EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition

Pierpaolo Basile*
Università degli Studi di Bari Aldo Moro

Malvina Nissim*
Rijksuniversiteit Groningen

Viviana Patti*
Università degli Studi di Torino

Rachele Sprugnoli*
Fondazione Bruno Kessler and
Università degli Studi di Trento

Francesco Cutugno*
Università degli Studi di Napoli
"Federico II"

EVALITA, the evaluation campaign of Natural Language Processing and Speech Tools for the Italian language, was organised for the fifth time in 2016. Six tasks, covering both re-reruns as well as completely new tasks, and an IBM-sponsored challenge, attracted a total of 34 submissions. An innovative aspect at this edition was the focus on social media data, especially Twitter, and the use of shared data across tasks, yielding a test set with layers of annotation concerning PoS tags, sentiment information, named entities and linking, and factuality information. Differently from the previous edition(s), many systems relied on a neural architecture, and achieved best results when used. From the experience and success of this edition, also in terms of dissemination of information and data, and in terms of collaboration between organisers of different tasks, we collected some reflections and suggestions that prospective EVALITA chairs might be willing to take into account for future editions.

1. Introduction

Shared tasks are a common tool in the Natural Language Processing community to set benchmarks for specific tasks and facilitate and promote the development of comparable systems. In practice, a group of researchers can set up a specific task, provide development and test data for it, and solicit the participation of research groups in the community, who will develop systems to address the task at hand. Such competitions often take place within larger frameworks, where multiple tasks are organised and coordinated at the same time. A prime example of such frameworks is SemEval[1], a well-known series of evaluation campaigns with a specific focus on semantic phenomena. In this contribution, we describe a framework for coordinated evaluation campaigns which rather than being focused on specific language processing phenomena, is centred on a variety of phenomena for a single language, namely Italian.

* Group - Address. E-mail: evalita2016@gmail.com
1 https://en.wikipedia.org/wiki/SemEval

EVALITA[2] is the evaluation campaign of Natural Language Processing and Speech Tools for Italian. Since its first edition in 2007, the aim of the campaign is to support the development and dissemination of NLP resources and technologies for Italian. To this end, many shared tasks, covering the analysis of both written and spoken language at various levels of processing, have been proposed within EVALITA.

EVALITA is an initiative of the Italian Association for Computational Linguistics[3] (AILC) and it is endorsed by the Italian Association of Speech Science[4] (AISV) and by the NLP Special Interest Group of the Italian Association for Artificial Intelligence[5] (AI*IA). Since 2014, EVALITA is organised in connection with the yearly Italian Conference on Computational Linguistics (CLiC-it), and co-located with it.

In 2016, EVALITA was organised around a set of six shared tasks and an application challenge, and included several novelties compared to previous years. Most of these novelties were introduced on the basis of the outcome of two questionnaires and of the fruitful discussion that took place during the panel "Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign" held in the context of the Second Italian Computational Linguistics Conference (CLiC-it 2015)[6]. For example, the 2016 edition saw a greater involvement of industrial companies in the organisation of tasks, the introduction of a task and a challenge that are strongly application-oriented, and the creation of cross-task shared data. Also, a strong focus has been placed on using social media data, so as to promote the investigation into the portability and adaptation of existing tools, up to now mostly developed for the news domain.

In just a few words, what characterised the 2016 edition is that *EVALITA went social*, from a range of perspectives: most tasks used social media data, and the newly introduced IBM challenge dealt with web-based applications. Also 'social' aspects within the community were enhanced: task organisers were encouraged to collaborate on the creation of a shared test set across tasks, and to eventually share all resources with everyone — this has resulted in the creation of a repository that is already accessible[7]. In addition to the standard webpage, EVALITA also appeared on social channels for the first time, by means of the regular use of a Facebook page[8] and a Twitter account [9] for updates and dissemination. We believe this has contributed to boost the number of interested teams and actual participants in the end.

Contributions. This paper offers an overview of the tasks at EVALITA 2016 including, for each, a brief description, a summary of the participating systems, and results, so as to provide a reliable overview of the state-of-the-art for Italian NLP in the targeted areas. For task-re-runs we also compare systems and results to those of previous years, and, whenever possible, also draw comparisons to similar tasks for other languages, especially within the SemEval campaigns. Additionally, we provide some general observations on two of the major innovations in 2016, namely the use of shared data across tasks, and on the use of data from social media. Focusing on the 'social' flavour of EVALITA 2016, we also devote some space to discuss the development of the EVALITA

2 http://www.evalita.it
3 http://www.ai-lc.it/
4 http://www.aisv.it/
5 http://www.aixia.it/
6 http://www.evalita.it/towards2016
7 https://github.com/evalita2016/data
8 https://www.facebook.com/evalita2016/
9 https://twitter.com/EVALITAcampaign

community, in terms of chairs, organisers, and participating teams. On the basis of this experience as well as on the observations gathered from the questionnaire's results (Sprugnoli, Patti, and Cutugno 2016), we finally offer some ideas and recommendations that the organisers of future EVALITA editions might want to take into account.

2. Tasks and Challenge

As in previous editions, both the tasks and the final workshop were collectively organised by several researchers from the community working on Italian language resources and technologies (see Section 4 for more details). As visible in Figure 1, the 2016 edition featured two re-runs of EVALITA 2014, namely sentiment analysis (SENTIPOLC), and pos tagging (PoSTWITA). However, while the former was an almost exact replica of the previous year (see Section 2.1 for specific differences), the latter shifted its focus on social media data from previously used newswire texts, thereby making this a substantially innovative task in the EVALITA panorama (also because using universal tagset). The other four tasks, and the challenge, were all newly developed in the context of the 2016 edition, though for some there are connections to previous tasks .

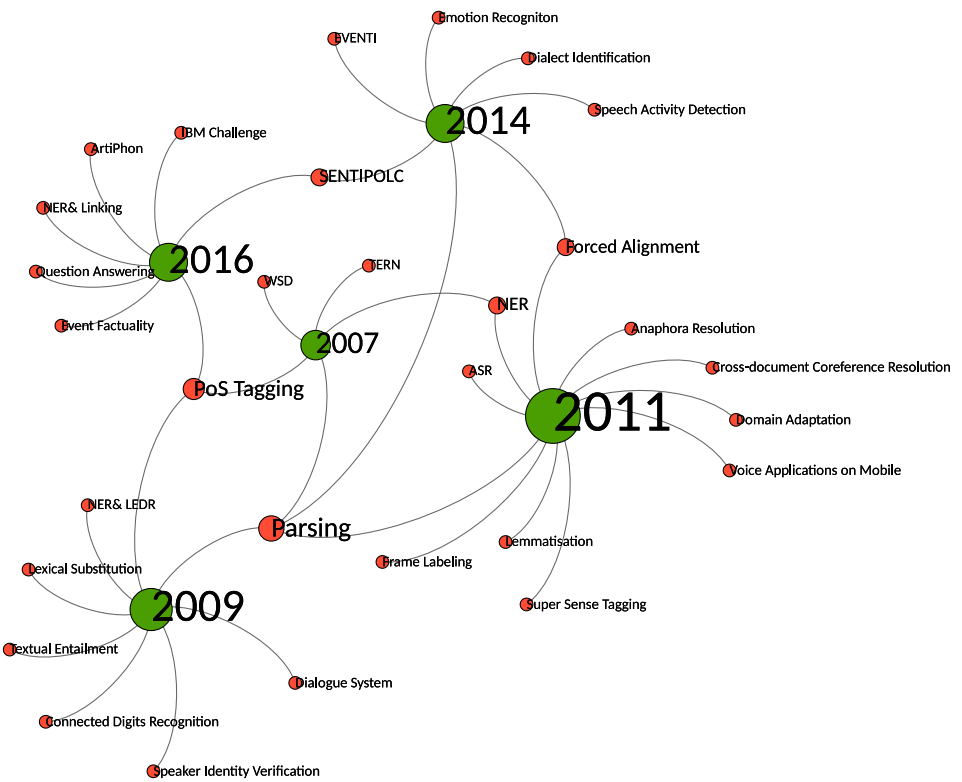

Figure 1
Overview of the tasks organised at EVALITA campaigns 2007–2016.

In the remainder of this section, we provide detailed information about the EVALITA 2016 tasks. First, we describe the four evaluation exercises that shared the

test set (i.e., SENTIPOLC, PoSTWITA, NEEL-It, and FactA—see also Section 3). Next, we report on the speech task (Artiphon), followed by the application-oriented task (QA4FAQ) and lastly on the IBM Challenge. The names of the groups used in the following subsections are directly taken from the reports written by task participants and organisers. We provide a mapping for the used abbreviations in Table 1. Please note that different names can refer to groups formed by the same members (e.g., ILC-CNR and ItaliaNLP) and that the same affiliation can cover different departments of the same institution (e.g. MicroNeel and fbk4faq).

Table 1
Mapping between participating groups and institutions

GROUP	TASK	INSTITUTION
ILC-CNR	PoSTWITA	Consiglio Nazionale delle Ricerche (CNR)
ItaliaNLP	SENTIPOLC	
samskara	SENTIPOLC	
ISTC	ArtiPhon	
MicroNeel	NEEL-it	Fondazione Bruno Kessler (FBK)
FBK-HLT-NLP	NEEL-it	
fbk4faq	QA4FAQ	
MARTIN	IBM Challenge	
UniPI	NEEL-it	University of Pisa
UniPisa	PoSTWITA	
UniPI	SENTIPOLC	
CoLingLab	SENTIPOLC	
UniBologna	PoSTWITA	University of Bologna
CoMoDI	SENTIPOLC	
UniBO	SENTIPOLC	
UniDuisburg	PoSTWITA	University of Duisburg-Essen
MIVOQ	PoSTWITA	Mivoq Srl
sisinflab	NEEL-it	Polytechnic University of Bari
UNIMIB	NEEL-it	University of Milano-Bicocca
UniGroningen	PoSTWITA	University of Groningen
ILABS	PoSTWITA	Integris Srl
EURAC	PoSTWITA	EURAC Research
NITMZ	PoSTWITA	National Institute of Technology
NLP-NITMZ	QA4FAQ	National Institute of Technology & IPN Mexico
chiLab4It	QA4FAQ	University of Palermo
ADAPT	SENTIPOLC	Adapt Centre
INGEOTEC	SENTIPOLC	CentroGEO/INFOTEC CONACyT
IntIntUniba	SENTIPOLC	University of Bari
IRADABE	SENTIPOLC	Uni. Pol. de Valencia & Uni. de Paris & Uni. of Turin
SwissCheese	SENTIPOLC	Zurich University of Applied Sciences
tweet2check	SENTIPOLC	Finsa s.p.a.
Unitor	SENTIPOLC	University of RomaTor Vergata
Appetitoso ChatBot	IBM Challenge	Kloevolution S.r.l. & University of Trento
Stockle	IBM Challenge	INRIA & SciLifeLab

2.1 SENTIPOLC

2.1.1 Task Description, Data, and Evaluation Metrics

SENTIPOC (*SENTIment POLarity Classification*) is a sentiment analysis task where systems are required to automatically annotate tweets with a tuple of boolean values indicating the message's subjectivity, its polarity (positive or negative), and whether

it is ironic or not (Barbieri et al. 2016). The SENTIPOLC task is indeed organised along three subtasks:[10]

- Task 1 – Subjectivity Classification: *a system must decide whether a given message is subjective or objective*. In Table 3 the value related to this task is expressed as `subj`, and allows for values {0,1}.
- Task 2 – Polarity Classification: *a system must decide whether a given message is of positive, negative, neutral or mixed sentiment.* In our data, positive and negative polarities are *not* mutually exclusive and each is annotated as a binary category. A tweet can thus be at the same time positive *and* negative, yielding a `mixed` polarity, or also neither positive nor negative, meaning it is a subjective statement with `neutral` polarity. Polarity is a valid field only in conjunction with subjectivity. See (Basile et al. 2014; Barbieri et al. 2016) for further details and examples. In Table 3, overall polarity values are expressed as `opos` and `oneg`, each of them allowing for presence or absence ({0,1}).
- Task 3 – Irony Detection: *a system must decide whether a given message is ironic or not.* Twitter communications include a high percentage of ironic messages (Reyes and Rosso 2014), and because of the polarity-reversing effect that irony can have (one says something "good" to mean something "bad"), systems are heavily affected by this. (Bosco, Patti, and Bolioli 2013; Ghosh et al. 2015). In Table 3,this value is reported as `iro` and allows for values {0,1}.

