ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 5, Number 2
december 2019

aA ccademia 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 979-12-80136-06-0

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_5_2

IJCoL

Volume 5, Number 2
december 2019

CONTENTS

Nota editoriale

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC - CNR

Il secondo numero del quinto anno della rivista *Italian Journal of Computational Linguistics* (*IJCoL*), la rivista italiana promossa dall'*Associazione Italiana di Linguistica Computazionale* (AILC - `www.ai-lc.it`), è un volume miscellaneo i cui articoli documentano una selezione di linee di ricerca attive nel panorama della Linguistica Computazionale italiana con risultati interessanti. Tra questi, vi sono articoli che documentano lavori di ricerca risultati particolarmente promettenti nell'ambito della Conferenza CLiC-it 2019 (Bari, 13–15 novembre 2019), così come contributi originali proposti per la pubblicazione sulla rivista. Tutti i contributi sono stati sottoposti a un processo di peer-review, iterativo nel caso degli articoli premiati come "Best Young Paper" e "Distinguished Young Paper" nell'ambito della conferenza. Chiude il volume un contributo dedicato alla memoria di Emanuele Pianta, un ricercatore che ha significativamente contribuito alla crescita della Linguistica Computazionale in Italia scomparso prematuramente nel 2012.

I temi affrontati coprono sviluppi recenti e fecondi della ricerca in linguistica computazionale, come ad esempio l'uso di tecniche di NLP complesse per la analisi dei fenomeni legati ai social media o l'adozione di algoritmi neurali per il trattamento di fenomeni audio (*speech profiles*) o visuali (video e immagini) e l'ottimizzazizone di compiti linguistici complessi, rappresentati dal cosiddetto "*captioning*" o "*speech recognition*".

Il lavoro di Polignano *e colleghi* presenta AlBERTo, un modello di lessico semantico per la lingua italiana addestrato sulla lingua dei Social Media, in particolare Twitter. In linea con i modelli basati sul paradigma dei *trasformers* (BERT *in primis*), AlBERTo è stato addestrato sfruttando la decomposizione del task di apprendimento nell'ambiente Google Cloud Platform e la disponiblità del corpus TWITA che raccoglie circa 200 milioni di tweet generalisti in lingua italiana. Il modello risultante è distribuito *open source* attraverso la piattaforma GitHub. La disponibilità di tale risorsa su larga scala è un risultato importante, in quanto rende possibili numerose ricerche e applicazioni di "*Computational Social Science*" per l'italiano, da parte di una sempre più vasta comunità di ricercatori.

Il lavoro di Ferro e Tamburini valida ed estende il ruolo di modelli di *smoothing*, discusso recentemente, negli algoritmi di *Pitch Detection* impiegati in sistemi di *Speech Recognition*. La ricerca dimostra che gli algoritmi neurali per lo *smoothing* possono migliorare le performances in modo significativo. In particolare, viene introdotto un *pitch smoother* basato su una architettura neurale che usa Keras come interfaccia di riferimento verso TensorFlow. Esso è in grado di incidere in modo eccellente su due *benchmark* standard per la lingua inglese, apprendendo il meccanismo di smoothing di un *pitch detector* in modo da eliminare completamente alcune classi di errori.

* 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`

Il lavoro di Scaiella *e colleghi* presenta l'applicazione di tecniche neurali per l'addestramento di un sistema di generazione di commenti testuali a immagini e testi. La ricerca sfrutta architetture in grado di codificare video (o immagini) in vettori numerici (*embeddings*) per alimentare un secondo sistema neurale (ricorrente) usato per generare il commento in lnguaggio naturale. Il lavoro descrive lo sviluppo semiautomatico di un corpus di video commentati per la lingua italiana, usando come sorgente la controparte in inglese.

Il lavoro di Monti e Di Buono descrive una risorsa originale e innovativa, il corpus PARSEME-It VMWE sviluppato all'interno della PARSEME COST Action che rappresenta il primo e l'unico corpus per la lingua italiana ad oggi arricchito con informazione relativa a una vasta a variegata tipologia di espressioni polirematiche (*MultiWord Expressions*, in breve MWE), che vanno da espressioni idiomatiche e composti a *light verb contructions* e locuzioni di varia natura (avverbiali, preposizionali, etc.). Il corpus PARSEME-It VMWE italiano rappresenta l'esito di un'analisi estensiva e linguisticamente motivata delle MWEs italiane, ed è accompagnato da specifiche dettagliate per la loro identificazione, classificazione e rappresentazione.

