IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 1
june 2016

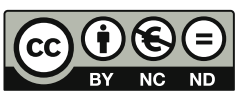
Emerging Topics at the Second Italian Conference on Computational Linguistics

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)

direttore responsabile
Michele Arnese



isbn 978-88-99200-99-2

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_02

aAccademia university press

IJCoL

Volume 2, Number 1
june 2016

Emerging Topics at the Second Italian Conference on Computational Linguistics

CONTENTS

Nota Editoriale

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC–CNR, Pisa

Siamo lieti di introdurre il secondo 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`). Con questo numero, la rivista continua, al servizio della comunità italiana, a contribuire alla promozione e alla diffusione dei risultati della ricerca nel campo della linguistica computazionale, affrontata da prospettive e culture diverse e complementari. Analogamente alla prima uscita, questo volume miscellaneo è dedicato a linee di ricerca particolarmente promettenti nel panorama della linguistica computazionale italiana e internazionale. In particolare, include il contributo invitato di Paola Merlo (Université de Genève) e la versione estesa di una selezione tra i lavori migliori che hanno contribuito alla conferenza annuale CLiC-it ("Italian Conference on Computational Linguistics") nella sua seconda edizione del 2015, tenutasi a Trento il 3 e 4 Dicembre, e che vedono come protagonisti giovani ricercatori.

I temi che ruotano attorno a linguaggio e computazione costituiscono un terreno fecondo per riflessioni intellettuali, scientifiche e tecnologiche su un duplice versante: da un lato, riguardano le capacità del computer di comprendere la struttura e il contenuto di produzioni linguistiche (scritte e orali) e di interagire col mondo esterno in linguaggio naturale, e dall'altro contribuiscono a una migliore comprensione del modo in cui il linguaggio funziona, viene appreso e cambia nel tempo, nello spazio, attraverso diverse situazioni comunicative oppure a seconda del mezzo o canale adottato per la comunicazione. Si tratta di due direzioni di ricerca complementari e in osmosi continua: se una più profonda comprensione del linguaggio ha consentito lo sviluppo di applicativi caratterizzati da una migliore gestione dell'informazione, metodi quantitativi e computazionali nati e sviluppati in modo autonomo si sono spesso rivelati utili per aprire nuovi orizzonti di ricerca nell'indagine linguistica. I contributi di questo volume sono rappresentativi di entrambe le direzioni di ricerca: ciò è in linea con lo spirito con cui è nata la rivista, ovvero di proporsi come forum in cui le diverse anime della linguistica computazionale possano dialogare e confrontarsi.

Il volume si apre con il contributo invitato di Paola Merlo dal titolo *Quantitative computational syntax: some initial results* che mostra in modo efficace e innovativo come tecniche e modelli computazionali basati su evidenza estratta da corpora arricchiti con annotazione linguistica multi–livello disponibili per diverse lingue consentano oggi di dare nuova linfa allo studio di numerose proprietà grammaticali astratte, oggetto tradizionale di indagine della linguistica teorica. Partendo dai risultati di ricerche dell'autrice su temi che spaziano dall'ordine delle parole alla semantica verbale,

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

l'articolo conferma ipotesi significative circa il ruolo della *frequenza* nelle spiegazioni che i modelli teorici hanno dato della complessità e della variabilità lessico–grammaticale di una lingua. L'evidenza che la linguistica computazionale sembra poter fornire alle riflessioni più teoriche è il segno concreto di un ruolo che pur attraversando discipline e conoscenze diverse è radicato nella linguistica come suo più autentico fondamento.

I due contributi che seguono possono essere collocati all'interno della stessa macro–area tematica focalizzata sulla modellazione, attraverso metodi di *machine learning*, di fenomeni linguistici tipici di varietà d'uso della lingua diverse, costituite rispettivamente da produzioni linguistiche in lingua straniera (o L2) o dialogiche. Nel contributo di Stehwien e Pado, il tema dell'identificazione della lingua materna (*Native Language Identification, o NLI*) viene discusso attraverso uno studio comparativo condotto su un corpus di letteratura scientifica in inglese da parte di parlanti nativi di lingue tipologicamente diverse. In particolare, vengono indagati la generalizzabilità di modelli di *Native Language Identification* basati su *learner corpora* e l'adattamento della tecnica a un nuovo genere testuale, costituito da articoli scientifici. Nell'articolo di Dragone e Lison viene discusso il riconoscimento di *Non Sentential Utterances* (NSUs) nel dialogo, cioè espressioni linguistiche che veicolano informazioni predicative cruciali che non sono però articolate in frasi grammaticalmente complete e ben formate. La combinazione di tecniche di *Active Learning* e di un nuovo inventario di caratteristiche linguistiche ha permesso di compensare l'esigua disponibilità di dati linguisticamente annotati e di sperimentare modelli probabilistici per la classificazione di NSUs rispetto alla tassonomia definita.

Il lavoro di Brunato e Dell'Orletta illustra la progettazione e la costruzione di un corpus di produzioni linguistiche scritte e orali, arricchito con annotazioni di varia natura, per l'analisi dello sviluppo linguistico di ragazzi in età scolare (scuola primaria). La risorsa presentata costituisce un supporto fondamentale per le ricerche sull'acquisizione linguistica dopo i cinque anni e rende possibile il confronto dell'evoluzione delle competenze a livello di lingua scritta e parlata. Analisi preliminari condotte dagli autori con strumenti di annotazione automatica e monitoraggio linguistico mostrano le potenzialità del corpus nell'indagine dell'evoluzione linguistica nella scuola primaria.