The data includes an additional layer of annotation, which specifies the *literal polarity* of a tweet. In non-ironic cases the values correspond to the overall polarity, while in the ironic tweets, they could be different (see examples in Table 3, values reported as `lpos` and `lneg`). This layer is not used in any evaluation directly, but it was provided in case teams wanted to make use of it, especially in dealing with the polarity reversing property of irony.

Development and Test Data. The full dataset released for the shared task comprises the whole of the SENTIPOLC 2014 dataset (training and test, TW-SENTIPOLC14, 6421 tweets (Basile et al. 2014)), 1500 tweets from TWitterBuonaScuola (TW-BS, (Stranisci et al. 2016)), and two brand new sets: 500 tweets selected from the TWITA 2015 collection (TW-TWITA15, (Basile and Nissim 2013)), and 1000 (filtered to 989) tweets collected in the context of the NEEL-IT shared task (TW-NEELIT, (Basile et al. 2016)).

The subsets of data extracted from existing corpora (TW-SENTIPOLC14 and TW-BS) were revised/enriched according to the new annotation guidelines specifically devised for this task (please consult (Barbieri et al. 2016) for details). The tweet from NEEL-IT and TWITA15, instead, were annotated completely from scratch using CrowdFlower[11], a crowdsourcing platform which has also been recently used for a similar annotation task (Nakov et al. 2016).

The TWITA15 collection, which comprises the 301 tweets also used as test data in the PoSTWITA (Tamburini et al. 2016), NEEL-IT-it (Basile et al. 2016) and FactA (Minard, Speranza, and Caselli 2016) EVALITA 2016 shared tasks (see Section 3 for

10 The three tasks are meant to be independent. For example, a team could take part in the polarity classification task without tackling Task 1.

11 `http://www.crowdflower.com/`

Table 2
Dataset breakdown for SENTIPOLC 2016. We specify if there was pre-existing annotation in the datasets we used (pre-annot), and what new annotations were added to comply with the SENTIPOLC 2016's guidelines. We also distinguish whether the new annotation was performed by the Crowd (C) or by Experts (E). "l-polarity" stands for literal polarity (`lpos`)

SOURCE	PRE-ANNOT	ADDED ANNOTATIONS	set	SIZE
TW-SENTIPOLC 2014	yes	l-polarity for ironic tweets (E)	train	6421
TW-NEELIT	no	all (C)	train	989
total train				7410
TW-BS	yes	polarity to ironic tweets (E) l-polarity to ironic tweets (E) potential subj to neutral (E)	test	1500
TW-TWITA15	no	all (C + E)	test	500
total test				2000
total				9410

further details on the cross-task shared data), was additionally annotated by experts, so that the resulting labels are a product of crowd and expert agreement.[12]

Table 3
Examples of tweets exhibiting a variety of annotation combinations according to the SENTIPOLC 2016 guidelines.

description and *example tweet in Italian*	`subj`	`opos`	`oneg`	`iro`	`lpos`	`lneg`
subjective with neutral polarity and no irony *Primo passaggio alla #strabrollo ma secondo me non era un iscritto*	1	0	0	0	0	0
subjective with mixed polarity and no irony *Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"* `http://t.co/kIKnbFY7`	1	1	1	0	1	1
subjective with negative polarity, and an ironic twist *Calderoli: Governo Monti? Banda Bassotti ..infatti loro erano quelli della Magliana.. #FullMonti #fuoritutti #piazzapulita*	1	0	1	1	0	1
subjective with negative polarity, an ironic twist, and positive literal polarity *Ho molta fiducia nel nuovo Governo Monti. Più o meno la stessa che ripongo in mia madre che tenta di inviare un'email.*	1	0	1	1	1	0
subjective with negative polarity, an ironic twist, and neutral literal polarity *arriva Mario #Monti: pronti a mettere tutti il grembiulino?*	1	0	1	1	0	0

In Table 3 we show a few examples and their annotation. For a comprehensive set of examples and an explanation of allowed combinations, please refer to the SENTIPOLC 2016's report (Barbieri et al. 2016).

12 The organisers of SENTIPOLC mention that the Crowdflower data had to undergo some post-validation for compliance with the guidelines. For all details, please refer to the SENTIPOLC 2016 report (Barbieri et al. 2016).

Evaluation. Evaluation is performed using precision, recall, and f-score, and is defined per subtask.

- *Task1* — Systems are evaluated on the assignment of a 0 or 1 value to the subjectivity field. A response is considered correct or wrong in comparison to the gold standard annotation. We compute precision (p), recall (r) and F-score (F) for each class (`subj`, `obj`), and the overall F-score is the average of the two F-scores.
- *Task2* — The coding system allows for four combinations of `opos` and `oneg` values: 10 (positive polarity), 01 (negative polarity), 11 (mixed polarity), 00 (no polarity). Accordingly, we evaluate positive and negative polarity independently by computing precision, recall and F-score for both classes (0 and 1). The F-score for the two polarity classes is the average of the F-scores of the respective pairs. Finally, the overall F-score for Task 2 is given by the average of the F-scores of the two polarities.
- *Task3* — Systems are evaluated on their assignment of a 0 or 1 value to the irony field. A response is considered fully correct or wrong when compared to the gold standard. We measure precision, recall and F-score for each class (`ironic`, `non-ironic`), similarly to the Task1. The overall F-score will be the average of the F-scores for ironic and non-ironic classes.

2.1.2 Participating Systems and Results

A total of 13 teams from 6 different countries participated in at least one of the three SENTIPOLC tasks. Almost all teams participated to both subjectivity and polarity classification subtasks. Each team had to submit at least a constrained run. Furthermore, teams were allowed to submit up to four runs (2 constrained and 2 unconstrained) in case they implemented different systems. Overall we have 19, 26, 12 submitted runs for the subjectivity, polarity, and irony detection tasks, respectively. Most of the submissions were constrained: three teams (*UniPI*, *Unitor* and *tweet2check*) participated with both a constrained and an unconstrained runs on the both the subjectivity and polarity subtasks. Unconstrained runs were submitted to the polarity subtask only by *IntIntUniba.SentiPy* and *INGEOTEC.B4MSA*. Differently from SENTIPOLC 2014, unconstrained systems performed better than constrained ones, with the only exception of *UniPI*, whose constrained system ranked first for the polarity classification subtask.

A single-ranking table was produced for each subtask, where unconstrained runs are properly marked. Notice that only the average F-score was used for global scoring and ranking. For each task, we ran a majority class baseline to set a lower-bound for performance.

Table 4 shows results for the subjectivity classification task. All participant systems show an improvement over the baseline. The highest F-score is achieved by *Unitor* at 0.7444, which is also the best unconstrained performance (Castellucci, Croce, and Basili 2016). Among the constrained systems, the best F-score is achieved by *samskara* with $F = 0.7184$ (Russo and Monachini 2016).

Table 5 shows results for the polarity classification task, which was again the most popular subtask with 26 submissions from 12 teams. Also in this case, all participant systems show an improvement over the baseline. The highest F-score is achieved by *UniPi* at 0.6638 (Attardi et al. 2016a), which is also the best score among the constrained

Table 4
Task 1 (Subjectivity Classification): F-scores for constrained ".c" and unconstrained runs ".u". After the deadline, two teams reported about a conversion error from their internal format to the official one. The resubmitted amended runs are marked with *.

TEAM ID	OBJ	SUBJ	F	METHOD	EMB	OTHER RESOURCES
Unitor.1.u	**0.6784**	**0.8105**	**0.7444**	CNN	word	Tw data, hm-Lex
Unitor.2.u	0.6723	0.7979	0.7351	CNN	word	Tw data, hm-Lex
samskara.1.c	0.6555	0.7814	**0.7184**	Naive Bayes	-	hm-Lex
ItaliaNLP.2.c	0.6733	0.7535	0.7134	LSTM-SVM	word	Lex
IRADABE.2.c	0.6671	0.7539	0.7105	SVM	-	hm-Lex
INGEOTEC.1.c	0.6623	0.7550	0.7086	SVM	-	Tw data
Unitor.c	0.6499	0.7590	0.7044	CNN	word	-
UniPI.1/2.c	**0.6741**	0.7133	0.6937	CNN	word	Lex
UniPI.1/2.u	0.6741	0.7133	0.6937	CNN	word	Tw data
ItaliaNLP.1.c	0.6178	0.7350	0.6764	LSTM-SVM	word	Lex
ADAPT.c	0.5646	0.7343	0.6495	-	-	-
IRADABE.1.c	0.6345	0.6139	0.6242	DNN-SVN	-	hm-Lex
tweet2check16.c	0.4915	0.7557	0.6236	-	-	yes (un)
tweet2check14.c	0.3854	**0.7832**	0.5843	-	-	yes (un)
tweet2check14.u	0.3653	0.7940	0.5797	-	-	yes (un)
UniBO.1.c	0.5997	0.5296	0.5647	-	-	-
UniBO.2.c	0.5904	0.5201	0.5552	-	-	-
Baseline	0.0000	0.7897	0.3949			
*SwissCheese.c_late	0.6536	0.7748	0.7142	CNN	word	-
*tweet2check16.u_late	0.4814	0.7820	0.6317	-	-	-

runs. As for unconstrained runs, the best performance is achieved by *Unitor* with $F = 0.6620$ (Castellucci, Croce, and Basili 2016) [13].

Table 6 shows results for the irony detection task. which attracted 12 submissions from 7 teams. The highest F-score was achieved by tweet2check at 0.5412 (constrained run) (Di Rosa and Durante 2016). The only unconstrained run was submitted by Unitor achieving 0.4810 as F-score. While all participating systems show an improvement over the baseline (F = 0.4688), many systems score very close to it, highlighting the complexity of the task [14].

Methods. All systems, except *CoMoDI*, exploited machine learning techniques in a supervised setting. Two main strategies emerged. One involves using linguistically principled approaches to represent tweets and provide the learning framework with valuable information to converge to good results. The other exploits state-of-the-art learning frameworks in combination with word embedding methods over large-scale corpora of tweets. On balance, the last approach achieved better results in the final ranks.

13 After the deadline, *SwissCheese* and *tweet2check* reported about a conversion error from their internal format to the official one. The resubmitted amended runs are shown in the table (marked by the * symbol), but the official ranking was not revised.