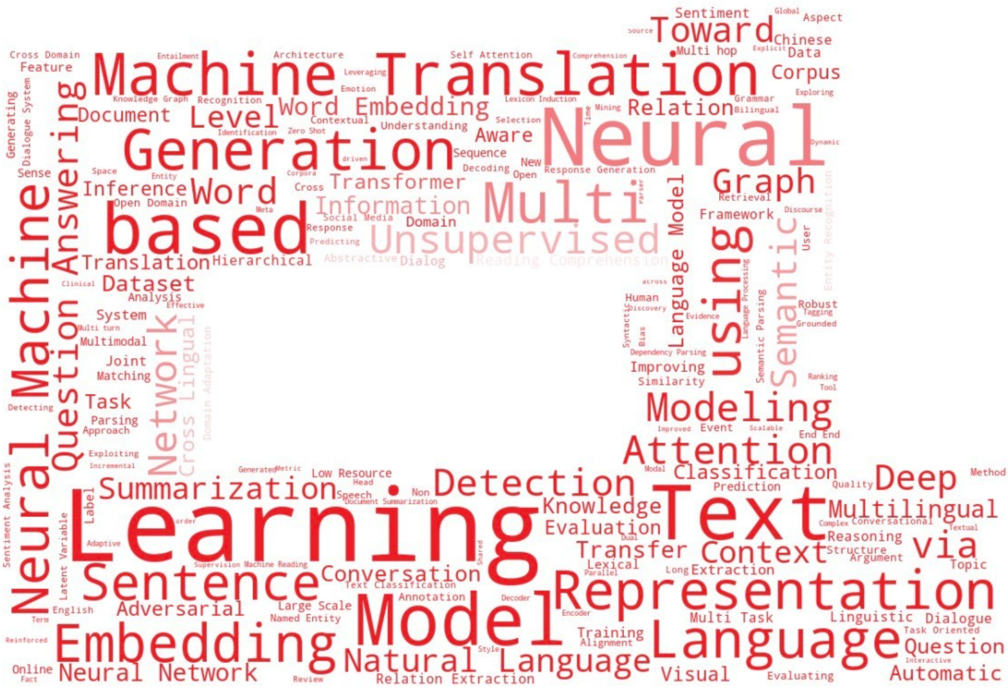
Infine, segue il contributo dedicato a Emanuele Pianta, eccellente studioso e ricercatore del settore della Linguistica Computazionale e in particolare del Trattamento Automatico della Lingua, prematuramente scomparso nel Novembre 2012. Magnini, Delmonte e Tonelli - tra i ricercatori che sono stati più vicini a Emanuele - ripercorrono i suoi contributi alla ricerca, che vengono presentati come un esempio vivido e fecondo per molti ricercatori, giovani e meno giovani. In riconoscimento del suo contributo, il Direttivo di AILC ha deciso di attivare un Premio intitolato alla sua memoria, assegnato annualmente alla miglior tesi di laurea magistrale nell'ambito della Linguistica Computazionale discussa in una università italiana.

Speriamo che questa sintesi del volume - inevitabilmente parziale - ispiri, come sempre, al lettore il desiderio di navigare, secondo i propri interessi, nelle pieghe delle pagine di questo volume, certamente più ricche di dettagli e sfumature.

Prima di chiudere questa nota vogliamo segnalare un importante evento che ha visto il coinvolgimento della comunità italiana della Linguistica Computazionale: il 57° Convegno Annuale dell'*Association for Computational Linguistics* (ACL), la più importante associazione scientifica internazionale per la Linguistica Computazionale, che si è svolto alla Fortezza da Basso a Firenze dal 28 luglio al 2 agosto 2019. Il convegno annuale dell'ACL è il momento in cui scienziati di tutto il mondo si confrontano per fare il punto sullo stato dell'arte della disciplina e sulle prospettive future di sviluppo. L'edizione italiana del 2019 è stata eccezionale per due motivi.

Prima di tutto è la prima volta che il convegno di ACL è stato organizzato in Italia. Organizzatori locali dell'evento sono stati Alessandro Lenci (Università di Pisa), Bernardo Magnini (Fondazione Bruno Kessler, Trento), Simonetta Montemagni (Istituto di Linguistica Computazionale "A. Zampolli" del CNR), che si sono avvalsi della collaborazione di un ampio segmento della comunità italiana in questo settore. Il fatto che ACL abbia scelto l'Italia come paese ospite del suo convengo annuale è stato un grande onore per tutta la nostra comunità e può essere visto come testimonianza della sua rilevanza nel panorama internazionale. L'Italia è in effetti da sempre protagonista delle ricerche in linguistica computazionale, che ha mosso i suoi primi passi proprio a Pisa ormai più di 50 anni fa. Oggi l'Italia conta molti centri di ricerca e ditte che contribuiscono attivamente all'avanzamento dello stato dell'arte nel settore. La ricchezza di attività della comunità nazionale è testimoniata dalla recente nascita dell'*Associazione Italiana di Linguistica Computazionale* (AILC), che ha supportato attivamente l'organizzazione del convegno.

