IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 4, Number 1
june 2018

Emerging Topics at the Fourth 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)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 978-88-31978-40-8

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_4_1

IJCoL

Volume 4, Number 1
june 2018

CONTENTS

Emerging Topics at the Fourth Italian Conference on Computational Linguistics

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC - CNR

1. Introduction

E' con gran piacere che introduciamo il primo volume del quarto anno della rivista *Italian Journal of Computational Linguistics* (*IJCoL*), la rivista italiana di linguistica computazionale promossa dall'*Associazione Italiana di Linguistica Computazionale* (AILC - www.ai-lc.it). La rivista, fino a oggi, è uscita regolarmente con cadenza semestrale e ha raccolto importanti contributi della comunità nazionale e internazionale della linguistica computazionale, con particolare attenzione a ricerca di frontiera condotta da parte di giovani ricercatori. I numeri pubblicati finora coprono un ampio spettro di temi che ruotano attorno alla dicotomia linguaggio-computazione, affrontata da prospettive diverse riconducibili alle "anime" umanistica e informatica della linguistica computazionale, con diverse finalità, sia teoriche sia applicative, e con particolare attenzione al trattamento automatico della lingua italiana nelle sue diverse varietà d'uso. Dalla fondazione, sono stati pubblicati due numeri speciali della rivista, dedicati all'approfondimento di aree di ricerca strategiche della disciplina, sul versante umanistico e informatico, riguardanti rispettivamente l'apporto di metodi e tecniche della linguistica computazionale alle "Digital Humanities" e i paradigmi dominanti nel panorama degli algoritmi di apprendimento automatico che stanno influenzando pesantemente gli sviluppi correnti della disciplina.

Dalla fase iniziale di rodaggio la rivista sta passando oggi a una fase più matura: ne è testimonianza il recente riconoscimento da parte dell'*Academic Committee* di OpenEdition, l'infrastruttura europea dedicata alla comunicazione e alla pubblicazione in Open Access della ricerca accademica in un ampio spettro di settori scientifici, che ha deliberato la pubblicazione della rivista sulla piattaforma *OpenEdition Journals*. Siamo consapevoli che per una nuova rivista la strada per guadagnare prestigio e autorevolezza è lunga e tutt'altro che scontata. Crediamo tuttavia che i risultati conseguiti finora siano promettenti e stiano creando i presupposti per la classificazione di IJCoL tra le riviste scientifiche del settore e, in prospettiva, tra le riviste in Classe A dell'ANVUR, e per la sua indicizzazione nei principali database internazionali rilevanti per i settori coperti dalla rivista (tra i quali, Scopus Bibliographic Database, ERIH Plus, Google Scholar, Web of Science). Tutto ciò, grazie alla passione e all'impegno di chi, a diverso titolo, sta contribuendo a questa importante impresa.

Alla Special Issue su *Natural Language and Learning Machines* (n. 3, vol. 2, 2017), segue oggi questo numero miscellaneo che, seguendo la tradizione di diversi precedenti

* 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

numeri, raccoglie lavori di ricerca ispirati da giovani ricercatori, che sono emersi come particolarmente promettenti nell'ambito della Conferenza CLiC-it 2017, tenutasi a Roma dall'11 al 13 dicembre 2017. Questo insieme corrisponde a una prima selezione di contributi di CLiC-it 2017, caratterizzata dalla ricerca su algoritmi di apprendimento profondo ("deep learning") per la soluzione di diversi e complessi compiti di inferenza linguistica. Si presenta quindi come una sorta di continuazione del precedente numero speciale della rivista, preludendo così a una seconda selezione dei lavori da CLiC-it 2017, la cui pubblicazione è prevista per il prossimo numero del 2018.

Come per altri numeri miscellanei della rivista, gli articoli di questo numero 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 2017; infine, nella versione estesa, come articolo di rivista scientifica. A questi articoli si aggiunge il contributo invitato dedicato alla rassegna della Conferenza CLiC-it 2017, a cura dei tre *co-chair*, con particolare attenzione alle novità introdotte per un coinvolgimento sempre maggiore della comunità italiana della linguistica computazionale: dei giovani all'interno di percorsi di formazione così come dei potenziali "stakeholders" - che vanno dalla Pubblica Amministrazione alle piccole e medie imprese - come beneficiari dell'apporto dei risultati della ricerca nazionale e internazionale nel settore della linguistica computazionale.