14 In all the tables above we marked with '-' all cases where the characteristic is not present or we have not clear information about its presence from the participants report. Moreover, notice that ADAPT, UniBO and twee2check didn't provide details about their systems

Table 5
Task 2 (Polarity Classification): F-scores for constrained ".c" and unconstrained runs ".u". Amended runs are marked with * .

TEAM ID	POS	NEG	F	METHOD	EMB	OTHER RESOURCES
UniPI.2.c	**0.6850**	0.6426	**0.6638**	CNN	word	-
Unitor.1.u	0.6354	**0.6885**	**0.6620**	CNN	word	Tw data, hm-Lex
Unitor.2.u	0.6312	0.6838	0.6575	CNN	word	Tw data, hm-Lex
ItaliaNLP.1.c	0.6265	**0.6743**	0.6504	LSTM-SVM	word	Lex
IRADABE.2.c	0.6426	0.6480	0.6453	SVM	-	hm-Lex
ItaliaNLP.2.c	0.6395	0.6469	0.6432	LSTM-SVM	word	Lex
UniPI.1.u	**0.6699**	0.6146	0.6422	CNN	word	Tw data
UniPI.1.c	0.6766	0.6002	0.6384	CNN	word	-
Unitor.c	0.6279	0.6486	0.6382	CNN	word	-
UniBO.1.c	0.6708	0.6026	0.6367	-	-	-
IntIntUniba.sentipy.c	0.6189	0.6372	0.6281	Linear SVC + Emoji Classifier	-	-
IntIntUniba.sentipy.u	0.6141	0.6348	0.6245	-	-	-
UniBO.2.c	0.6589	0.5892	0.6241	-	-	
UniPI.2.u	0.6586	0.5654	0.6120	CNN	word	TW data
CoLingLab.c	0.5619	0.6579	0.6099	SVM	-	hm-Lex
IRADABE.1.c	0.6081	0.6111	0.6096	SVM, DNN	-	hm-Lex
INGEOTEC.b4msa.u	0.5944	0.6205	0.6075	SVM	-	Tw data
INGEOTEC.2.c	0.6414	0.5694	0.6054	SVM	-	-
ADAPT.c	0.5632	0.6461	0.6046	-	-	-
IntIntUniba.sentiws.c	0.5779	0.6296	0.6037	Rocchio Naive Bayes	-	-
tweet2check16.c	0.6153	0.5878	0.6016	-	-	-
tweet2check14.u	0.5585	0.6300	0.5943	-	-	-
tweet2check14.c	0.5660	0.6034	0.5847	-	-	-
samskara.1.c	0.5198	0.6168	0.5683	Naive Bayes	-	Lex
Baseline	0.4518	0.3808	0.4163			
*SwissCheese.c_late	0.6529	0.7128	0.6828	CNN	word	-
*tweet2check16.u_late	0.6528	0.6373	0.6450	-	-	-

However, with F-scores of 0.744 (unconstrained) and 0.7184 (constrained) in *subjectivity recognition* and 0.6638 (constrained) and 0.6620 (unconstrained) in *polarity recognition*, we are still far from having solved sentiment analysis on Twitter. For the future, we envisage the definition of novel approaches, for example by combining neural network-based learning with a linguistic-aware choice of features.

Many teams adopted learning methods already investigated in SENTIPOLC 2014; in particular, Support Vector Machine (SVM) is the most adopted learning algorithm. The SVM is generally based over specific linguistic/semantic feature engineering, as discussed for example by *ItaliaNLP*, *IRADABE*, *INGEOTEC* or *ColingLab*. Micro-blogging specific features such as emoticons and hashtags are also adopted, for example by *ColingLab*, *INGEOTEC*) or *CoMoDi*. In addition, some teams (e.g. *ColingLab*) adopted Topic Models to represent tweets. *Samskara* also used feature modelling with a communicative and pragmatic value. *CoMoDi* is one of the few systems that investigated irony-specific features. Other methods have been also used, as a Bayesian approach by *samskara* (achieving good results in polarity recognition) combined with linguistically motivated feature modelling. *CoMoDi* is the only participant that adopted a rule based approach in combination with a rich set of linguistic cues dedicated to irony detection. Approaches based on Convolutional Neural Networks (CNN) have been investigated at 2016 SENTIPOLC for the first time by a few teams. Deep learning methods adopted by

Table 6
Task 3 (Irony detection): F-scores for constrained ".c" and unconstrained runs ".u". Amended runs are marked with *.

TEAM ID	NON IRO	IRO	F	METHOD	EMB	OTHER RESOURCES
tweet2check16.c	0.9115	0.1710	**0.5412**	-	-	-
CoMoDI.c	0.8993	0.1509	0.5251	Rule-based	-	Lex
tweet2check14.c	0.9166	0.1159	0.5162	-	-	-
IRADABE.2.c	0.9241	0.1026	0.5133	SVM	-	hm-Lex
ItaliaNLP.1.c	0.9359	0.0625	0.4992	LSTM-SVM	word	Lex
ADAPT.c	0.8042	**0.1879**	0.4961	-	-	-
IRADABE.1.c	0.9259	0.0484	0.4872	SVM,DNN	-	hm-Lex
Unitor.2.u	**0.9372**	**0.0248**	**0.4810**	CNN	word	TW data
Unitor.c	0.9358	0.0163	0.4761	CNN	word	-
Unitor.1.u	0.9373	0.0084	0.4728	CNN	word	Tw data
ItaliaNLP.2.c	**0.9367**	0.0083	0.4725	LSTM-SVM	word	Lex
Baseline	0.9376	0.000	0.4688			
*SwissCheese.c_late	0.9355	0.1367	0.5361	CNN	word	-

some teams, such as *UniPi* and *SwissCheese* required to model individual tweets through geometrical representation of tweets, i.e. vectors. Words from individual tweets are represented through *Word Embeddings*, mostly derived by using the Word2Vec tool or similar approaches. *Unitor* extends this representation with additional features derived from Distributional Polarity Lexicons.

The majority of teams also used external resources, such as lexicons specific for Sentiment Analysis tasks. Some teams used already existing lexicons (referred as *Lex* in the tables above), such as *Samskara*, *ItaliaNLP*, *CoLingLab*, or *CoMoDi*, while others created their own task specific resources, such as *Unitor*, *IRADABE*, *CoLingLab* (referred as *hm-Lex* in the tables above).

Unconstrained runs. Some teams submitted unconstrained results, as they used additional Twitter annotated data for training their systems (*Tw data* in the above tables). In particular, *UniPI* used a silver standard corpus made of more than 1M tweets to pre-train the CNN; this corpus is annotated using a polarity lexicon and specific polarised words. Also *Unitor* used external tweets to pre-train their CNN. This corpus is made of the contexts of the tweets populating the training material and automatically annotated using the classifier trained only over the training material, in a semi-supervised fashion. Moreover, *Unitor* used distant supervision to label a set of tweets used for the acquisition of their so-called Distribution Polarity Lexicon. Distant supervision is also adopted by *INGEOTEC* to extend the training material for the their SVM classifier. For a deeper comparison between participating systems and approaches see (Barbieri et al. 2016). As a final note, we would like to mention that the distinction between constrained and unconstrained runs, that we still maintained in this edition, becomes less meaningful when we consider that, as shown the tables above, many *constrained* systems exploited word embeddings built on huge amounts of additional (Twitter) data. The traditional distinction, which normally focuses on using or not using additional training data annotated according to the task guidelines, was meant to guarantee a fair comparison.

However, as this distinction might generally become a bit blurred, it is worth reflecting on whether it makes sense to use it in future editions.

2.1.3 Links to Other Shared Tasks

Previous EVALITA tasks. SENTIPOLC 2016 is a re-run of SENTIPOLC 2014, which had been introduced then for the first time, and had attracted the highest number of participants among EVALITA tasks. The main differences between the 2014 and the 2016 editions lie in the data and in the best performing algorithms. Regarding annotation, two new annotation fields which express *literal polarity* have been added, in order to provide insights into the mechanisms behind polarity shifts in the presence of figurative usage. Also, a portion of the data was annotated via Crowdflower rather than by experts. Regarding the source of data, the test set is drawn from Twitter, but it is composed of a portion of random tweets and a portion of tweets selected via keywords which do not exactly match the selection procedure that led to the creation of the training set. This was intentionally done as a novelty in 2016 to observe the portability of supervised systems, in line with what suggested in (Basile et al. 2015).

Finally, concerning systems, for the first time at SENTIPOLC neural models were used with success, achieving best results especially in the open runs. Although evaluated over a different dataset, the best systems also show better, albeit comparable, performance for subjectivity with respect to 2014's systems, and outperform them for polarity (if we include late submissions). The use of a progress set, as already done at SemEval, would allow for a proper evaluation across the various editions, and would definitely be a welcome innovation at next edition.

In contrast to the improvement in performances from 2014 to 2016, irony detection appears truly challenging, and the systems' performance drops in 2016 w.r.t the previous edition. The task's complexity does not depend (only) on the inner structure of irony, but also on unbalanced data distribution (1 out of 7 examples is ironic in the training set). Examples in the dataset are probably not sufficient to generalise over the structure of ironic tweets. Future campaigns could consider including a larger and more balanced dataset of ironic tweets in future campaigns.

Non-EVALITA tasks. Sentiment classification on Twitter is by now an established task internationally. Such solid and growing interest is reflected in the fact that the Sentiment Analysis tasks at SemEval (where they constitute now a whole track) have attracted the highest number of participants in the last years (Rosenthal et al. 2014, 2015; Nakov et al. 2016). It is interesting to highlight that the Swiss team *SwissCheese*, which achieved the best score in polarity classification (including late submissions) at SENTIPOLC 2016 was the top-scoring team also at the 'twin task' for English at Semeval2016-Task4 (Nakov et al. 2016). Task 10 at SemEval 2015 was concerned with irony in Twitter, but rather than as an irony *detection* task, it was designed as a polarity detection task in tweets that were already known to be ironic. This is also an avenue that could be explored for Italian. More generally, anyway, tasks revolving around the use of non literal language are becoming more popular: two of the five tasks of the sentiment analysis track at SemEval in 2017 are organised around humor-related topics.

2.2 PoSTWITA

2.2.1 Task Description, Data and Evaluation Metrics

PoSTWITA consists in developing systems for the Part-Of-Speech (PoS) tagging of tweets: in other words, it concerns with the domain adaptation of tools built for stan-

Table 7
Size of PoSTWITA datasets and an annotated example

	Development Set	Test Set	Example
# tweets	6,438	300	___579013335921885184___ @LudovicaCagnino MENTION
# tokens	114,967	4,759	Grazie INTJ AMORE PROPN

dardized texts to social media data (Tamburini et al. 2016). As for the other EVALITA 2016 tasks devoted to the automatic processing of Twitter posts, the final aim of PoSTWITA is to promote research in the automatic extraction of knowledge from social media texts written in Italian. In order to deal with these type of data, it is crucial to have annotation guidelines and resources that take into consideration the linguistic peculiarities of Twitter language. To this end, a specific tagset was defined and new datasets were released.

