Further Topics Emerging at the Fourth Italian Conference on Computational Linguistics

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC - CNR

Il secondo numero del quarto anno della rivista *Italian Journal of Computational Linguistics* (*IJCoL*), la rivista italiana promossa dall'*Associazione Italiana di Linguistica Computazionale* (AILC - www.ai-lc.it), costituisce un volume miscellaneo che completa la raccolta degli *Emerging Topics at the Fourth Italian Conference on Computational Linguistics* (n. 4, vol. 1, 2018) pubblicata nel primo volume dell'anno: in particolare, raccoglie un secondo insieme di lavori di ricerca che sono risultati particolarmente promettenti nell'ambito della Conferenza CLiC-it 2017, tenutasi a Roma dall'11 al 13 dicembre 2017. Questo insieme corrisponde a una selezione di contributi che documenta ricerca condotta in diversi ambiti, talora con interessanti ricadute applicative, che spaziano dallo studio dei Social Media, alla stilometria e alla musicologia, per arrivare a moderni approcci alla semantica lessicale e alla lessicografia. I temi affrontati sono alla base di recenti e interessanti sviluppi della linguistica computazionale, come ad esempio la Social Computational Science o le Digital Humanities.

Gli articoli sui temi emergenti sono stati, come sempre, selezionati attraverso un processo iterativo di *peer-review*. Ogni articolo è stato sottoposto a tre valutazioni da parte di comitati diversi: come contributo alla conferenza; come candidato ai premi di "Best Young Paper" e "Distinguished Young Paper" di CLiC-it 2017; infine, nella versione estesa, come articolo di rivista scientifica.

Aprono il volume due contributi che illustrano progettazione e costruzione di risorse linguistiche, lessicali e testuali. Il primo contributo di Micheli e Litta presenta una risorsa lessicale di vaste dimensioni per lo studio della lingua latina, denominata *Word Formation Latin* (WFL). Si tratta di un un lessico di forme derivate latine, all'interno del quale i lemmi sono segmentati nelle loro componenti formative e le relazioni tra di esse sono stabilite sulla base di regole di formazione di parole. Questa risorsa risponde all'attuale crescente interesse per lo studio della formazione delle parole, che presenta ripercussioni sui versanti teorico e applicativo. In particolare, il contributo si focalizza sulla rappresentazione dei composti all'interno di un lessico derivazionale, aspetto spesso trascurato in analoghe risorse recenti per altre lingue. L'originalità del contributo riguarda dunque anche il versante metodologico, ovvero le modalità di rappresentazione dei lemmi composti nella risorsa.

Nell'articolo di Cignarella e colleghi vengono presentati la progettazione e lo sviluppo di una risorsa testuale annotata con informazione relativa all'ironia, costituita da testi dei social media. Tale risorsa è stata sviluppata come parte del corpus Twitter TWITTIRÒ. L'annotazione è stata condotta secondo uno schema multi-livello già appli-

* Dept. of Enterprise Engineering - Via del Politecnico 1, 00133 Roma
E-mail: basili@info.uniroma2.it
** Istituto di Linguistica Computazionale "A. Zampolli", CNR - Via Moruzzi 1, 56124 Pisa
E-mail: simonetta.montemagni@ilc.cnr.it

cato a corpora di altre lingue (francese e inglese). L'esito della ricerca non è circoscritto al solo versante metodologico, ma include anche un nuovo corpus "gold" per la lingua italiana arricchito con informazione relativa all'ironia. Tale risorsa arricchisce l'attuale portafoglio di insiemi di dati multilingui disponibili per l'impegnativo compito di rilevamento dell'ironia, ed è pronta per essere usata come corpus annotato di riferimento all'interno di esperimenti e campagne di valutazione incentrate su di esso.

Segue il contributo di Pannitto e colleghi, incentrato sullo studio di metodi e tecniche per il riconoscimento automatico di relazioni di iperonimia con metodi di semantica distribuzionale. All'interno di tale paradigma sono stati proposti diversi modelli non supervisionati per l'identificazione di relazioni di iperonimia tra coppie di lemmi. Rispetto a questi, l'originalità dell'approccio proposto consiste nell'utilizzo di operazioni di algebra lineare applicate alle rappresentazioni vettoriali dei termini lessicali: tale approccio si basa su una variante della cosiddetta "distributional inclusion hypothesis" (DIH), che tiene in considerazione i vincoli pragmatici con cui la relazione di iperonimia è espressa nell'uso linguistico. L'ampia evidenza sperimentale riportata relativamente al dataset BLESS conferma che il metodo è in grado di fornire migliori prestazioni rispetto alla versione standard della DIH.

L'articolo di Della Moretta e colleghi riguarda una risorsa lessicale per la lingua italiana con strutture predicato-argomento, T-PAS (Typed Predicate-Argument Structure), di cui viene indagato l'uso nell'etichettatura semantica di testi. Il processo di annotazione semantica è condotto in due fasi, corrispondenti all'etichettatura dell'argomento e all'assegnazione del corrispondente tipo semantico di T-PAS. La metodologia di annotazione è stata testata all'interno di un esperimento pilota riguardante una selezione di verbi caratterizzati da diversi gradi di polisemia in T-PAS. Per verificare efficacia e affidabilità della metodologia proposta, un sottoinsieme del corpus è stato annotato da più persone e il livello di accordo tra gli annotatori è stato calcolato per ciascuna fase. Attraverso l'analisi dei casi di disaccordo tra annotatori sono stati identificati fenomeni e costruzioni particolarmente problematici da trattare e aree di miglioramento nelle specifiche e nel processo di annotazione.