Apre il volume il paper di Lakew e colleghi, che discute un modello neurale per la traduzione automatica in grado di affrontare la sfida posta da lingue caratterizzate da una scarsa disponibilità di risorse di *training*. L'approccio proposto si basa sulla creazione di uno spazio semantico multilingue che permette il trasferimento dei parametri usati tra lingue diverse, migliorando così le condizioni di addestramento per i casi in cui i dati per una specifica coppia di lingue siano limitati. Nel lavoro, questa ipotesi è verificata mostrando i risultati di esperimenti condotti su tre lingue (inglese, italiano e rumeno): la metodologia proposta migliora le prestazioni rispetto a sistemi bilingui, evitando al contempo la complessità insita nell'addestramento di tali sistemi.

Nel lavoro di Bizzoni et al., rappresentazioni vettoriali sono utilizzate per la classificazione di espressioni idiomatiche e non, in condizioni di limitata disponibilità di dati: l'obiettivo è quello di verificare se l'informazione convogliata dal vettore distribuzionale associato a una data espressione sia sufficiente alla rete per inferire la sua potenziale idiomaticità. La sperimentazione presentata conferma il ruolo cruciale di rappresentazioni distribuzionali in questo compito. Diversamente da quanto rilevato in precedenza per il riconoscimento di espressioni metaforiche, l'impiego di rappresentazioni vettoriali associate all'espressione nel suo complesso si dimostra essere più efficace rispetto alla concatenazione dei vettori associati alle singole parole dell'espressione.

Il lavoro di Masotti e colleghi, si occupa di un tema piuttosto nuovo nel panorama italiano: la generazione automatica di didascalie per immagini, processo che coinvolge in modo integrato competenze di tipo visuale (nel riconoscimento dei tipi di oggetti ritratti nella immagine) e linguistico (nella generazione di frasi corrette che descrivono gli oggetti e la situazione ritratta che li coinvolge). L'architettura neurale presentata integra due reti distinte: una prima rete dedicata all'*embedding* grafico, e una seconda rete ricorrente per la generazione automatica della didascalia che utilizza l'*embedding* prodotto dalla prima come input (stato iniziale). Tale architettura, già applicata con successo alla lingua inglese, è stata addestrata su un dataset esteso per la lingua italiana ottenuto attraverso strumenti di traduzione automatica applicati alle descrizioni in inglese di una collezione di immagini.

Infine, il lavoro di Cardillo e colleghi affronta il problema dell'induzione di conoscenza morfologica seguendo un approccio che, piuttosto che presupporre una segmentazione delle parole in morfemi, si basa sulle connessioni tra forme flesse all'interno di reti lessicali associative. Secondo questo approccio, l'identificazione della cella paradigmatica appropriata per una forma flessa sconosciuta è guidata dall'evidenza offerta da forme conosciute. La novità del contributo consiste nell'utilizzo di reti neurali per modellare il processo di flessione delle parole come inferenza paradigmatica. A tal fine, sono state utilizzate reti di tipo *Long Short Term Memory* (LSTM) che si sono mostrate particolarmente flessibili ed efficaci nel combinare diversi tipi di informazione (relativa alla struttura morfologica, all'organizzazione paradigmatica e al grado di (ir)regolarità nella formazione del tema), e in grado di adattarsi alle specificità e ai diversi livelli di complessità caratterizzanti ciascun sistema morfologico.

Questa breve vista d'insieme non esaurisce i molti aspetti di interesse che emergono dai lavori che compongono il presente volume per quanto concerne l'adozione di tecnologie di apprendimento automatico per il trattamento della lingua. Lasciamo quindi al lettore l'onere e il piacere di approfondirli direttamente negli articoli qui raccolti.

2. Editorial Note Summary

It is with great pleasure that we introduce the first 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`). Until today, the journal has been regularly published biannually and has collected important contributions from the national and international communities of computational linguistics, with particular attention to frontier research carried out by young researchers. The volumes published so far cover a wide spectrum of themes revolving around the language-computation dichotomy, addressed from the humanistic and computational perspectives, with both theoretical and applicative purposes, and with particular emphasis on the automatic processing of Italian in its different varieties of use.