As for the tagset, the main labels are inherited from the ones adopted within the Universal Dependencies (UD) project for Italian[15] so to make the resources annotated for the task compliant with the UD treebanks. Anyway, novel tags are introduced to cover three cases: (i) articulated prepositions (`ADP_A`, *della*); (ii) clusters made by a verb and a clitic pronoun (`VERB_CLIT`, *mangiarlo*); (iii) Twitter-specific elements i.e., emoticons (`EMO`, *:-)*), web addresses (`URL`, *http://www.site.it*), email addresses (`EMAIL`, *name@domain.it*), hashtags (`HASHTAG`, *#staisereno*) and mentions (`MENTION`, *@someone*). The use of the first two labels described above, allow to tokenize by maintaining words unsplitted rather than splitted as in the original UD format.

Once defined the tagset, development (DS) and test sets (TS) have been collected and annotated. DS data are taken from the EVALITA 2014 SENTIPOLC dataset while TS is shared with other EVALITA 2016 tasks (see Section 3). As for the annotation, in the first step tokenisation and annotation were performed automatically then two expert annotators manually corrected the same tweets in parallel. At the end, an adjudicator resolved disagreements. Table 7 reports the size of DS and TS together with an annotated example. No additional resource was distributed to the participants who, however, by following an open task approach, had the opportunity to use external data to develop their systems.

Systems output is evaluated in terms of tagging accuracy, that is the number of correct PoS tags divided by the total number of tokens in TS.

2.2.2 Participating Systems and Results

Although 16 teams registered to the task, only 9 submitted a run to be evaluated. Among these groups, 7 are affiliated to universities or other research centers located in Italy and abroad (India, The Netherlands, and Germany) and 2 are made by representatives of Italian private companies. Table 8 presents the official results of PoSTWITA runs with an accuracy ranging from 0.7600 to 0.9319. Three systems are based on traditional machine learning algorithms (i.e., CRF, HMM , and SVM) while all the others employ Deep Neural Networks: perceptron algorithm in one case and Long Short-Term Memo-

15 `http://universaldependencies.org/it/pos/index.html`

Table 8
PoSTWITA results in terms of accuracy (ACC.) with details about the main characteristics of the participating systems

TEAM ID	ACC.	METHOD	EMBEDDINGS	OTHER RESOURCES
ILC-CNR	0.9319	two-branch Bi-LSTM NN	word&char	yes
UniDuisburg	0.9286	CRF classifier	-	yes
MIVOQ	0.9271	CRF + HMM + SVM	-	yes
UniBologna	0.9246	Stacked Bi-LSTM NN + CRF	word	yes
UniGroningen	0.9225	Bi-LSTM NN	word	yes
UniPisa	0.9157	Bi-LSTM NN + CRF	word&char	yes
ILABS	0.8790	Perceptron algorithm	-	yes
NITMZ	0.8596	HMM bigram model	-	-
EURAC	0.7600	LSTM NN	word&char	yes

ries (LSTM) in the remaining systems. More specifically, bidirectional LSTM (BiLSTM) proves to be optimal being used by 4 out of the 6 top-performing systems. In addition, experiments carried on during the development of the best-scoring system show that a two-branch architecture of BiLSTM performs clearly better than a single bi-LSTM with an improvement of about 0.5 points on DS and 0.84 on TS (Cimino and Dell'Orletta 2016). The majority of systems use word-level or character-level embeddings as inputs for their systems: with respect to the former, the latter performs well giving a finer representation of words and thus better coping with the noisy language of tweets (Attardi and Simi 2016). In all but one case, additional corpora or resources were employed: word clusters (Horsmann and Zesch 2016), morphological analysers (Tamburini 2016), external dictionaries (Paci 2016), annotated and non-annotated corpora (Stemle 2016). As for this last point, Plank and Nissim (2016) show that adding a small amount of in-domain data is more effective than using data form different genres.

Error analysis on systems output highlights that the most challenging distinction in terms of tag assignment is between nouns and proper nouns. In addition, the performances on the DS and on the TS register a large difference: on DS, top systems reach an accuracy above 0.95 thus more than 0.3 points with respect to the results on the TS. Indeed, if compared with state-of-the art systems built for other languages (Owoputi et al. (2013) reports an accuracy of 0.93 for English), results on DS seems particularly good. This may be due to the strong homogeneity of the training data. On the other side, tweets in the TS covered different topics with respect to the DS: for example, there are only two mentions (*@Youtube* and *@matteorenzi*) and four very generic hashtags (*#governo*, *#news*, *#rt*, *#lavoro*) that are present in both the DS and the TS.

2.2.3 Links to Other Shared Tasks

Previous EVALITA Tasks. Both EVALITA 2007 and EVALITA 2009 hosted an evaluation exercise on PoS tagging but their focus was on texts with standard forms and syntax. During the first edition of the campaign, the task was designed around a corpus of different genres (newspaper articles, narrative prose, academic and legal texts) and organized in two subtasks based on two tagsets: i.e., EAGLES and DISTRIB. The former has a long tradition in computational linguistics while the latter is distributionally and syntactically oriented (Monachini 1995; Bernardi et al. 2005). Among the 11 submitted systems, the majority uses SVM or a combination of taggers plus additional resources. In particular, morphological analysers play a crucial role in many of these systems.

Accuracy ranges from 0.8871 to 98.04 with the EAGLES tagset and from 0.9142 to 0.9768 with the DISTRIB tagset (Tamburini 2007). The 2009 task had a higher complexity due to the adoption of a larger tagset (37 tags with morphological variants and 336 morphed tags) and to the fact that the training and test corpora consisted of texts belonging to different genres (newspaper articles for training and Wikipedia pages for test). Accuracy was measured on morphed tags and also on the tags without morphology following both a closed and an open approach. As in the previous year, most of the 8 participating systems were based on a combination of taggers. Results in the open subtask are all above 0.95 while the closed subtask saw a greater variability with an accuracy ranging from 0.9164 to 0.9691. By looking at these previous EVALITA PoS evaluation exercise, it is easy to see that the performances of systems annotating tweets are lower than when applied to standardized texts thus there is room for improvements. It is also worth noting that in EVALITA 2016 has witnessed the breakthrough of deep learning techniques also for the PoS tagging task.

Non EVALITA Tasks. EmpiriST 2015 shared task on automatic linguistic annotation of computer-mediated communication (CMC) and social media had a subtask on PoS tagging of CMC discourse in German. The dataset is made of data from different CMC genres including tweets (Beißwenger et al. 2016). Performances are lower than the ones registered in EVALITA 2016 with the best system obtaining an accuracy of 0.9028.

2.3 NEEL-it

2.3.1 Task Description, Data and Evaluation Metrics

The NEEL-it[16] task consists in automatically annotating each named entity mention (belonging to the following categories: `Thing`, `Event`, `Character`, `Location`, `Organization`, `Person` and `Product`) in a tweet by linking it to the DBpedia knowledge base (Basile et al. 2016). Tweets represent a great wealth of information useful to understand recent trends and user behaviour in real-time. Usually, natural language processing techniques would be applied to such pieces of information in order to make them machine-understandable. Named Entity rEcongition and Linking (NEEL) is a particularly useful technique aiming to automatically annotate tweets with named entities. However, due to the noisy nature and shortness of tweets, this technique is more challenging in this context than elsewhere.

NEEL-it followed a setting similar to the NEEL challenge for English Micropost on Twitter (Rizzo et al. 2016). Specifically, each task participant is required to: 1) recognize and typing each entity mention that appears in the text of a tweet; 2) disambiguate and link each mention to the canonicalized DBpedia 2015-10, which is used as referent Knowledge Base; 3) cluster together the non linkable entities, which are tagged as NIL, in order to provide a unique identifier for all the mentions that refer to the same named entity. In the annotation process, a named entity is a string in the tweet representing a proper noun that: 1) belongs to one of the categories specified in a taxonomy and/or 2) can be linked to a DBpedia concept. This means that some concepts have a NIL DBpedia reference[17].

From the annotation are excluded the preceding article (like *il, lo, la*, etc.) and any other prefix (e.g. *Dott., Prof.*) or post-posed modifier. Each participant is asked to produce an annotation file with multiple lines, one for each annotation. A line is a tab

16 `http://neel-it.github.io/`

17 These concepts belong to one of the categories but they have no corresponding concept in DBpedia

separated sequence of tweet id, start offset, end offset, linked concept in DBpedia, and category. For example, given the tweet with id 288976367238934528:

Chameleon Launcher in arrivo anche per smartphone: video beta privata su Galaxy Note 2 e Nexus 4: Chameleon Laun...

the annotation process is expected to produce the output as reported in Table 9.

Table 9
Example of annotations.

id	begin	end	link	type
288...	0	18	NIL	`Product`
288...	73	86	http://dbpedia.org/resource/Samsung_Galaxy_Note_II	`Product`
288...	89	96	http://dbpedia.org/resource/Nexus_4	`Product`
290...	1	15	http://dbpedia.org/resource/Carlotta_Ferlito	`Person`

The annotation process is also expected to link Twitter mentions (@) and hashtags (#) that refer to a named entities, like in the tweet with id 290460612549545984:

@CarlottaFerlito io non ho la forza di alzarmi e prendere il libro! Help me

the correct annotation is also reported in Table 9.

Participants were allowed to submit up to three runs of their system as TSV files. We encourage participants to make available their system to the community to facilitate reuse.

As for the NEEL-IT corpus, it consists of both a development set (released to participants as training set) and a test set. Both sets are composed by two TSV files: (1) the tweet id file, i.e, a list of all tweet ids used for training; (2) the gold standard, containing the annotations for all the tweets in the development set following the format showed in Table 9. The development set was built upon the dataset produced by (Basile, Caputo, and Semeraro 2015). This dataset is composed by a sample of 1,000 tweets randomly selected from the TWITA dataset (Basile and Nissim 2013). We updated the gold standard links to the canonicalized DBpedia 2015-10. Furthermore, the dataset underwent another round of annotation performed by a second annotator in order to maximize the consistency of the links. Tweets that presented some conflicts were then resolved by a third annotator. Data for the test set was generated by randomly selecting 1,500 tweets from the SENTIPOLC test data (Barbieri et al. 2016). From this pool, 301 tweets were randomly chosen for the annotation process and represents our Gold Standard (GS). The GS was choose in coordination with the task organizers of SENTIPOLC (Barbieri et al. 2016), POSTWITA (Tamburini et al. 2016) and FacTA (Minard, Speranza, and Caselli 2016) with the aim of providing a unified framework for multiple layers of annotations (see Section 3). The tweets were split in two batches, each of them was manually annotated by two different annotators. Then, a third annotator intervened in order to resolve those debatable tweets with no exact match between annotations. The whole process has been carried out by exploiting BRAT[18] web-based tool (Stenetorp et al. 2012).

By looking at the annotated data, `Person` results the most populated category among the NIL instances, along to `Organization` and `Product`. In the development

18 `http://brat.nlplab.org/`

Table 10
NEEL-it results in terms of final score with details about the main characteristics of the participating systems

TEAM ID	SCORE	METHOD	EMBEDDINGS	OTHER RESOURCES
UniPI	0.5034	BiLSTM + text similarity	word	yes
MicroNeel	0.4967	Tint + The Wiki Machine		no
FBK-HLT-NLP	0.4932	EntityPro + NewsReader		yes
Sisinflab	0.3418	ensemble	word	no
UNIMIB	0.2220	CRF + supervised linking	word	no

set, the least represented category is `Character` among the NIL instances and both `Thing` and `Event` between the linked ones. A different behaviour can be found in the test set where the least represented category is `Thing` in both NIL and linked instances.

