise), (Sadness-Joy), (Disgust-Trust). For each timestep, or the word, the contribution is the expected emotion,

$$E[e] = e * p(e) \quad (4)$$

which $p(e)$ is the occurrence probability of the emotion in the lexicon and e is the Bernoulli random variable, either 0 or 1, either has the emotion

or not. The cumulative sum of a emotional contrast pair,

$$C(\{e_1, e_2\}, T) = \sum_{t=0}^T (E[e_2, t] - E[e_1, t]) \quad (5)$$

which $\{e_1, e_2\}$ are the random variables for emotional contrast pairs. The cumulative total for each pair is specified for 5 curves. Cumulative sums of (Negative-Positive), (Sadness-Joy), (Disgust-Trust) pairs for all four plays constantly increases. (Fear-Anger), (Anticipation-Surprise) are more neutral. All in all emotionally, the temporal word distributions for tragedies are similar.

4 Conclusion

The traces of the tensions between characters are observable from the emotional aspect. As we show, the emotional positions of the protagonist, the antagonist and the main cluster gives much insight about the greatness such pieces. Any great tragedy needs emotional contrast between the main characters and there is always *the main cluster*. The temporal-emotional characteristics of the plays are also important and very much similar to each other. There are constantly increasing emotions as well as neutrals. Each play grows towards positive, joyful and trusty emotional state. This might be the reason behind a followable play. Positive feelings should accumulate.

Either it is computer-assisted or fully automated *writing machine*, artificial literature needs emotional aspect. The emotional aspect of literary works should be conditional of such generative models. The common acceptance on this early

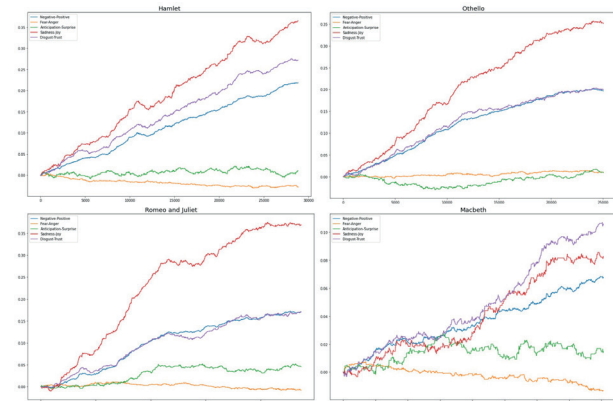


Figure 3: Emotional landscape of each character is represented, Hamlet, Othello, Romeo and Juliet, Macbeth, respectively.

modern author is his greatness as a tragedy writer. The theatrical pieces by Shakespeare are in dialog form, each character express themselves clearly. Therefore, it is shown that any dramatic antagonism is also emotional. Any artificial dramatic work should have a similar emotional resonance with such tragedies. With this analyses, we try to further develop evaluation metrics for artificial literature. A baseline metric to emotionally evaluate such theatrical forms.

On the way to Artificial Literature (ALit), there needs more criteria and more complex tools to analyze literariness of such pieces (literariness, 2020). As we shown so far, emotion aspect of the plays are very crucial at establishing antagonisms.

