IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 3, Number 1
june 2017

Emerging Topics at the Third Italian Conference on Computational Linguistics and EVALITA 2016

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 978-88-99982-64-5

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_3_1

aAccademia university press

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

IJCoL

Volume 3, Number 1
june 2017

Emerging Topics at the Third Italian Conference on Computational Linguistics and EVALITA 2016

CONTENTS

Nota Editoriale

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC–CNR, Pisa

Eccoci al quarto numero dell'*Italian Journal of Computational Linguistics* (IJCoL), la *Rivista Italiana di Linguistica Computazionale* edita dall' "Associazione Italiana di Linguistica Computazionale" (AILC - www.ai-lc.it). La rivista, al suo terzo anno di pubblicazione, si sta affermando come importante occasione per la promozione e la diffusione della ricerca in linguistica computazionale condotta all'interno della comunità nazionale da prospettive diverse e complementari, che integrano in un rapporto dialettico i punti di vista umanistico e matematico–formale, teorico e applicato.

Il ruolo sempre più centrale che il linguaggio svolge nei processi di comunicazione odierna è quotidianamente certificato da riflessioni e annunci che anche i media tradizionali hanno fatto propri. La natura essenzialmente linguistica della comunicazione nelle reti sociali on–line, così come il vantaggio competitivo che oggi l'industria associa alla capacità computazionale di trattare il linguaggio in volumi crescenti di dati confermano l'importanza strategica della ricerca in questa area che costituisce in prospettiva un fecondo terreno di innovazione sociale, industriale ed economica. I temi che ruotano attorno a linguaggio e computazione forniscono opportunità uniche per comprendere da un lato i modi con cui le macchine possono elaborare le produzioni linguistiche (scritte e orali) e definire processi avanzati di tipo applicativo, e dall'altro contribuire a una sempre migliore comprensione dei modi con cui il linguaggio opera e cambia nel tempo, nello spazio e attraverso diversi canali e mezzi di comunicazione. I contributi di questo volume ben si collegano a queste due dimensioni di ricerca, in linea con lo spirito della rivista che intende proporsi come forum in cui le diverse anime della linguistica computazionale dialogano e si confrontano.

Seguendo la tradizione dei primi due numeri, questo è un volume miscellaneo che raccoglie lavori di ricerca ispirati da giovani ricercatori che sono emersi nell'ambito della Conferenza CLiC-it 2016, tenutasi a Napoli il 5 e 6 dicembre 2016, come particolarmente promettenti nel panorama della linguistica computazionale italiana. Questi contributi, selezionati tra le diverse aree tematiche della conferenza, testimoniano linee di ricerca originali e innovative della comunità italiana, e in modo particolare dei suoi più giovani protagonisti. Gli articoli sono stati 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 2016; infine, nella versione estesa, come articolo di rivista scientifica. A questi articoli si aggiunge un contributo invitato che propone una rassegna e una riflessione critica sull'esperienza di EVALITA 2016, la campagna di valutazione delle tecnologie del linguaggio per la lingua italiana

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

scritta e parlata che, come da tradizione, contribuisce a fare il punto sull'efficacia e la qualità dei metodi di analisi della lingua italiana all'interno di un ricco e vario repertorio di task applicativi. Anche nell'edizione del 2016, EVALITA ha costituito un raccordo essenziale tra la fase empirica e applicativa della ricerca e la riflessione metodologica e teorica.

I contributi del volume si articolano in due macro–sezioni: la prima raccoglie contributi di ricerca originali e innovativi, la seconda fornisce un resoconto della campagna di valutazione EVALITA 2016 delineandone al contempo linee di sviluppo per il futuro.

All'interno della prima macro–sezione, i primi quattro contributi sono accomunati dall'utilizzo di modelli di semantica distribuzionale: è interessante notare che nel 2016 *Distributional Semantics* è risultata essere la parola chiave usata più frequentemente dagli autori nella caratterizzazione dei propri contributi a CLiC-it. Non meraviglia quindi che 4 dei 5 contributi selezionati siano riconducibili a questo paradigma di ricerca. Essi differiscono a vari livelli, che vanno dalle finalità della ricerca e le lingue trattate, ai modelli delle distribuzioni statistiche delle parole nei corpora, alle proprietà semantiche considerate e alla caratterizzazione dei contesti linguistici usati nella determinazione degli spazi semantici. In questa sede ci limitiamo a segnalare i diversi scenari – sia teorici sia applicativi – all'interno dei quali tali tecniche sono state utilizzate.