Each participant was asked to submit up to three different runs and the evaluation was based on the following three metrics:

STMM (*Strong_Typed_Mention_Match*). This metrics evaluates the micro average F-1 score for all annotations considering the mention boundaries and their types. This is a measure of the tagging capability of the system.

SLM (*Strong_Link_Match*). This metrics is the micro average F-1 score for annotations considering the correct link for each mention. This is a measure of the linking performance of the system.

MC (*Mention_Ceaf*). This metrics, also known as Constrained Entity-Alignment F-measure (Luo 2005), is a clustering metric developed to evaluate clusters of annotations. It evaluates the F-1 score for both NIL and non-NIL annotations in a set of mentions.

The final score for each system is a combination of the aforementioned metrics and is computed as follows:

$$score = 0.4 \times MC + 0.3 \times STMM + 0.3 \times SLM. \quad (1)$$

All the metrics were computed by using the TAC KBP scorer[19].

2.3.2 Participating Systems and Results

The task was well received by the NLP community and was able to attract 17 expressions of interest. Five groups participated actively to the challenge by submitting their system results, each group presented three different runs, for a total amount of 15 runs submitted. Table 10 summarizes the methodology followed by each group and the best performance achieved by each participant.

The best result was reported by *UniPI* (Attardi et al. 2016b), while *MicroNeel.base* (Corcoglioniti et al. 2016) and *FBK-HLT-NLP* (Minard, Qwaider, and Magnini 2016) ob-

19 https://github.com/wikilinks/neleval/wiki/Evaluation

tain remarkable results very close to the best performance. It is interesting to notice that all these systems (*UniPI*, *MicroNeel* and *FBK-HLT-NLP*) developed specific techniques for dealing with Twitter mentions reporting very good results for the tagging metric (with values always above 0.46).

All participants have made used of supervised algorithms at some point of their tagging/linking/clustering pipeline. *UniPi*, *Sisinflab* and *UNIMIB* have exploited word embeddings trained on the development set plus some other external resources (manual annotated corpus, Wikipedia, and Twita). *UniPI* and *FBK-HLT-NLP* built additional training data obtained by active learning and manual annotation. The use of additional resources is allowed by the task guidelines, and both the teams have contributed to develop additional data useful for the research community.

2.3.3 Links to Other Shared Tasks

Previous EVALITA Tasks. This is the first edition of the NEEL-it task in EVALITA, however several tasks about Named Entity Recognition (NER) were organized in previous editions of EVALITA (Speranza 2007, 2009; Bartalesi Lenzi, Speranza, and Sprugnoli 2013). The noisy nature and shortness of tweets make the NER task more difficult, in fact we report a mention ceaf of about 59%. During EVALITA 2007 and 2009 the best performance was about 82%, while in EVALITA 2011 the best performance was 61%. It is important to underline the in EVALITA 2011 the task was based on automatic transcription of broadcast news.

Non EVALITA Tasks. The #Microposts (Making Sense of Microposts) workshop organizes a NEEL challenge since 2014. The NEEL-it task is inspired by #Microposts and fellows its guidelines and annotation schema. Other tasks related to entity linking in tweets are the Knowledge Base Population (KBP2014) Entity Linking Track[20] and the Entity Recognition and Disambiguation Challenge (ERD 2014) (Carmel et al. 2014). It is important to underline that the ERD challenge is not focused on tweets but the short text track is performed in the context of search engine queries.

2.4 FactA

2.4.1 Task Description, Data and Evaluation Metrics

FactA (*Event Factuality Annotation*)[21] aims at evaluating the automatic assignment of factuality values to events (Minard, Speranza, and Caselli 2016). In this task, factuality is defined as the committed belief expressed by relevant sources, either the utterer of direct and reported speech or the author of the text, towards the status of an event (Saurí and Pustejovsky 2012), i.e. a situation that happen or occur but also a state or a process (Pustejovsky et al. 2003). Specific linguistic markers and constructions help identifying the factuality of an event that can be classified according to 5 values: `factual`, `counterfactual`, `non-factual`, `underspecified` and `no factuality`. This classification is assigned by combining the values given to three attributes namely, *Polarity*, *Certainty* and *Time*. The first attribute expresses how sure the source is about the event, the second distinguishes future and underspecified events from all the others, and the third attribute specifies whether an event is affirmed or negated (Tonelli et al. 2014). The following example shows the factuality annotation of the event *usciti*/"went out".

20 `http://nlp.cs.rpi.edu/kbp/2014/`
21 `http://facta-evalita2016.fbk.eu`

Probabilmente i ragazzi sono usciti di casa tra le 20 e le 21. ("The guys went probably out between 8 and 9 pm.")

- *Certainty*: `no_certain`
- *Time*: `past/present`
- *Polarity*: `positive`
- *Factuality Value*: `non-factual`

FactA is organised around a Main Task focusing on the newswire domain and a Pilot Task dedicated to a particular type of social media texts, i.e., tweets. The training set of the Main Task consists of 169 news of the Fact-ItaBank corpus (Minard, Marchetti, and Speranza 2014) while the test set is made of 120 Wikinews articles taken from the Italian section of the NewsReader MEANTIME corpus (Speranza and Minard 2015). As for the Pilot Task, the idea was to measure how well systems built for standard language perform on new text types, like tweets. For this reason only a test set of 301 tweets is provided and corresponds to a subsection of the EVALITA 2016 SENTIPOLC task (Barbieri et al. 2016). An official score and a baseline system are defined: the first one calculates the micro-average F1 score (corresponding to the accuracy) on the overall factuality value and also on the single attributes while the baseline system assigns the most frequent value per attribute on the basis of its frequencies in the training data (that is, `certain`, `positive` and `past/present`) thus all events are annotated as `factual`.

2.4.2 Participating Systems and Results

Although 13 teams registered for the task, none of them took part in it by submitting a run. However, after the official evaluation, the system developed by one of the organisers was tested on both Main and Pilot data. The system, called FactPro, performs multi-class classification: three classifiers, one for each factuality attribute, are build using a Support Vector Machines algorithm exploiting lexical, syntactic and semantic features plus manually defined trigger lists (e.g., list of linguistic particles that are polarity markers). FactPro outperforms the baseline on both datasets reaching 0.72 F1 in the assignment of the factuality value on news (+ 0.5 points with respect to the baseline) and 0.56 F1 on tweets (+ 0.9 points with respect to the baseline). The first result can be compared with the performances of *De Facto*, the tool developed by Saurí and Pustejovsky (2012) that achieves an F1 of 0.80 (micro-averaging) and 0.70 (macro-averaging). Error analysis reveals four main source of errors: (i) the unbalanced distribution of some attribute values; (ii) the semantic complexity of some sentences; (iii) the incompleteness of the trigger lists; (iv) the analysis of nominal events. As for tweets, the peculiarities of their language, such as their fragmentary style and the high frequency of imperatives and interrogatives, pose additional challenges to the task. Indeed, the accuracy in the classification of each attribute registered a consistent drop (*Polarity*:-0.13, *Certainty*: -0.17, *Time*: -0.14) on Twitter data with respect to news data.

2.4.3 Links to Other Shared Tasks

Previous EVALITA Tasks. FactA is the first evaluation task for factuality profiling of events in Italian thus it is not possible to made a direct comparison with any previous exercise organised in EVALITA. Anyway, FactA is strictly connected with the EVALITA 2014 EVENTI task (Caselli et al. 2014). EVENTI is the acronym of *EValuation of Events aNd Temporal Information* and its aim was to evaluate the performance of Temporal

Table 11
Performance of FactPro

		ACCURACY			
		polarity	certainty	time	Factuality Value
MAIN TASK	baseline	0.94	0.86	0.71	0.67
	FactPro	0.92	0.83	0.74	0.72
PILOT TASK	baseline	0.80	0.69	0.55	0.47
	FactPro	0.79	0.66	0.60	0.56

Information Processing systems on Italian texts. To this end, a corpus of news annotated with temporal expressions, events and temporal relations was released. This corpus is a revised and simplified version of Ita-TimeBank (Caselli et al. 2011) from which the training set of FactA was taken. In other words, EVENTI and FactA share the same broad notion of event as inherited from TimeML.

Non EVALITA Tasks. The task on Event Factuality Classification for Dutch was organised in the context of CLIN26, the 26th Meeting of Computational Linguistics in the Netherlands, in 2016. This task and the one run in EVALITA, shared in part the same type of data (i.e. Wikinews articles) and the same annotation guidelines. However, in CLIN26 participants were asked to classify only the certainty and polarity of events. Two groups developed rule-based systems and submitted results: the best system (RuGGed) obtained an F-score of 96.10 for certainty and of 88.20 for polarity (Minard et al. 2016). Other issues related to factuality such as subjectivity, hedging and modality, are instead the focus of other evaluation exercises. In 2005, the ACE Event Detection and Recognition Task[22] asked participants to distinguish between asserted, hypothetical, desired, and promised events by assigning the correct modality value. This evaluation covered different types of texts, such as news and weblogs, both in English and Chinese. The same languages, plus Spanish, are taken into consideration in the TAC KBP Event Tracks organised in 2015, 2016 and 2017[23]. In particular, the Event Nugget task is based on the Rich ERE Annotation Guidelines where each event has an attribute, called `Realis`, indicating whether or not that event occurred (Mitamura et al. 2015). Uncertainty, negation and speculation in the domain of biology have been addressed both in the CoNLL-2010 Shared Task (Farkas et al. 2010) and, since 2009, in the BioNLP evaluation (Kim et al. 2009) while the detection of modality of clinically significant events is one of the topic in the 2012 i2b2 challenge and the SemEval Clinical TempEval tasks in 2015, 2016 and 2017. The systems participating to these tasks adopt different machine learning approaches to detect clinical event modality (e.g., SVM and CRF) reaching an F1 of 0.86 both in the i2b2 challenge and in Clinical TempEval (Sun, Rumshisky, and Uzuner 2013; Bethard et al. 2016). In contrast, the detection of negation and speculation in BioNLP is still far from a level of practical use with the best system having an F1 below 0.30 (Kim, Wang, and Yasunori 2013).

22 `http://www.itl.nist.gov/iad/mig/tests/ace/2005/`
23 `https://tac.nist.gov/`

2.5 ArtiPhon

2.5.1 Task Description, Data, and Evaluation Metrics

In the *Articulatory Phone Recognition* (ArtiPhon) task (Badino 2016b), participants had to build a speaker-dependent phone recognition system. Train set is delivered in form of simple audio files and related orthographic transcriptions and, for each audio file, a series of articulatory data aligned with the acoustic stimulus is available. Artiphon task at EVALITA 2016 aimed at resolving two different dilemmas in speech sciences: the first one was to evaluate recognition systems trained with a speech corpus presenting a given speaking style (read speech in this case, where speaker were asked to keep a constant speech rate) and tested with a further corpus presenting a mismatched speaking style (test set ranged from slow and hyper-articulated speech to fast and hypo-articulated speech). Training and testing data were from the same speaker. The second goal, which motivates the presence of measured articulatory movements data in the train and test corpus, was to investigate to what extent the increase of representational power obtained by adding an articulatory feature set could help in building ASR systems that are more robust to the effects of the mismatch problem and to other noise source in ASR domain. Recently, Badino (2016a) and Badino et al. (2016) have proposed an "articulatory" ASR based on deep neural networks (DNNs) extending also to this specific domain the influence that DNN are having in many other fields of speech technologies. Task organizers also made available to participants a set of tools to apply DNNs to this challenge, please see the original paper (Badino 2016b) for further details on this. The evaluation metrics here used are taken from the SCLITE method, based on Levenstein distance evaluation algorithm and contained in the SCTK Toolkit[24]. In particular, results are expressed in terms of phoneme correct rate, percentage of substitutions, insertions, deletion, and average phone error rate (PER henceforth).

