ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 6, Number 1
june 2020

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 PDF 9791280136404

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_6_1

IJCoL

Volume 6, Number 1
june 2020

CONTENTS

Nota editoriale

Roberto Basili*
Università di Roma, Tor Vergata

Simonetta Montemagni**
ILC - CNR

Il primo numero del sesto 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 linee di ricerca attive nel panorama della Linguistica Computazionale italiana con risultati interessanti. Gli articoli raccolti in questo volume documentano una prima selezione di lavori di ricerca risultati particolarmente promettenti nell'ambito della sesta Conferenza CLiC-it, tenutasi a Bari, dal 13 al 15 novembre 2019. Tutti i contributi sono stati sottoposti a un processo di peer-review iterativo, prima nell'ambito della conferenza, poi come candidati ai premi di "Best Young Paper" e "Distinguished Young Paper", infine come contributo su rivista.

I temi affrontati coprono sviluppi recenti e fecondi della ricerca in linguistica computazionale, che vanno dalla costruzione di modelli e algoritmi cognitivamente motivati per l'analisi del significato, all'utilizzo di "word embeddings" per studi diacronici del lessico della lingua latina e a metodi di profilazione linguistica applicati a generi testuali e registri linguistici, per giungere all'applicazione di metodi e tecniche sviluppate in ambito linguistico-computazionale all'interno di diversi scenari, che spaziano dalla sottotitolazione di film al riconoscimento di espressioni di odio.

Apre il volume il contributo di Preissner ed Herbelot, insignito del "Best Student Paper Award" nell'ambito di CLiC-it 2019. In reazione alle costose architetture tipiche dei sistemi di learning per NLP che sembrano legate a enormi risorse di calcolo sia in addestramento che in esecuzione, gli autori esplorano la sfida di riprodurre la biodiversità del mondo reale all'interno di modelli neurali di trattamento automatico della lingua. In particolare, un algoritmo di ispirazione cognitiva viene qui presentato come un interessante punto di partenza per l'esplorazione di nuove architetture. L'articolo valida ed estende il lavoro originale degli autori sull'algoritmo denominato "Fruit Fly Algorithm", ispirato al processo naturale di adattamento negli organismi viventi, cioè la diminuzione nell'intensità della risposta a uno stimolo frequente nell'ambiente. Le potenzialità dell'algoritmo modificato sono discussse in un compito di apprendimento di modelli vettoriali del lessico, con particolare attenzione al suo comportamento strettamente incrementale e al suo livello di interpretabilità rispetto ad altri metodi.

Il contributo di Sprugnoli, Moretti e Passarotti presenta una nuova serie di "Lemma Embeddings" per la lingua latina costruiti a partire da testi appartenenti all'era classica lemmatizzati manualmente. A tal fine sono stati testati e valutati diversi modelli, architetture e dimensioni, utilizzando il nuovo benchmark per l'attività di selezione dei sinonimi. Attraverso il confronto degli embeddings relativi all'era classica con quelli pre-addestrati sull'"Opera Maiora" di Tommaso d'Aquino, è stato condotto un

* 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

esperimento volto a un'analisi diacronica dell'uso lessicale. L'articolo mostra il duplice impatto dei risultati raggiunti: se da un lato gli embeddings prodotti possono essere utilizzati per il trattamento automatico di testi latini, sia a livello di strumenti (e/o modelli) sia di nuove risorse linguistiche, dall'altro possono essere utilmente sfruttati in analisi lessicali diacroniche da parte di studiosi di scienze umane che lavorano nell'area delle lingue classiche.

L'articolo di Buongiovanni et al. propone un'analisi contrastiva volta a identificare similarità e differenze tra diversi generi testuali e/o registri linguistici, che si basa sui prerequisiti della metodologia di "profiling" linguistico con lo scopo specifico di modellare la forma stilistica delle diverse parti in cui è articolato un testo. L'indagine comparativa, svolta su quattro generi testuali e due livelli di complessità per ciascuno, ha mostrato che la distribuzione delle caratteristiche morfosintattiche e sintattiche si differenzia tra sottosezioni di testi rappresentativi di un dato genere e varietà linguistica. L'approccio, testato sulla lingua italiana utilizzando un ampio spettro di caratteristiche linguistiche estratte automaticamente da corpora analizzati, dimostra che è possibile modellare il grado di varianza stilistica all'interno dei testi secondo il genere e la complessità del linguaggio utilizzato.

La seconda parte del volume, dedicata ad applicazioni specifiche basate su metodi e tecniche di trattamento automatico della lingua, si apre con il contributo di Karakanta e colleghi, che affronta il problema della sottotitolazione di film ricorrendo alla Neural Machine Translation (NMT), una pratica consolidata che permette di ridurre costi e tempi di consegna. Contrariamente alla traduzione del testo, i sottotitoli sono soggetti a vincoli spaziali e temporali, che aumentano notevolmente lo sforzo di post-elaborazione richiesto. Partendo da un lavoro precedente, gli autori hanno identificato diversi elementi mancanti nei corpora disponibili per l'addestramento di sistemi di NMT, specificamente adottati per i sottotitoli. Attraverso un'analisi comparativa dei risultati ottenuti con diverse risorse di sottotitolazione parallele esistenti, gli autori giungono alla conclusione che una combinazione intelligente di vecchie e nuove risorse di sottotitolazione potrebbe essere vantaggiosa per la creazione di soluzioni NMT per la sottotitolazione, in particolare per supportare più domini, come film, serie, interviste, conferenze, documentari ed eventi con relatori singoli o multipli.