Il secondo fattore di eccezionalità del convegno fiorentino sono state le sue dimensioni. Tutti i numeri di ACL 2019 sono da record: 660 presentazioni accettate alla conferenza principale in 6 sessioni parallele, 21 tra workshop e conferenze satellite per più di 3200 iscritti in totale. L'edizione 2019 è in assoluto quella col maggior numero di partecipanti nella storia di ACL. Tra i temi più presenti nel programma ci sono stati gli agenti conversazionali in grado di dialogare con gli esseri umani, la traduzione automatica, l'uso dei metodi di "deep learning", ma anche aspetti etici legati alle tecnologie del linguaggio. Di grande attualità sono risultate le nuove tecnologie per combattere l'uso improprio e offensivo del linguaggio sui Social Media, o per gestire conversazioni tra medico e paziente, solo per menzionarne alcune. La varietà dei temi affrontati emerge chiaramente dal "word cloud" che segue, che contiene le parole chiave emergenti dall'analisi dei titoli degli articoli presentati nell'ambito della conferenza.



La conferenza è stata sponsorizzata da tutte le "Big Companies" del Web (Amazon, Apple, Baidu, Facebook, Google, ecc.) e da molte altre imprese italiane e straniere, segno del grande interesse commerciale di un settore di mercato in continua espansione.

Qualcuno ha definito la Conferenza *ACL-2019* tenutasi a Firenze "*The best ACL ever!*": ciò può anche essere visto come testimonianza dell'eccellenza del contributo alla ricerca internazionale che la comunità italiana della Linguistica Computazionale fornisce da anni e che AILC intende rafforzare e promuovere sin dai suoi esordi, solo qualche anno fa. La strada percorsa è già lunga, nonostante la giovane età dell'associazione: ne siamo orgogliosi.

1. Editorial Note Summary

The second volume of the fifth year of the *Italian Journal of Computational Linguistics* (IJCoL) promoted by the *Associazione Italiana di Linguistica Computazionale* (AILC - `www.ai-lc.it`) integrates original research results as well as papers that have been presented at CLiC-it 2019 held in Bari, in November 2019. The major themes discussed by the papers are hot topics that include complex learning methods for the analysis of Social Media texts, neural algorithms for integrated audio, visual and language learning. Some works are on NLP resources developed for teh Italian language.

The interetsing paper by Polignano *et al.* presents AlBERTo, a lexical semantic model based on the Transformer paradigm, trained over Social Media material in Italian. AlBERTo exploits the availability of the TWITA corpus, that includes about 200 millions generalist tweets in Italian. The resulting model is distributed under an *open source* scheme on the GitHub platform. The availability of this resource is inspiring a number of further studies on "*Computational Social Science*" over Web sources in Italian involving a growing research community.

Ferro and Tamburini study how neural *smoothing models* can be adopted to improve the *Pitch Detection* stage in *Speech Recognition* systems. The work shows how a *pitch smoother* based on a Keras API towards interfaccia Tenworflow is able to limit error rates of a *pitch detector* on a large English speech corpus.

The work by Scaiella and colleagues presents the application of convolutional and recurrent neural networks to the task of automatic captioning of images an video. The architecture develops on methods already experimented for English, and shows how on Italian similar performances can be achieved. The work also describes the semi-automatic development and the release of an annotated corpus of video captions for the Italian language, through the automatic translation of the English counterpart.

The work by Monti and Di Buono presents the PARSEME-It corpus for the analysis of multiword expressions (MWE). It is a linguistically principled annotated corpus, that embody a comprehensive study for the annotation of MWE in italian: it specializes the PARSEME COST Action framework.

Finally, one contribute to this volume is dedicated to Emanuele Pianta, brilliant researcher in Computational Linguistics and Natural Language Processing, whose untimely death, in November 2012, has left a sad gap in the Italian CL community. Bernardo Magnini, Rodolfo Delmonte and Sara Tonelli, who were closer to Emanuale during his studies, go through his own major research contributions in the paper. It surveys thus a lively and fruitful example for all of us, younger or senior researchers. Accordingly, in an attempt to emphasize his contributions, the Steering Committee of AILC decided to dedicate to Emanuele Pianta a Prize, yearly assigned to the best Master Degree thesis in the Computational Linguistics area defended during the year in one Italian University.

This very synthetic view serves only to survey the focus of the papers. We leave the reader the pleasure to navigate across the valuable pages of our volume and discover there all the interesting details.