Segue l'articolo di Basile e colleghi che discute le potenzialità degli approcci di apprendimento automatico nel riconoscere argomenti controversi all'interno di testi del Web come i post sui social media. Una batteria di modelli supervisionati basati su reazioni di Facebook sono proposti per la previsione delle polemiche generate da una notizia: l'approccio si basa sull'assunto che una controversia possa essere modellata basandosi sull'entropia della distribuzione delle reazioni a un post. Un corpus basato su Facebook, I-CONTRO, è presentato nell'articolo, sul quale è stata verificata la validità dell'approccio proposto e dei diversi modelli di riconoscimento di controversie.

Chiude il volume il contributo di Unguendoli e colleghi che illustra un metodo di attribuzione automatica dell'autore applicato al campo della musicologia. Vengono presentati i risultati dell'implementazione di una tecnica di attribuzione stilometrica automatica a un corpus di monodie liturgiche medievali. La natura rigorosamente lineare dei repertori musicali consente l'adozione di metriche di pseudo-distanza tra vettori di frequenza di n-grammi di simboli consecutivi. I risultati ottenuti suggeriscono che l'approccio quantitativo proposto può essere efficacemente utilizzato per supportare anche l'indagine di problemi più complessi in musicologia.

Questa breve sintesi non fa giustizia ai molti stimoli attivati dai lavori nel presente volume e alla loro variegata rassegna di metodologie e applicazioni innovative del trattamento automatico della lingua. Lasciamo come sempre al lettore il piacere di approfondirli direttamente nell'interezza di questo volume.

1. Editorial Note Summary

It is with great pleasure that we introduce the second volume of the fourth year of the *Italian Journal of Computational Linguistics* (IJCoL) promoted by the *Associazione Italiana di Linguistica Computazionale* (AILC - `www.ai-lc.it`). This is a miscellaneous volume which follows the First Issue on *Emerging Topics at the Fourth Italian Conference on Computational Linguistics* (n. 4, vol. 1, 2018). It collects a selection of particularly promising research contributions inspired by young researchers at the CLiC-it 2017 Conference, held in Rome from 11 to 13 December 2017. This second selection of contributions from CLiC-it 2017 focuses on a large set of applications, such as Social Media analysis, stylometry and musicology, and research areas of modern Computational Linguistics, such as lexical semantics and lexicography.

As for the other miscellaneous issues, the papers have been selected through an iterative peer-review process. Each article underwent three evaluations: as a contribution to the conference; as a candidate for the "Best Young Paper" and "Distinguished Young Paper" awards of CLiC-it 2017; finally, in the extended version, as a journal article.

The volume opens with the paper by Micheli and colleagues that present a large lexical resource for the study of Latin. The *Word Formation Latin* (WFL) is a derivational lexicon for Latin based on a principled system of formative components and word formation rules (WFRs). WFL presents itself as an answer to the current increased interest in both the theoretical and applied aspects of word formation and represents an important lexical resource. The paper discusses its impact non only on the study of Latin derivational morphology, but mostly on compounding. It presents the methodology and workflow employed to represent compound lemmata in the resource, an aspect which is often neglected in recent resources for other languages.

In the paper by Cignarella et al., the development of a textual resource dedicated to model irony in social media texts is presented. The overall process has been designed within the development of the Twitter corpus TWITTIRÒ whereas a multi-layered scheme for fine-grained annotation of irony is proposed. It follows a multilingual setting, previously applied also to French and English datasets. The outcome of the reported research is not only on the methodological side, but it also includes a novel gold standard corpus for irony detection in Italian. It enriches the current portofolio of available multilingual datasets for the challenging irony detection task, ready to be used as a benchmark in experiments and evaluation campaigns.

In their work, Pannitto and colleagues study the automatic detection of hypernym relations using distributional semantics methods. This paradigm has inspired several unsupervised methods for the detection of hypernym relationship between pairs of lexical entries according to linear algebra operations applied to lexical vector representations. In the paper a new approach is proposed, represented by a smoothed version of the distributional inclusion hypothesis, where pragmatically inspired constraints are imposed to achieve a better discrimination between co-hyponym and hypernym word pairs. The extensive experimental evidence reported over the BLESS dataset confirms that the method is able to outperform previous standard application of this method.

The paper by Della Moretta and colleagues focuses on a rich lexical resource and investigates its use in the semantic tagging of Italian texts. A number of linguistic issues are discussed about the exploitation of the semantic types provided by the T-PAS (Typed Predicate-Argument Structure) resource in tagging Italian texts. The largely recognized hypothesis that the fillers of a certain argument position are characterized by a set of common semantic features/constraints is studied by investigating how human annotators tend to agree or disagree on the annotation of these typed structures. Dis-

agreement cases are here used to identify phenomena that represent aspects to improve the annotation model and process.

In their paper, Basile and colleagues discuss the potential of machine learning approaches in recognizing controversial topics in Web texts such as Social Media posts. A battery of distant supervised models based on Facebook reactions are proposed as proxies for predicting news controversy. A Facebook-based corpus is presented, which has been used to test the validity of the proposed approach to model controversies and as a test bed for benchmarking different controversy recognition models. Results show that controversy and reactions can be modelled successfully at various degrees of granularity.

The volume closes with the paper by Unguendoli and colleagues that discusses an automatic attribution method applied to the field of musicology. The results of the implementation of an automatic stylometric attribution technique to a corpus of liturgical monodies of medieval origin is presented. The strictly linear nature of the musical repertoires allows the adoption of pseudo-distance metrics between frequency-vectors of n-gram of consecutive symbols. Results suggest that the proposed quantitative approach can be effectively used to support the investigation of refined and more complex problems in musicology.

This synthetic view does not exhaust the suggestions and nuances inspired by the papers here collected. We leave the reader the pleasure to discover them through a thoughtful sailing across the rest of the volume contents.