Il lavoro di Rodda, Senaldi e Lenci propone un metodo per lo studio del cambiamento semantico a partire dalla variazione degli spazi semantici distribuzionali, dimostrando il potenziale contributo della semantica distribuzionale in una prospettiva diacronica. Il metodo è stato sperimentato all'interno di uno studio finalizzato all'identificazione delle aree coinvolte da cambiamenti semantici nel lessico del greco antico tra l'epoca precristiana e cristiana.

L'articolo di Ponti, Jezec e Magnini utilizza tecniche di semantica distribuzionale per esplorare questioni aperte in ambito linguistico sull'interfaccia tra sintassi e semantica, riguardanti le proprietà lessico–semantiche degli argomenti verbali. Lo studio, condotto sulla classe dei verbi ad alternanza causativa–incoativa, ha portato alla luce importanti differenze nelle proprietà degli argomenti che rappresentano categorie non uniformi, la cui distribuzione attorno a un prototipo varia in misura significativa tra diverse posizioni argomentali, suggerendo l'esistenza di diverse restrizioni di selezione.

Nel contributo di Senaldi, Lebani e Lenci, la similarità distribuzionale tra il vettore di una data espressione e il vettore delle sue varianti lessicali è alla base di una serie di indici finalizzati a discriminare tra espressioni idiomatiche e composizionali: viene dimostrato che gli indici proposti presentano una maggiore efficacia rispetto a quanto proposto nella letteratura distribuzionale sulla composizionalità.

Diversamente dai precedenti contributi che mostrano come modelli di semantica distribuzionale contribuiscano significativamente a fare luce su questioni aperte della linguistica teorica, Vanzo, Croce, Basili e Nardi utilizzano la semantica distribuzionale all'interno di uno scenario applicativo di Human-Robot Interaction. In particolare, dimostrano come essa svolga un ruolo rilevante nell'apprendimento strutturato di un processo di *Semantic Role Labeling* per interfacce robotiche.

Chiude la macro–sezione il contributo di Bompolas, Ferro, Marzi, Cardillo e Pirrelli che mostra ruolo e impatto derivante dall'utilizzo di approcci basati sulla nozione di paradigma sull'analisi e sull'apprendimento delle parole. Con esperimenti condotti su tre lingue (greco moderno, italiano e tedesco) gli autori dimostrano che diverse classi verbali sono apprese in funzione del loro grado di trasparenza e predicibilità. Tali risultati, in linea con evidenza psicolinguistica, contribuiscono significativamente a rafforzare l'ipotesi del lessico mentale come sistema emergente integrato.

Il volume si chiude con il contributo invitato degli organizzatori della quinta edizione della campagna di valutazione EVALITA. Nel 2016 sono stati organizzati 6 *shared tasks*, tra cui alcuni nuovi, e una competizione sponsorizzata dall'IBM, che hanno attratto globalmente 34 partecipanti. Una delle novità di questa edizione è rappresentata dal focus sulla lingua dei social media, che ha portato alla creazione di una risorsa con annotazioni multi–livello (PoS tags, sentiment information, named entities and linking, e factuality information) che è stata condivisa tra diversi tasks, permettendone anche una maggiore compenetrazione. Sul versante dei metodi e delle tecniche alla base dei sistemi che hanno partecipato a EVALITA 2016, è interessante segnalare che *Deep Learning* è risultata la parola chiave usata più frequentemente dagli autori nella caratterizzazione dei propri contributi, mostrando anche la complementarità, per il 2016, di EVALITA rispetto a CLiC-it. A partire dall'analisi dell'edizione 2016, gli organizzatori di EVALITA concludono delineando interessanti prospettive di sviluppo della campagna di valutazione per gli anni a venire.

Al lettore dunque il piacere di approfondire i temi e gli stimoli che la rivista – anche in questo numero – continua a raccogliere.

Editorial Note Summary

We are pleased to announce the fourth issue of the *Italian Journal of Computational Linguistics* (IJCoL), published by the Italian Association of Computational Linguistics (AILC, `www.ai-lc.it`). The journal, in its third year of publication, is becoming an important opportunity for promoting and disseminating research results achieved by the Italian computational linguistics community from different and complementary – e.g. humanistic vs. computational or theoretical vs. applied – perspectives.

