Third Italian Conference on Computational Linguistics CLiC-it 2016

Poster session

Monday, December 5

Linguistic Issues in CL and NLP

Poster n. 1: Erica Tusa, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi *Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità*

In this work, we will investigate whether and to what extent algorithms typically used to assess the reliability of the output of syntactic parsers can be used to study the correlation between processing complexity and the linguistic notion of markedness. Although still preliminary, achieved results show the key role of features such as dependency direction and length in defining the markedness degrees of a given syntactic construction.

Machine Translation and Multilingual Applications

Poster n. 2: Johanna Monti, Federico Sangati, Francesca Chiusaroli, Martin Benjamin, Sina Mansour *Emojitalianobot and EmojiWorldBot - New online tools and digital environments for translation into emoji*

Emojitalianobot and EmojiWorldBot are two new online tools and digital environments for translation into emoji on Telegram, the popular instant messaging platform. Emojitalianobot is the first open and free Emoji-Italian and Emoji-English translation bot based on Unicode descriptions. The bot was designed to support the translation of Pinocchio into emoji carried out by the followers of the "Scritture brevi" blog on Twitter and contains a glossary with all the uses of emojis in the translation of the famous Italian novel. EmojiWorldBot, an off-spring project of Emojitalianobot, is a multilingual dictionary that uses Emoji as a pivot language from dozens of different languages. Currently the emoji-word and word-emoji functions are available for 72 languages imported from the Unicode tables and provide users with an easy search capability to map words in each of these languages to emojis, and vice versa. This paper presents the projects, the background and the main characteristics of these applications.

Information Retrieval and Question Answering

Poster n. 3: Gianni Barlacchi, Azad Abad, Emanuele Rossinelli, Alessandro Moschitti *Appetitoso: A Search Engine for Restaurant Retrieval based on Dishes*

Recent years have seen an impressive development and diffusion of web applications to food domains, e.g., Yelp, TripAdvisors. These mainly exploit text for searching and retrieving food facilities, e.g., *restaurants*, *caffé*, *pizzerias*. The main features of such applications are: the location and quality of the facilities, where quality is extrapolated by the users' reviews. More recent options also enable search based on restaurant categorization, e.g., *Japanese*, *Italian*, *Mexican*. In this work, we introduce Appetitoso, an innovative approach for finding restaurants based on the dishes a user

would like to taste rather than using the name of food facilities or their general categories.

Poster n. 4: Giorgio Maria Di Nunzio, Maria Maistro, Daniel Zilio *Gamification for IR: The Query Aspects Game*

The creation of a labelled dataset for IR purposes is a costly process. For this reason, a mix of crowdsourcing and active learning approaches have been proposed in the literature in order to assess the relevance of documents of a collection given a particular query at an affordable cost. In this paper, we present the design of the gamification of this interactive process that draws inspiration from recent works in the area of gamification for IR. In particular, we focus on three main points: i) we want to create a set of relevance judgements with the least effort by human assessors, ii) we use interactive search interfaces that use game mechanics, iii) we use NLP to collect different aspects of a query.

Linguistic Resources

Poster n. 5: Linda Alfieri, Fabio Tamburini *(Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format*

This paper describes the automatic procedure we developed to convert an Italian dependency treebank into a different format. We defined about 4,250 formal rules for rewriting dependencies and token tags as well as an algorithm for treebank rewriting able to avoid rule interference. At the end of this process a large portion of the whole treebank was automatically converted, with very few errors, leaving only a small amount of work to be done manually.

Poster n. 6: Alice Bracchi, Tommaso Caselli, Irina Prodanof *Enrichring the Ita-TimeBank with Narrative Containers*

This paper reports on an annotation experiment to enrich an existing temporally annotated corpus of Italian news articles with Narrative Containers, annotation devices representing temporal windows in text and marking up very informative temporal relations between temporal entities. The annotation has shown that the distribution of Narrative Containers is sensitive to the text genre and may be used to facilitate the creation of informative timelines.