Since its foundation, two special issues have been published, dedicated to strategic areas of the discipline: namely, the contribution of methods and techniques of computational linguistics to "Digital Humanities" and the dominant Machine Learning paradigms that are currently influencing the developments of the discipline. The journal is now entering into a mature phase, as confirmed e.g. by the recent recognition by the *OpenEdition Academic Committee* which has deliberated the publication of IJCoL on the *OpenEdition Journals* platform. We are aware that it takes time to earn prestige for a new journal. However, we believe that the results achieved so far are promising and are creating the prerequisites for the classification of IJCoL among the scientific journals in the computational linguistics area and for its indexing in the main international bibliographic databases. All this has been possible thanks to the passion and commitment of those who, in different ways, are contributing to this important enterprise.

This miscellaneous volume follows the Special Issue on *Natural Language and Learning Machines* (No. 3, Volume 2, 2017); as in the previous issues, it collects a selection of research contributions inspired by young researchers which emerged as particularly promising at the CLiC-it 2017 Conference, held in Rome from 11 to 13 December 2017. This first selection of contributions from CLiC-it 2017 shares the use of deep learning algorithms for the solution of different and challenging linguistic problems. It thus presents itself as a sort of continuation of the previous special issue of the journal, which will be followed by another miscellaneous volume with a second selection of papers from CLiC-it 2017, whose publication is scheduled for the second issue of 2018.

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. This set of papers also includes an invited contribution devoted to a retrospective of CLiC-it 2017 by the conference co-chair, with particular attention to the innovations introduced for an increasing involvement of the Italian community of computational linguistics: in particular, young researchers and potential "stakeholders", ranging from public administrations to small and medium-sized companies.

The volume opens with the paper by Lakew and colleagues, discussing a neural model for Machine Translation (NMT) that addresses the challenge of low-resourced languages. The proposed approach is multilingual: i.e. it is based on the creation of hidden representations of words in a shared semantic space across multiple languages, thus enabling a positive parameter transfer across languages. Results of experiments carried out on three languages (English, Italian and Romanian) are reported: compared to bilingual NMT systems, the system significantly improves its performance, while avoiding the complexity inherent in training systems for single language pairs.

In the paper by Bizzoni et al. vector representations are used for the classification of Italian idiomatic and non-idiomatic phrases under constrains of data scarcity. The goal is to assess whether and to what extent the information conveyed by the distributional vector associated with a phrase (whether idiomatic or not) is sufficient for the network to infer its potential idiomaticity. Reported experiments confirm the crucial role of distributional representations in this task. Contrary to what previously reported for metaphorical expressions, the use of phrase-based vector representations proves to be more effective than the concatenation of the vectors associated with the individual words of the expression.

The work by Masotti and colleagues tackles a recent topic in the Italian landscape: the automatic generation of image captions, a process that involves both visual and linguistic skills. The neural architecture presented for this purpose integrates two distinct networks: a first network dedicated to the vector representation of the image, and a second recurrent network for the automatic generation of the caption that uses the *embedding* produced by the first as input. This architecture, already successfully applied to English, is trained on an extended data set for Italian obtained through automatic translation of English descriptions of a collection of images.

Finally, the work by Cardillo et al. addresses the problem of the induction of morphological knowledge following an approach that, rather than presupposing a segmentation of words into morphemes, is based on the connections between inflected forms within associative lexical networks. According to this approach, the identification of the appropriate paradigmatic cell for an unknown inflected form is guided by the evidence offered by known forms. The novelty of the contribution consists in the use of neural networks to model word inflection as a paradigmatic inference. To this end, *Long Short Term Memory* (LSTM) networks were used which proved to be particularly flexible and effective in combining different types of information and able to adapt to the peculiarities of each morphological system.

This 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 thoughtful sailing across the rest of the volume contents. We think this volume sheds further light on achievements regularly emerging from the worldwide dimensions of the computational linguistics research, with particular emphasis on the contributions by the Italian community.