Chiude il volume l'articolo di Capozzi et al. che descrive la piattaforma Web realizzata nell'ambito del progetto "Contro l'Odio", per il monitoraggio e il contrasto alla discriminazione e all'incitamento all'odio nei confronti degli immigrati in Italia. Il lavoro applica una combinazione di tecniche di trattamento automatico della lingua per il riconoscimento dei discorsi di incitamento all'odio e di strumenti di visualizzazione dei dati tratti da Twitter, consentendo agli utenti di accedere a un'enorme quantità di informazioni attraverso mappe interattive, di visualizzare i tweet più virali e di ridurre in modo interattivo la complessità intrinseca dei dati. La piattaforma Web sviluppata è stata oggetto di corsi di formazione per studenti delle scuole superiori finalizzati alla decostruzione degli stereotipi negativi contro gli immigrati, i rom e le minoranze religiose, e alla creazione di narrazioni positive. I dati raccolti e analizzati dalla piattaforma sono attualmente utilizzati anche per attività di benchmarking all'interno di una campagna di valutazione e per aprire la strada a nuovi progetti contro l'odio.

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

1. Editorial Note Summary

The first volume of the sixth year of the *Italian Journal of Computational Linguistics* (IJCoL) promoted by the *Associazione Italiana di Linguistica Computazionale* (AILC - `www.ai-lc.it`) reports original research results that have been presented at CLiC-it 2019 held in Bari, in November 2019. The articles collected in this miscellaneous volume document a first selection of papers that turned out to be particularly promising in the context of the conference. 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 2019 "Best Young Paper" and "Distinguished Young Paper" awards; finally, in its extended version, as a scientific journal article.

The investigated topics cover recent and fruitful developments in computational linguistics research, ranging from the construction of cognitively motivated models and algorithms for the analysis of meaning, to the use of word embeddings for diachronic lexical studies of the Latin language, and methods of linguistic profiling applied to investigate similarities and differences across textual genres and linguistic registers; the volume closes with two papers focusing on the application of NLP methods and techniques within different scenarios, ranging from film subtitling to the recognition of hate expressions in social media.

The contribution by Preissner and Herbelot, awarded the "Best Student Paper Award" within CLiC-it 2019 opens the volume. In reaction to current expensive NLP architectures that require enormous computing resources, the authors address the challenge of reproducing real-world biodiversity within artificial models of automatic language processing. In particular, cognitively inspired algorithms are presented as an interesting starting point for the exploration of novel architectures. The article validates and extends the original work of the authors on the so-called "Fruit Fly Algorithm", modeling neural learning as the natural process of adaptation in living organisms, i.e. decrease in response to a frequent stimulus. The potential of the modified algorithm is illustrated with respect to a word vector learning task, with particular attention to its fully incremental behavior and its level of interpretability compared to other methods.

The contribution by Sprugnoli, Moretti and Passarotti presents a new series of "Lemma Embeddings" for the Latin language constructed starting from manually lemmatized texts belonging to the classical era. To this end, different models, architectures and dimensions were tested and evaluated, using the new benchmark for the activity of selecting synonyms. By comparing the embeddings relating to the classical era with those pre-trained on the "Opera Maiora" by Thomas Aquinas, an experiment was conducted aimed at a diachronic analysis of lexical use. The article shows the double impact of the results achieved: if on the one hand the embeddings produced can be used for the automatic processing of Latin texts, both at the level of tools (and / or models) and of new linguistic resources, on the other hand they can be usefully exploited for lexical analysis in a diachronic perspective by scholars working in the area of classical languages.

The article by Buongiovanni et al. proposes a contrastive analysis aimed at identifying similarities and differences between different textual genres and / or linguistic registers, which is based on the prerequisites of the linguistic "profiling" methodology, with the specific aim of modeling the stylistic form of different parts in which a text is articulated. The comparative investigation, carried out on four textual genres and two levels of complexity for each, showed that the distribution of morphosyntactic and syntactic characteristics differs between subsections of representative texts of a given genre and linguistic variety. The approach, tested on the Italian language using a

broad spectrum of linguistic characteristics automatically extracted from the analyzed corpora, demonstrates that it is possible to model the degree of stylistic variance within the texts according to the genre and complexity of the language used.

The second part of the volume, dedicated to specific applications based on automatic language processing methods and techniques, opens with the contribution of Karakanta and colleagues, who tackles the problem of film subtitling using Neural Machine Translation (NMT) techniques. This consolidated practice allows to reduce costs and delivery times. Unlike text translation, subtitles are subject to spatial and temporal constraints, which greatly increase the required post-processing effort. Building on previous work, the authors identified several missing elements in the corpora available for training NMT systems, when specifically adopted for subtitles. Through a comparative analysis of the results obtained with several existing parallel subtitling resources, the authors come to the conclusion that a clever combination of old and new subtitling resources could be beneficial for creating NMT solutions for subtitling, particularly to support more domains, such as films, series, interviews, conferences, documentaries and events with single or multiple speakers.

The paper by Capozzi and colleagues describes the Web platform built within the project "Contro l'Odio", for monitoring and contrasting discrimination and hate speech against immigrants, Rom and religious minorities in Italy. By combining NLP techniques for hate speech detection and data visualization tools on data drawn from Twitter, the "Contro l'Odio" Web platform allows users to access a huge amount of information through interactive maps, also tuning their view, e.g. visualizing the most viral tweets and interactively reducing the inherent complexity of data. Since October 2018 the platform analyzes daily Twitter posts and exploits temporal and geo-spatial information related to messages in order to ease the summarization of the hate detection outcome. The data collected and analyzed by the platform have also been used for benchmarking activities within a national evaluation campaign, and for paving the way to new projects against hate. The platform has also been used by the civil society organization partners for educational purposes in courses for high school students.

After this synthetic view of the papers in this issue, we leave the reader the pleasure to navigate across the valuable pages of the volume.