L'ultima sezione del volume include due lavori particolarmente significativi della conferenza legati ad aspetti applicativi molto attuali, quali l'analisi di microblog e il riconscimento di fenomeni semantici (in particolare, le citazioni di entità e relazioni) in essi presenti. Bonadiman e colleghi descrivono una *Deep Neural Network* (DNN) per la modellazione del compito di *Named Entity Recognition* (NER) su testi in lingua italiana. Il modello neurale presentato è computazionalmente vantaggioso, anche per la politica di apprendimento in *backpropagation* che si basa esplicitamente sulle associazioni tra le *label* di output e le dipendenze grammaticali. Chiude il volume il lavoro di Pierpaolo Basile e colleghi, che discute il problema dell'adattamento di un algoritmo per il *Named Entity Linking* al contesto specifico dell'analisi di *tweet* in italiano; il contributo include anche un'illustrazione del corpus di riferimento sviluppato.

Lasciamo a questo punto al lettore l'onere, e speriamo il piacere, di approfondire direttamente i diversi temi degli articoli di questo secondo volume, qui brevemente delineati: esso conferma l'ampiezza, la qualità e la diversità delle prospettive di ricerca presentate a CLiC-it 2015, segno tangibile del potenziale della comunità italiana ma anche internazionale a cui questa rivista intende contribuire pienamente nel tempo.

Editorial Note Summary

We are pleased to announce the second issue of the *Italian Journal of Computational Linguistics* (IJCoL), published by the "Italian Association of Computational Linguistics" (AILC, `www.ai-lc.it`). With this issue, the journal continues to contribute to the promotion and dissemination of research results in the field of computational linguistics, tackled from different and complementary perspectives. As for the first issue, this is a miscellaneous volume reporting the results of promising lines of computational linguistics research in Italy and abroad. In particular, it includes an invited contribution by Paola Merlo (University of Geneva) and the extended version of a selection among the best papers involving junior researchers and presented at CLiC-it 2015, the second edition of the "Italian Conference on Computational Linguistics" which was held in Trento in December 2015.

The issues revolving around natural language and computation relate, on the one hand, to the computer's ability to understand the structure and content of linguistic productions (both written and oral) and to interact with the outside world in natural language, and, on the other hand, to a better understanding of the way in which the language works, is learned and changes in time, in space, through different communicative situations or channels. These are two directions of research which are complementary and in constant osmosis: if a deeper understanding of language has enabled the development of better knowledge management applications, quantitative and computational analysis methods which were developed independently have often proved useful to open new horizons in linguistics research. The contributions of this volume are representative of both directions of research: this is in line with the goal of the journal of acting as a forum in which the various "souls" of computational linguistics can meet and discuss.

This issue of the journal opens with the contribution by Paola Merlo, that follows from her invited talk at CLiC-it 2015. Interestingly, the work outlines the feedback that computational models of complex linguistic phenomena allow to empirically reinforce or shed new light onto significant challenges targeted traditionally by theoretical linguistics. This is illustrated in three case studies focusing on word order and verb meaning properties across languages. This work is fully in line with the spirit of the journal that aims at establishing explicit connections across areas (computational vs. theoretical linguistics) keeping a central focus on natural languages and their evolution.

The two papers that follow belong to the same macro–thematic area focusing on computational modeling through machine learning methods of linguistic phenomena characterizing different varieties of language use, represented respectively by linguistic productions of non–native speakers and dialogs. The paper by Stehwien and Pado focuses on the task of Native Language Identification (NLI), studied comparatively across different types of text collections, such as learner vs scientific corpora. The emergence of information is found to strictly depend on corpus specific aspects such as language–specific transfer models and topic indicators. In Dragone and Lison, the recognition of non sentential utterances in dialogue is studied. These convey crucial clausal meaning without exhibiting a complete and well formed sentential form. Against the inherent problem of annotated data scarcity, the paper proposes the combination of a wide inventory of linguistic features and probabilistic active learning techniques that allowed to carry on successfull experiments against a gold standard based on a taxonomy of non sentential utterance classes.

Brunato and Dell'Orletta present a new corpus of oral and written productions by Italian children in primary school. The resource aims at supporting research on

"later language acquisition", with a specific view to the comparative assessment of the evolution of oral and written language competencies in early school grades.

The last group of papers report research on machine learning applied in NLP applications. Bonadiman and colleagues describe a Deep Neural Network (DNN) for the modeling of Named Entity Recognition (NER) over raw texts in Italian. The proposed neural modeling relies on a new recurrent feedback mechanism to ensure that the dependencies between the output tags are properly modeled, making it simple and computationally efficient. The evaluation is carried out against the Evalita 2009 benchmark. Finally, the paper by Basile and colleagues propose an adaptation of an existing Named Entity Linking algorithm to the context of Italian tweets. For evaluating the proposed algorithm, a new dataset made of Italian tweets is also presented.

The above synthetic view does not exhaust the wide range of issues touched by the papers and this leaves the reader the pleasure to discover them through a thoughtfull sailing across the rest of the volume contents. We think this volume sheds further light on achievements regularly emerging from the world–wide dimensions of the computational linguistics research, with particular emphasis on the contributions by the Italian community.