2.5.2 Participating Systems and Results

Only one out of the 6 research groups that expressed an interest in the task actually participated. This system, henceforth *ISTC* (Cosi 2016), did not make use of articulatory data. Two sets of results are presented for the evaluation of this system. In the first one, the system was considered as speaker-dependent as both training and test datasets were recorded by the same speaker. In the second set of runs, *ISTC* was trained using a different training corpus and tested on the ArtiPhon test set, thus producing a speaker independent answer set. ISTC adapted for Italian the KALDI ASR toolkit[25] which contains two different DNNs, namely Dan's and Karel's (Veselỳ et al. 2013). For both answer sets, a variety of features, feature combinations, and processing methods were used, giving rise to a complex scenario of possible evaluations. Obviously, results obtained in the case of Speaker Dependent sessions are higher than those obtained in the Speaker Independent one. In the former case, PER was brought down to 15.1% when a combination of Gaussian Mixture Model (GMM) and DNN (Dan's) methods were used. In the Speaker Independent case, the combination GMM+DNN (Dan's) also performs well in terms of PER, however, the absolute best result is obtained with Karel's DNN (PER at 26.9%). A direct comparison between these results and the baseline obtained by the organiser is not simple because the set of phones used in the evaluation is different (i.e., greater) from that used by *ISTC*, furthermore the task organiser presented his results considering performances across speaking style (from hyper- to hypo- articulated

24 https://www.nist.gov/itl/iad/mig/tools
25 http://kaldi-asr.org/

speech). This said, the best performance obtained by the organiser has a PER of 23.5% on hyper-articulated speech using both acoustic and actual articulatory features.

2.5.3 Links to Other Shared Tasks

Previous EVALITA Tasks. ASR and speech technologies related tasks have been organised within EVALITA since its second edition in 2009. Tasks ranged from continuous digit recognition to forced alignment (Matassoni, Brugnara, and Gretter 2013; Cutugno, Origlia, and Seppi 2013; Cosi et al. 2014). In these cases, however, the evaluation exercises were approached in a more classic way, with the ultimate aim of establishing a benchmark that before the EVALITA times had never been available. The speaker independence dimension introduced at ArtiPhon this year, constitutes a novelty in speech tasks for Italian.

Non EVALITA Tasks. In recent years, ASR international evaluation campaigns have become sparser [26], and focused more on applications[27] than on ASR in itself. In particular, a great attention is now given to new domains such as emotion recognition and speech measurements for affective computing[28], elimination of reverberation[29], speech analytics[30], automatic spoken translation [31]. ASR involving articulatory features is still a developing field, and the evaluation is made difficult by the overall limited availability of annotated training and test sets.

2.6 QA4FAQ

2.6.1 Task Description, Data and Evaluation Metrics

The goal of this task is to develop a system retrieving a list of relevant FAQs and corresponding answers related to a query issued by an user (Caputo et al. 2016). For defining an evaluation protocol, we need a set of FAQs, a set of user questions and a set of relevance judgements for each question. In order to collect these data, we exploit an application called *AQP Risponde*, developed by QuestionCube for the Acquedotto Pugliese. A*QP Risponde* provides a back-end that allows to analyze both the query log and the customers' feedback to discover, for instance, new emerging problems that need to be encoded as FAQ. AQP received about 25,000 questions and collected about 2,500 user feedback. We rely on these data to build the dataset for the task. In particular, we provide:

- a knowledge base of 406 FAQs. Each FAQ is composed of a question, an answer, and a set of tags;
- a set of 1,132 user queries. The queries are collected by analysing the *AQP Risponde* system log. From the initial set of queries, we removed queries that contains personal data;
- a set of 1,406 pairs $< query, relevant faq >$ that are exploited to evaluate the contestants. We build these pairs by analysing the user feedback provided by real users of *AQP Risponde*. We manually check the user feedback in order to remove

26 https://www.nist.gov/itl/iad/mig/past-hlt-evaluation-projects
27 See, for example, https://asru2017.org/Challenges.asp
28 http://emotion-research.net/sigs/speech-sig/is16-compare
29 https://reverb2014.dereverberation.com/
30 https://www.nist.gov/itl/iad/mig/nist-2017-pilot-speech-analytic-technologies-evaluation
31 http://www.iwslt.org

noisy or false feedback. The check was performed by two experts of the AQP customer support.

We provided a little sample set for the system development and a test set for the evaluation. We did not provide a set of training data: AQP is interested in the development of unsupervised systems because *AQP Risponde* must be able to achieve good performance without any user feedback. Following, an example of FAQ is reported:

Question "Come posso telefonare al numero verde da un cellulare?" *How can I call the toll-free number by a mobile phone?*

Answer "È possibile chiamare il Contact Center AQP per segnalare un guasto o per un pronto intervento telefonando gratuitamente anche da cellulare al numero verde 800.735.735. Mentre per chiamare il Contact Center AQP per servizi commerciali 800.085.853 da un cellulare e dall'estero è necessario comporre il numero +39.080.5723498 (il costo della chiamata è secondo il piano tariffario del chiamante)." *You can call the AQP Contact Center to report a fault or an emergency call without charge by the phone toll-free number 800 735 735...*

Tags *canali, numero verde, cellulare*

For example, the previous FAQ is relevant for the query: "Si può telefonare da cellulare al numero verde?" *Is it possible to call the toll-free number by a mobile phone?* FAQs are provided in both XML and CSV format using ";" as separator. The file is encoded in UTF-8 format. Each FAQ is described by the following fields:

id a number that uniquely identifies the FAQ
question the question text of the current FAQ
answer the answer text of the current FAQ
tag a set of tags separated by ","

Test data are provided as a text file composed by two strings separated by the *TAB* character. The first string is the user *query id*, while the second string is the text of the user query. For example: "1 Come posso telefonare al numero verde da un cellulare?" and "2 Come si effettua l'autolettura del contatore?".

Moreover, we provided a simple baseline based on a classical information retrieval model. The baseline is built by using Apache Lucene (ver. 4.10.4)[32]. During the indexing for each FAQ, a document with four fields (*id, question, answer, tag*) is created. For searching, a query for each question is built taking into account all the question terms. Each field is boosted according to the following score *question=4*, *answer=2* and *tag=1*. For both indexing and search the *ItalianAnalyzer* is adopted. The top 25 documents for each query are provided as result set. The baseline is freely available on GitHub[33] and it was released to participants after the evaluation period.

The participants must provide results in a text file. For each query in the test data, the participants can provide 25 answers at the most, ranked according by their systems. Each line in the file must contain three values separated by the TAB character: $< queryid >< faqid >< score >$.

Systems are ranked according to the accuracy@1 (c@1). We compute the precision of the system by taking into account only the first correct answer. This metric is used for the final ranking of systems. In particular, we take into account also the number of

32 http://lucene.apache.org/
33 https://github.com/swapUniba/qa4faq

Table 12
QA4FAQ results in terms of accuracy with details about the main characteristics of the participating systems

TEAM ID	c@1	METHOD	EMBEDDINGS	OTHER RESOURCES
chiLab4It	0.4439	QuASIt (cognitive model)		Wiktionary
baseline	*0.4076*	*Lucene BM25*		*no*
fbk4faq	0.3746	Vector similarity	word	no
NLP-NITMZ	0.2125	VSM + Apache Nutch		no

unanswered questions, following the guidelines of the CLEF ResPubliQA Task (Peñas et al. 2009). The formulation of c@1 is:

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n}) \tag{2}$$

where n_R is the number of questions correctly answered, n_U is the number of questions unanswered, and n is the total number of questions.

The system should not provide result for a particular question when it is not confident about the correctness of its answer. The goal is to reduce the amount of incorrect responses, keeping the number of correct ones, by leaving some questions unanswered. Systems should ensure that only the portion of wrong answers is reduced, maintaining as high as possible the number of correct answers. Otherwise, the reduction in the number of correct answers is punished by the evaluation measure for both the answered and unanswered questions.

2.6.2 Participating Systems and Results

Thirteen teams registered in the task, but only three of them submitted the results for the evaluation. A short description of each system with its best performance is reported in Table 12.

All the systems adopt different strategies, while only one system (*chiLab4It*) is based on a typical question answer module. The best performance is obtained by the *chilab4it* (Pipitone, Tirone, and Pirrone 2016) team, that is the only one able to outperform the baseline. Moreover, the *chilab4it* team is the only one that exploits question answering techniques: the good performance obtained by this team proves the effectiveness of question answering in the FAQ domain. All the other participants had results under the baseline. Another interesting outcome is that the baseline exploiting a simple VSM model achieved remarkable results.

In (Fonseca et al. 2016), the authors have built a custom development set by paraphrasing original questions or generating a new question (based on original FAQ answer), without considering the original FAQ question. The interesting result is that their system outperformed the baseline on the development set. The authors underline that the development set is completely different from the test set which contains sometime short queries and more realistic user's requests. This is an interesting point of view since one of the main challenge of our task concerns the variety of language expressions adopted by customers to formulate the information need.

Previous EVALITA Tasks. No other tasks related to QA4FAQ was organized in the previous editions of EVALITA.

Non EVALITA Tasks. The QA4FAQ task is strongly related to the "Answer Selection in Community Question Answering" task recently organized in the context of Semeval 2015 and 2016 (Nakov et al. 2015). This task helps to automate the process of finding good answers to new questions in a community-created discussion forum (e.g., by retrieving similar questions in the forum and by identifying the posts in the answer threads of similar questions that answer the original one as well). Moreover, the QA4FAQ has some common points with the Textual Similarity task (Agirre et al. 2015) that received an increasing amount of attention in recent years.

2.7 Application Challenge

In 2016, for the first time, EVALITA included an application challenge organised by IBM Italy[34]. The aim of the challenge was to award the most innovative apps employing the services available on Bluemix, the platform of IBM APIs. More specifically, participants were required to build their apps on top of at least one the Watson's APIs for cognitive computing supporting the Italian language. Submitted systems were evaluated by a judge panel made by academics and IBM representatives by taking into consideration the creativity and viability of the use case, the intuitiveness of the user experience, the value of the app, the feasibility and uniqueness of the implementation.

The first three best submissions, described below and listed following their final ranking position, received a monetary prize.

1. **Stockle** is a sentiment analysis API and web application focused on the stock trade market. The APIs of AlchemyData News, Yahoo Finance, reddit and Twitter are used to retrieve comments, tweets and news related to a set of companies selected by the user and the sentiment towards these companies is analysed by the sentiment analysis module of Alchemy API[35].
2. **MARTIN** (*Monitoring and Analysing Real-time Tweets in Italian Natural language*)[36] is a stand-alone application that allows to scan real-time information on Twitter, compare tweets by pairs of Twitter accounts, analyse the language of tweets and visualize the output of these analyses. To this end, Twitter APIs are combined with the natural language understanding modules for sentiment analysis and keyword extraction provided by Alchemy APIs.
3. **Appetitoso ChatBot** is a dialog system connected to the mobile application and the website of Appetitoso[37], a recommendation service that searches for restaurants on the basis of the dishes desired by the user. The IBM Watson Conversation Module is used to create a conversation between the user and the application through an exchange of written messages.