The increasingly central role that language plays in today's communication processes is daily acknowledged. The essentially linguistic nature of communication in social networks, as well as the competitive advantage that industry today associates with the computational capacity to handle language in big data confirm – together – the strategic importance of the research in this area, creating a fertile ground for social, industrial and economic innovation. The topics revolving around language and computation provide unique opportunities, on the one hand, to understand the ways in which machines can process language productions (both written and oral) and to define advanced applications, and on the other hand to contribute to a deeper understanding of the ways in which language works and changes over time, space and through different channels and means of communication. Contributions to this volume are well linked to these two dimensions of research, in line with the spirit of the journal that presents itself as a forum where the different souls of computational linguistics confront each other and are combined.

As in the case of the first two IJCoL issues, this is a miscellaneous volume that collects research work inspired by young researchers who emerged as particularly promising within the CLiC-it 2016 Conference, held in Naples on 5–6 December 2016. These contributions testify original and innovative research lines of the Italian computational linguistics community, and in particular of its youngest protagonists. The articles were evaluated through an iterative peer–review process carried out by different committees: as a contribution to the CLiC-it conference; as a candidate for the CLiC-it 2016 "Best Young Paper" and "Distinguished Young Paper" awards; finally, in its extended version, as a scientific journal article. This issue also includes an invited contribution by the organizers of the EVALITA 2016 evaluation campaign.

The contributions in this issue are organized into two macro–sections, with the first one collecting original and innovative research contributions, and the second one providing an overview of EVALITA 2016 and outlining further developments.

The first four papers illustrate different applications of distributional semantic models: it is interesting to note that in 2016 *Distributional Semantics* was the keyword most frequently used by CLiC-it authors for characterizing their contribution. They differ at different levels, ranging from the research goal and the language(s) dealt with to the computational techniques used to model word co-occurrence statistics, the semantic properties taken into account and the contexts used in determining semantic spaces.

The work by Rodda, Senaldi and Lenci proposes an innovative method for studying semantic change starting from the variation of distributional semantic spaces, thus demonstrating the potential contribution of distributional semantics to diachronic studies. The method was tested in a case study aimed at identifying the areas of semantic change in the ancient Greek lexicon between the pre-Christian and Christian era.

The paper by Ponti, Jezec and Magnini uses distributional semantics to investigate open linguistic issues on the syntax–semantics interface, with particular emphasis on the lexical-semantic properties of verbal arguments. The study, carried out on the class of causative–inchoative verbs, has brought to light important differences in the properties of arguments.

In the contribution of Senaldi, Lebani and Lenci, distributional similarity between the vector of a given expression and the vector of its lexical variants is used as the basis of a set of indices aimed at discriminating between idiomatic vs. compositional expressions, which turned out to be more effective than those proposed so far in the distributional literature on compositionality.

Unlike the previous contributions showing how distributional semantics models can significantly contribute to shed light on open theoretical linguistics issues, Vanzo, Croce, Basili and Nardi use distributional semantic models within an application scenario: Human–Robot Interaction. In particular, they show the beneficial impact of distributional semantic lexicons on the structured learning of a *Semantic Role Labeling* component in robotic interfaces.

This section is closed by the paper by Bompolas, Ferro, Marzi, Cardillo and Pirrelli, who demonstrate role and impact of paradigm–based approaches for word processing and learning. The results of a case study simulating the acquisition of Modern Greek conjugation, compared with evidence from German and Italian, support a view of the mental lexicon as an emergent integrative system.

The volume closes with the invited contribution of the organizers of the fifth edition of the EVALITA evaluation campaign. In 2016, 6 *shared tasks* were organized together with a challenge sponsored by IBM, which globally attracted 34 participants. One of the novelties of this edition is the focus of social media language, which led to the creation of a resource with multi-level annotations (PoS tags, sentiment information, named entities and linking, and factuality information) that has been used across different tasks. For what concerns methods and techniques underlying EVALITA 2016 participant systems, it is worth reporting that *Deep Learning* was the keyword most frequently used by the authors in characterizing their system; this also demonstrates the complementarity of EVALITA with respect to CLiC–it in 2016. EVALITA's organizers conclude their paper by outlining interesting lines of development for future EVALITA campaigns.

The synthetic view provided above does not exhaust the wide range of topics touched by the papers in this issue; this leaves the reader the pleasure to discover the themes and stimuli that the journal continues to collect – even in this issue.