References

- Ardanuy, M. C., & Sporleder, C. (2014, April). Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 31-39).
- Clement, T., Steger, S., Unsworth, J. and Uszkalo, K. (2008). *How not to read a million books*. Available online at <http://people.brandeis.edu/unsworth/hownot2read.html>
- Crane, G. (2006). *What do you do with a million books?* D-Lib Magazine. Available online at <http://www.dlib.org/dlib/march06/crane/03crane.html>
- Dalgleish, T., Power, M. (Eds.). (2000). *Handbook of cognition and emotion*. John Wiley Sons.
- Dennerlein, K. (2015). Measuring the average population densities of plays. A case study of Andreas Gryphius, Christian Weise and Gotthold Ephraim Lessing. *Semicerchio. Rivista di poesia comparata* LIII: 80–88.
- Ekman, P. (1992). An argument for basic emotions. *Cognition emotion*, 6(3-4), 169-200.
- Golub, G. H.; Reinsch, C. (1970). "Singular value decomposition and least squares solutions". *Numerische Mathematik*. 14 (5): 403–420. doi:10.1007/BF02163027. MR 1553974.
- Gao, J., Galley, M., Li, L. (2018, June). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval* (pp. 1371-1374).
- Hirsch, B., & Craig, H. (2014). "Mingled Yarn": The State of Computing in Shakespeare 2.0. In T. Bishop, & A. Huang (Eds.), *The Shakespearean International Yearbook* (Vol. 14: Special Section, Digital Shakespeares, pp. 3-35). United Kingdom: Ashgate Publishing Limited.
- Hope, J., & Witmore, M. (2010). The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. *Shakespeare Quarterly*, 61(3), 357-390. Retrieved from <http://www.jstor.org/stable/40985589>
- Krautter, B. (2018). Quantitative microanalysis? Different methods of digital drama analysis in comparison. *Book of Abstracts, DH 2018*. Mexico-City, Mexico, pp. 225-228.
- Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, 39(2), 124–129. <https://doi.org/10.1037/0003-066X.39.2.124>
- Le, Q., Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).
- Lebrun, T. (2017, May). Who Is the Artificial Author?. In *Canadian Conference on Artificial Intelligence* (pp. 411-415). Springer, Cham.
- literariness. Oxford Reference. Retrieved 8 Jun. 2020, from <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100108912>.
- Mohammad, S. M., Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Moretti, F. (2000). Conjectures on world literature. *New left review*, 54-68.
- Mueller, Martin. (2008). Digital Shakespeare, or towards a literary informatics. *Shakespeare*. 4. 284-301. 10.1080/17450910802295179.
- Nalisnick, E. T., Baird, H. S. (2013, August). Character-to-character sentiment analysis in Shakespeare's plays. In *extitProceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 479-483).
- Plutchik, R. (1985) On emotion: The chicken-and-egg problem revisited. *Motiv Emot* 9, 197–200 . <https://doi.org/10.1007/BF00991576>
- Plutchik, R. (1994) *The psychology and biology of emotion*. HarperCollins College Publishers.
- Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.
- Romanska, M. (2015). Drametrics: what dramaturgs should learn from mathematicians. In Romanska, M. (ed.), *The Routledge Companion to Dramaturgy*. Routledge, pp. 472-481.
- Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. (2019). Katharsis – A Tool for Computational Drametrics. In *Book of Abstracts, DH 2019*.

- Schmidt, T., Burghardt, M. (2018, August). An evaluation of lexicon-based sentiment analysis techniques for the plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 139-149).
- Schmidt, T. and Burghardt, M. (2018b). Toward a Tool for Sentiment Analysis for German Historic Plays. In Piotrowski, M. (ed.), *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*. Lausanne, Switzerland: Laboratoire lausannois d'informatique et statistique textuelle. pp. 46-48.
- Schmidt, T., Burghardt, M. and Dennerlein, K. (2018a). Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. Sandra Kübler, Heike Zinsmeister (eds.), *extitProceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)* Sofia, Bulgaria. pp. 47-52.
- Schmidt, T., Burghardt, M. and Dennerlein, K. (2018b). Kann man denn auch nicht lachend sehr ernsthaft sein?“ – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In *Book of Abstracts, DHd 2018*, Cologne, Germany.
- Schöch, Christof. (2016). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*. <http://doi.org/10.5281/zenodo.166356>
- Trilcke, P., Fischer, F. and Kampkaspar, D. (2015). Digital Network Analysis of Dramatic Texts. *Book of Abstracts, DH 2015*. Sidney, Australia
- Underwood, T., Black, M.L., Auvil, L., & Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. *2013 IEEE International Conference on Big Data*, 95-103.
- Wilhelm, T., Burghardt, M., and Wolff, C. (2013). “To See or Not to See” - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, and J. Sieck (Eds.), *Kultur und Informatik: Visual Worlds & Interactive Spaces*. Glückstadt: Verlag Werner Hülsbusch, pp. 175–185.
- Yavuz, M. C.: (2019, November) Analyses of Literary Texts by Using Statistical Inference Methods. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics, CLiC-it'19*. Bari, Italy., CEUR-WS.org, online <http://ceur-ws.org/Vol-2481/paper75.pdf>.
- Yavuz, M. C.: (2020, March) Analyses of Characters in Dramatic Works by Using Document Embeddings. In: *Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020)* co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020), Évora, Portugal., CEUR-WS.org, online <http://ceur-ws.org/Vol-2481/paper75.pdf>.
- Yavuz, M. C. (2020, November) “Analyses of Character Networks in Dramatic Works by Using Graphs” *The 7th International Conference on Behavioural and Social Computing*, Bournemouth UK.
- Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing* 23(3): 327-343.
- Xanthos, A., Pante, I., Rochat, Y and Grandjean, M. (2016). Visualising the dynamics of character networks. *Book of Abstracts, DH 2016*. Kraków, Poland, pp. 417-419.
- Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, 39(2), (1984). 117–123.