34 http://www.evalita.it/2016/tasks/ibm-challenge
35 https://gingerbeard.alwaysdata.net/stockle/
36 https://dh.fbk.eu/technologies/martin
37 http://www.appetitoso.it/

Table 13
Number of tweets of cross-task shared data. A * indicates instead the number of sentences from newswire documents.

TRAIN				
	SENTIPOLC	NEEL-it	PoSTWITA	FactA
SENTIPOLC	7410	989	6412	0
NEEL-it	989	1000	0	0
PoSTWITA	6412	0	6419	0
FactA	0	0	0	2723*
TEST				
	SENTIPOLC	NEEL-it	PoSTWITA	FactA
SENTIPOLC	2000	301	301	301
NEEL-it	301	301	301	301
PoSTWITA	301	301	301	301
FactA	301	301	301	597*+301

3. Cross-task Shared Data

One of the greatest benefits of evaluation campaigns is the creation of benchmark data for a variety of tasks. This requires quite some effort on the part of the organisers, with the development of guidelines and, mostly, with manual annotation towards the creation of gold standard sets. One little exploited advantage of such data creation efforts is the possibility of adding layers of annotation related to different phenomena *over exactly the same data*, so as to facilitate and promote the development and testing of end-to-end systems (Basile et al. 2015). With this in mind, we encouraged task organisers to share datasets so as to annotate the same instances, each task with their respective layer. The involved tasks were SENTIPOLC, PoSTWITA, NEEL-it and FactA. In this Section we provide an overview of the shared data in terms of number of instances and an example of how the different annotations of the same data look like like over a sample tweet.

The matrix in Table 13 shows both the total number of test instances per task (diagonally) as well as the number of overlapping instances for each task pair. Please note that while the datasets of SENTIPOLC, NEEL-it, and PoSTWITA were composed entirely of tweets, both as training and test data, FactA included tweets only in one of their test set, as a pilot task. FactA's training and standard test sets are composed of newswire data, which we report in terms of number of sentences. For this reason the number of instances in Table 13 is broken down for FactA's test set: 597 newswire sentences and 301 tweets, the latter being the same as the other tasks.

This first attempt at creating shared data across tasks was completely successful in terms of test data, as the testsets for all four tasks comprise *exactly* the same 301 tweets, although for SENTIPOLC and FactA this is only a portion of a larger test set.

Regarding the training sets, which are obviously larger, there are overlaps, but these are not complete. Specifically, the training sets of PoSTWITA and NEEL-it are almost entirely subsets of SENTIPOLC. In addition, 989 tweets from the 1000 that make NEEL-it's training set are in SENTIPOLC, and 6412 of PoSTWITA (out of 6419) also are

Table 14
Annotation of the tweet *@juventusfc E come se vogliamo vincerla, forza ragazzi!!!!!!!*, with id 601071129810309120

FactA

id	begin	end	element	polarity	certainty	time
601...	31	39	EVENT	POS	NON_CERTAIN	FUTURE

NEEL-it

id	begin	end	link	type
601...	1	11	http://dbpedia.org/resource/Juventus_F.C.	Organization

SENTIPOLC

id	subj	opos	oneg	iro	lpos	lneg
601...	1	1	0	0	1	0

PoSTWITA

601...	
@juventusfc	MENTION
E	CONJ
come	ADP
se	SCONJ
vogliamo	AUX
vincerla	VERB_CLIT
,	PUNCT
forza	NOUN
ragazzi	NOUN
!!!!!!!	PUNCT

included in the SENTIPOLC training set, and only the training data the SENTIPOLC shares with NEEL-it is not included in PoSTWITA.

In Table 14 we show how the very same tweet – *@juventusfc E come se vogliamo vincerla, forza ragazzi!!!!!!!* (with id 601071129810309120, from the EVALITA 2016 distribution) – has been annotated according to the guidelines of the four tasks.

Currently, the annotations are overlapping but are still encoded on separate, unconnected files in practice. In the next future we plan to develop and share specific standards and tools that will allow for such annotations to be practically linked and knit together, so that current and future single annotations for different phenomena over the same data will be exploited simultaneously.

We believe the cross-task data produced within EVALITA 2016 is an excellent starting point towards making more data that is enriched with as many layers of annotation as possible, especially related to the EVALITA shared tasks. In order to further promote this, we have also set up a repository to collect and keep track of data creation and development for Italian. All 2016 datasets are available online on the github account of EVALITA 2016: `https://github.com/evalita2016/data`.

4. The EVALITA Community

The tasks and the challenge of EVALITA 2016 attracted the interest of a large number of researchers, for a total of 96 single registrations. Overall, 34 teams composed of more than 60 individual participants submitted their results to one or more different tasks of the campaign. A breakdown of the figures per task is shown in Table 15. As for the

members of the teams, they were affiliated to more than 20 different institutions, 4% belong to private companies and 26% work outside Italy [38].

Table 15
Registered and actual participants of EVALITA 2016

task	registered	actual
ARTIPHON	6	1
FactA	13	0
NEEL-IT	16	5
QA4FAQ	13	3
PoSTWITA	18	9
SENTIPOLC	24	13
IBM Challenge	6	3
total	96	34

With respect to the 2014 edition, we collected a significantly higher number of registrations (96 registrations *vs* 55 registrations collected in 2014), which can be interpreted as a signal that we succeeded in reaching a wider audience of researchers interested in participating in the campaign. This result could be also be positively affected by the novelties introduced in this edition to improve the dissemination of information on EVALITA, e.g. the use of social media such as Twitter and Facebook. Also the number of teams that actually submitted their runs increased in 2016 (34 teams *vs* 23 teams participating in the 2014 edition), even if we reported a substantial gap between the number of actual participants and those who registered. In order to better investigate this issue and gather some insights on the reasons of the significant drop in the number of participants w.r.t. the registrations collected, we ran an online questionnaire specifically designed for those who did not submit any run to the task to which they were registered. In two weeks we collected 14 responses which show that the main obstacles to the actual participation in a task were related to personal issues ("I had an unexpected personal or professional problem outside EVALITA" or "I underestimated the effort needed") or personal choices ("I gave priority to other EVALITA tasks"). As for this last point, NEEL-it and SENTIPOLC were preferred to FactA, which did not have any participant. Another problem mentioned by some of the respondents is that the evaluation period was too short: this issue is highlighted mostly by those who registered to more than one task. However, the gap between registered and actual participants is not new for EVALITA but affected also all the previous editions of the campaign as shown in Figure 2. Another general trend in EVALITA is the high percentage of new participants (always above 60%): in particular, Figure 3 highlights that in 2016 70% of the members of participating teams were at their first experience in the campaign.

Twenty-five researchers took part in the task organization: 16% were affiliated to foreign institutions and 20% were representatives of private companies. This last percentage demonstrates that it is easier to involve industrial companies in the organization of tasks than in the participation. This can be due to some reluctance on the part of companies to expose themselves to potential bad publicity if resulting in the lower portions of the ranks. Organizers had 18 affiliations and all tasks had at least two organizers from two different institutions. This indicates that organizing a task is a

38 In Brazil, France, Germany, India, Ireland, Mexico, The Netherlands, Spain, and Switzerland.

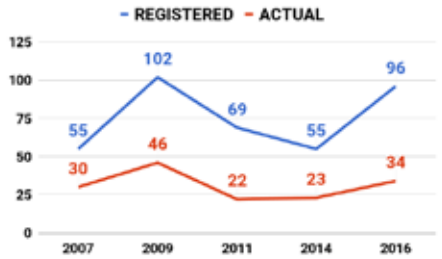

Figure 2
Registered and actual participants of EVALITA campaigns

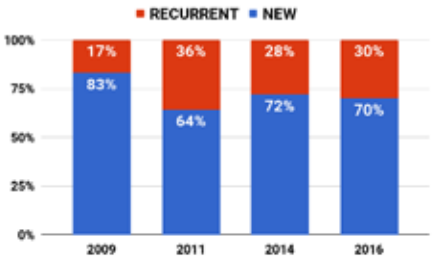

Figure 3
Percentage of new *versus* recurrent participants in EVALITA

great way to boost the cooperation as shown in Figure 4 where node colour represents affiliations while edge colour indicates tasks[39].

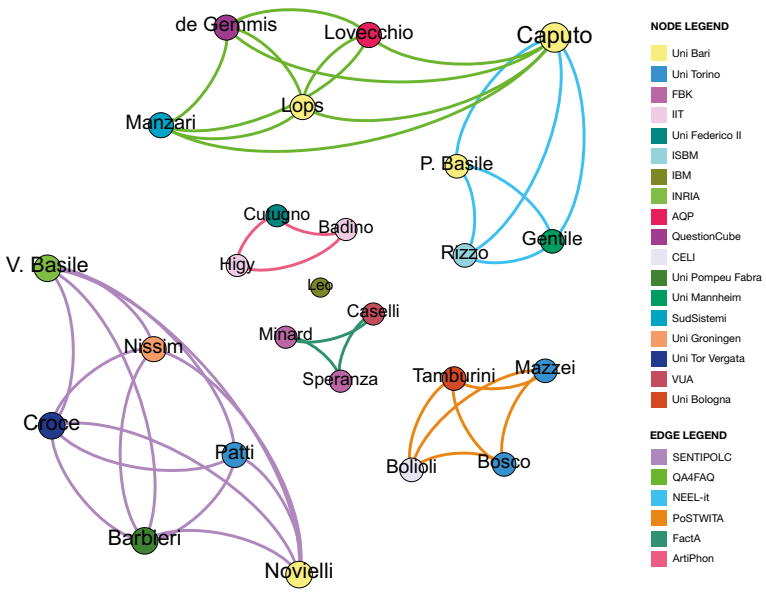

Figure 4
Co-organisation of EVALITA 2016 tasks. Node colours characterise affiliations, so that same colour means same institution.

5. Conclusions

EVALITA 2016 was a successful edition, surely in terms of participation and obtained results, but also in terms of data creation and dissemination, especially with the now available github repository, and in terms of collaboration across tasks.

39 A video showing a dynamic network of task organisers in EVALITA from 2007 to 2016 is available online: `http://www.evalita.it/EVALITAcommunity`

As for future editions, we suggest to prospective EVALITA chairs to continue on the path started in 2016 by working towards an always stronger involvement of representatives from companies in the organisation of tasks, a balance between research and application tasks, and an ever-increasing development of shared and open data. These three aspects proved to be useful to boost the cooperation between different private and public institutions and to attract new researchers in the EVALITA community.

Moreover, social media texts turned out to be a very attractive domain, and there is still plenty to be explored in this domain. Even within Twitter, sampling data using different strategies has proved potentially challenging for systems, especially for some tasks (e.g PoSTWITA), so that future editions could involve some more explicit domain adaptation tasks, still in the general domain of social media (see for example what has been done for author profiling at PAN 2016 (Rangel et al. 2016)). Obviously, domains other than social media could be explored as well. For instance, Humanities resulted as one of the most appealing domains in the questionnaires for industrial companies and former participants and other countries are organising evaluation exercises on it (see, for example, the *Translating Historical Text* shared task at CLIN 27[40]).

Other innovations can be envisaged for the next campaigns, too, also from an organisational perspective. For example, different evaluation windows for different tasks could be planned instead of having the same evaluation deadlines for all. This flexibility would have an impact on the work load of the EVALITA's organisers but, on the other side, would help teams to participate in multiple tasks without making them choose to concentrate their effort only on one task due to lack of time.