Poster n. 7: Valeria Caruso, Anna De Meo, Vincenzo Norman Vitale *Increasing information accessibility on the Web: a rating system for specialized dictionaries*

The paper illustrates the features of the WLR (Web Linguistic Resources) portal, which collects specialized online dictionaries and asses their suitability for different functions using a specifically designed rating system. The contribution aims to demonstrate how the existing tool has improved

the usefulness of lexicographical portals and how its effectiveness can be further increased by transforming the portal into a collaborative resource.

Poster n. 8: Elisa Corino, Claudio Russo *Parsing di corpora di apprendenti di italiano: un primo studio su VALICO*

Modern learner corpora are now routinely PoS tagged, whereas syntactic parsing is much less frequent. This paper proposes a first attempt of parsing applied to a subcorpus of VALICO, in an effort to identify key elements to be further used to parse corpora of Italian as a foreign language in a proper way.

Poster n. 9: Anna Fantini *Spammare senza pietà - Corpus based analysis of English, unacclimatised verb loans in Italian and creation of a reference lexicon*

We describe the lexical resource created to investigate the semantic changes of 90 English, un-acclimatised verb loans in Italian. Final results and interesting observations concerning the annotation task are discussed.

Poster n. 10: Antonio Lieto, Enrico Mensa, Daniele P. Radicioni *Taming Sense Sparsity: a Common-Sense Approach*

We present a novel algorithm and a linguistic resource named CLOSeST after 'Common SEnse STrainer'. The resource contains a list of the main senses associated to a given term, and it was obtained by applying a simple set of pruning heuristics to the senses provided in the NASARI vectors for the set of $15K$ most frequent English terms. The preliminary experimentation provided encouraging results.

Poster n. 11: Alessandro Mazzei *Building a computational lexicon by using SQL*

This paper presents some issues about a computational lexicon employed in a generation system for Italian. The paper has three goals: (i) to describe the SQL resources produced during the construction of the lexicon; (ii) to describe the algorithm for building the lexicon; (iii) to present an ongoing work for enhancing the lexicon by using the syntactic information extracted from a treebank.

Poster n. 12: Lucia C. Passaro, Alessandro Bondielli, Alessandro Lenci *FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis*

In this paper we present the FB-NEWS15 corpus, a new Italian resource for sentiment analysis and emotion detection. The corpus has been built by crawling the Facebook pages of the most important newspapers in Italy and it has been organized into topics using LDA. In this work we provide a preliminary analysis of the corpus, including the most debated news in 2015.

Poster n. 13: Irene Russo, Simone Pisano, Claudia Soria *Sardinian on Facebook: Analysing Diatopic Varieties through Translated Lexical Lists*

Presence of regional and minority languages over digital media is an indicator of their vitality. In this paper, we want to investigate quantitative aspects of the use on Facebook of the Sardinian language. In particular, we want to focus on the co-existence of diatopic varieties. We extracted linguistic data from public pages and, through the translation of the most frequent words, we find out similarities and differences between varieties.

Poster n. 14: Anna Corazza, Valerio Maggio, Giuseppe Scanniello *A new dataset for source code comment coherence*

Source code comments provide useful insights on a codebase and on the intent behind design decisions and goals. Often, the information provided in the comment of a method and in its corresponding implementation may be not coherent with each other (i.e., the comment does not properly describe the implementation). Several could be the motivations for this issue (e.g., comment and source code do not evolve coherently). In this paper, we present the results of a manual assessment on the coherence between comments and implementations of 3,636 methods, gathered from 4 Java open-source software. The results of this assessment has been collected in a dataset that we made publicly available on the web. We also sketch here the protocol to create this dataset.

NLP for Digital Humanities

Poster n. 15: Martina A. Rodda, Marco S.G. Senaldi, Alessandro Lenci *Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek*

We present a method to explore semantic change as a function of variation in distributional semantic spaces. In this paper we apply this approach to automatically identify the areas of semantic change in the lexicon of Ancient Greek between the pre-Christian and Christian era. Distributional Semantic Models are used to identify meaningful clusters and patterns of semantic shift within a set of target words, defined through a purely data-driven approach. The results emphasize the role played by the diffusion of Christianity and by technical languages in determining semantic change in Ancient Greek and show the potentialities of distributional models in diachronic semantics.