As in the past editions, EVALITA 2016 served as the optimal forum for the creation and discussion of the most challenging tasks for Italian NLP. Additionally, collaboration between task organisers and between academia and industry was more fruitful than ever. We hope that this kind of active and open cooperation will be brought forward in future editions, too, and that the repository of shared data that was created in the context of the 2016 edition will continue to be populated, so as to form the reference benchmark for Italian data in a variety of Natural Language Processing tasks.

Acknowledgements

EVALITA 2016 would not have been possible without the invaluable work of those who proposed and organised the tasks, without the participants, and without the support of AILC and especially IBM, who organised and sponsored the challenge included in this edition. We are enormously grateful to all of them. We also would like to thank the organisers of CLiC-it 2016 for hosting the final workshop in Naples, co-located with the conference, and Walter Daelemans for being our invited speaker. Finally, reviewers provided insightful comments which we have benefited from in the final version of this paper.

References

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, USA, June 4-5.

40 http://www.ccl.kuleuven.be/CLIN27/

Attardi, Giuseppe, Daniele Sartiano, Chiara Alzetta, and Federica Semplici. 2016a. Convolutional neural networks for sentiment analysis on italian tweets. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Attardi, Giuseppe, Daniele Sartiano, Maria Simi, and Irene Sucameli. 2016b. Using Embeddings for Both Entity Recognition and Linking in Tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Attardi, Giuseppe and Maria Simi. 2016. Character Embeddings PoS Tagger vs HMM Tagger for Tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Badino, Leonardo. 2016a. Phonetic Context Embeddings for DNN-HMM Phone Recognition. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016): Understanding Speech Processing in Humans and Machines*, San Francisco, California, USA, September 8-12.

Badino, Leonardo. 2016b. The ArtiPhon Challenge at Evalita 2016. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Badino, Leonardo, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. 2016. Integrating articulatory data in deep neural network-based acoustic modeling. *Computer Speech & Language*, 36:173–195.

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Bartalesi Lenzi, Valentina, Manuela Speranza, and Rachele Sprugnoli. 2013. EVALITA 2011: Description and Results of the Named Entity Recognition on Transcribed Broadcast News Task. In *Evaluation of Natural Language and Speech Tools for Italian. Revised Selected Papers of the EVALITA 2011 International Workshop*, pages 86–97. Springer. Rome, Italy, January 24-25, 2012.

Basile, Pierpaolo, Valerio Basile, Malvina Nissim, and Nicole Novielli. 2015. Deep Tweets: from Entity Linking to Sentiment Analysis. In Cristina Bosco, Sara Tonelli, and Fabio Massimo Zanzotto, editors, *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 41–46, Trento, Italy, December 3-4. aAccademia University Press.

Basile, Pierpaolo, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro. 2015. Entity Linking for Italian Tweets. In Cristina Bosco, Sara Tonelli, and Fabio Massimo Zanzotto, editors, *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 36–40, Trento, Italy, December 3-4. aAccademia University Press.

Basile, Valerio, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop*

EVALITA 2014, pages 50–57, Pisa, Italy, 9-11 December. Pisa University Press.

Basile, Valerio 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, Atlanta, Georgia, June. Association for Computational Linguistics.

Beißwenger, Michael, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 44–56, Berlin, Germany, August 7-12.

Bernardi, Raffaella, Andrea Bolognesi, Corrado Seidenari, and Fabio Tamburini. 2005. Automatic induction of a POS tagset for Italian. In *Proceedings of the Australasian Language Technology Workshop*, pages 176–183, Sydney, Australia, 10-11 December.

Bethard, Steven, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, June. Association for Computational Linguistics.

Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems, Special Issue on Knowledge-based Approaches to Content-level Sentiment Analysis*, 28(2):55–63.

Caputo, Annalina, Marco de Gemmis, Pasquale Lops, Franco Lovecchio, and Vito Manzari. 2016. Overview of the EVALITA 2016 Question Answering for Frequently Asked Questions (QA4FAQ) Task. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Carmel, David, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June Paul Hsu, and Kuansan Wang. 2014. ERD'14: entity recognition and disambiguation challenge. In *ACM SIGIR Forum*, volume 48, pages 63–77. ACM.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Portland, Oregon, USA, June 23-24. Association for Computational Linguistics.

Caselli, Tommaso, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, pages 27–34, Pisa, Italy, 9-11 December. Pisa University Press.

Castellucci, Giuseppe, Danilo Croce, and Roberto Basili. 2016. Context-aware convolutional neural networks for Twitter sentiment analysis in Italian. 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)*, Naples, Italy, December 5-7. aAccademia University Press.

Cimino, Andrea and Felice Dell'Orletta. 2016. Building the state-of-the-art in POS tagging of Italian Tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Corcoglioniti, Francesco, Alessio Palmero Aprosio, Yaroslav Nechaev, and Claudio Giuliano. 2016. MicroNeel: Combining NLP Tools to Perform Named Entity Detection and Linking on Microposts. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Cosi, Piero. 2016. Phone Recognition Experiments on ArtiPhon with KALDI. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December

5-7. aAccademia University Press.
Cosi, Piero, Vincenzo Galatà, Francesco Cutugno, and Antonio Origlia. 2014. Forced Alignment on Children Speech. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, pages 124–126, Pisa, Italy, 9-11 December. Pisa University Press.
Cutugno, Francesco, Antonio Origlia, and Dino Seppi. 2013. EVALITA 2011: Forced alignment task. In *Evaluation of Natural Language and Speech Tools for Italian*. Springer, pages 305–311.
Di Rosa, Emanuele and Alberto Durante. 2016. Tweet2Check evaluation at Evalita Sentipolc 2016. 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)*, Naples, Italy, December 5-7. aAccademia University Press.
Farkas, Richárd, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 1–12, Uppsala, Sweden, 15-16 July. Association for Computational Linguistics.
Fonseca, Erick R., Simone Magnolini, Anna Feltracco, Mohammed R. H. Qwaider, and Bernardo Magnini. 2016. Tweaking Word Embeddings for FAQ Ranking. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.
Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and Jhon Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–475, Denver, Colorado, USA, June 4-5.
Horsmann, Tobias and Torsten Zesch. 2016. Building a Social Media Adapted PoS Tagger Using FlexTag–A Case Study on Italian Tweets. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. aAccademia University Press.
Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9, Boulder, Colorado, USA, June 4-5. Association for Computational Linguistics.
Kim, Jin-Dong, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria, August 8-9. Association for Computational Linguistics.
Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, B.C., Canada, October 6-8. Association for Computational Linguistics.
Matassoni, Marco, Fabio Brugnara, and Roberto Gretter. 2013. EVALITA 2011: Automatic Speech Recognition Large Vocabulary Transcription. In *Evaluation of Natural Language and Speech Tools for Italian. Revised Selected Papers of the EVALITA 2011 International Workshop*, pages 274–285. Springer. Rome, Italy, January 24-25, 2012.
Minard, Anne-Lyse, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Italy, 9-10 December.
Minard, Anne-Lyse, R. H. Mohammed Qwaider, and Bernardo Magnini. 2016. FBK-NLP at NEEL-IT: Active Learning for Domain Adaptation. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.
Minard, Anne-Lyse, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 Event Factuality Annotation Task (FactA). In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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), Naples, Italy, December 5-7. aAccademia University Press.

Minard, Anne-Lyse, Manuela Speranza, Marieke van Erp, Antske Fokkens, Marten Postma, Piek Vossen, Eneko Agirre, Itziar Aldabe, German Rigau, and Ruben Urizar. 2016. Evaluation tasks in open competitions Deliverable D10. 4. Technical report, NewsReader EU Project.

Mitamura, Teruko, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event nugget annotation: Processes and issues. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 66–76, Denver, Colorado, USA, May 31 - June 5.

Monachini, M. 1995. ELM-IT: An Italian incarnation of the EAGLES-TS. definition of lexicon specification and classification guidelines. Technical report.

Nakov, Preslav, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 Task 3: Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281, Denver, Colorado, June. Association for Computational Linguistics.

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, pages 380–390, Atlanta, Georgia, USA, June, 9-15. Association for Computational Linguistics.

Paci, Giulio. 2016. Mivoq EVALITA 2016 PosTwITA tagger. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Peñas, Anselmo, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forăscu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. 2009. Overview of ResPubliQA 2009: question answering evaluation over European legislation. In *Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum*, pages 174–196, Corfu, Greece, 30 September - 2 October. Springer.

Pipitone, Arianna, Giuseppe Tirone, and Roberto Pirrone. 2016. ChiLab4It System in the QA4FAQ Competition. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Plank, Barbara and Malvina Nissim. 2016. When silver glitters more than gold: Bootstrapping an Italian part-of-speech tagger for Twitter. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Pustejovsky, James, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Rangel, Francisco, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, pages 750–784, Évora, Portugal, 5-8 September.

Reyes, Antonio and Paolo Rosso. 2014. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. *Knowledge and Information Systems*, 40(3):595–614.

Rizzo, Giuseppe, Marieke van Erp, Julien Plu, and Raphaël Troncy. 2016. Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. In *Proceeding of the 6th Workshop on Making Sense of Microposts (#Microposts2016) co-located with*

WWW 2016, Montreal, Canada, 11 April.
Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, SemEval '2015, Denver, Colorado, June.
Rosenthal, Sara, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proc of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August.
Russo, Irene and Monica Monachini. 2016. Samskara minimal structural features for detecting subjectivity and polarity in italian tweets. 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)*, Naples, Italy, December 5-7. aAccademia University Press.
Saurí, Roser and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.
Speranza, Manuela. 2007. EVALITA 2007: The Named Entity Recognition Task. *Intelligenza Artificiale*, 4(2):66–68. Proceedings of the EVALITA 2007 Final Workshop, Rome, Italy, September 10.
Speranza, Manuela. 2009. The Named Entity Recognition Task at EVALITA 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December 9-12.
Speranza, Manuela and Anne-Lyse Minard. 2015. Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC). In Cristina Bosco, Sara Tonelli, and Fabio Massimo Zanzotto, editors, *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 252–257, Trento, Italy, December 3-4. aAccademia University Press.
Sprugnoli, Rachele, Viviana Patti, and Franco Cutugno. 2016. Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.
Stemle, Egon W. 2016. bot. zen@ EVALITA 2016-A minimally-deep learning PoS-tagger (trained for Italian Tweets). In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.
Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Avignon, France. Association for Computational Linguistics.
Stranisci, Marco, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2892–2899, Portoroz, Slovenia, 23-28 May. ELRA.
Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.
Tamburini, Fabio. 2007. EVALITA 2007: The Part-of-Speech tagging task. *Intelligenza artificiale*, 4(2):57–73.
Tamburini, Fabio. 2016. A BiLSTM-CRF PoS-tagger for Italian tweets using morphological information. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.
Tamburini, Fabio, Cristina Bosco, Alessandro Mazzei, and Andrea Bolioli. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In Pierpaolo Basile, Anna Corazza,

Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*, Naples, Italy, December 5-7. aAccademia University Press.

Tonelli, Sara, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. Newsreader guidelines for annotation at document level. Technical report, Fondazione Bruno Kessler.

Veselỳ, Karel, Arnab Ghoshal, Lukás Burget, and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013): Speech in Life Sciences and Human Societies*, pages 2345–2349, Lyon, France, 25-29 August.