Third Italian Conference on Computational Linguistics CLiC-it 2016

Poster session

Tuesday, December 6

NLP for Digital Humanities

Poster n. 1: David Alfter, Yuri Bizzoni *Hybrid Language Segmentation for Historical Documents*

Language segmentation, i.e. the division of a multilingual text into monolingual fragments has been addressed in the past, but its application to historical documents has been largely unexplored. We propose a method for language segmentation for multilingual historical documents. For documents that contain a mix of high- and low-resource languages, we leverage the high availability of high-resource language material and use unsupervised methods for the low-resource parts. We show that our method outperforms previous efforts in this field.

Poster n. 2: Franca Orletti, Felice Dell'Orletta, Rossella Iovino *La leggibilità dei testi di ambito medico rivolti al paziente: il caso dei bugiardini di farmaci senza obbligo di prescrizione medica*

In this paper we present the first results of an exploratory analysis of simplification of the package leaflets of medicines, considered representative texts of doctor-patient communication. It will be shown how natural language processing tools can be used to reconstruct the linguistic profile of these texts and to guide their simplification.

Poster n. 3: Silvia Piccini, Andrea Bellandi, Giulia Benotto, Emiliano Giovannetti *La Modellazione Diacronica di Risorse Termino-Ontologiche nell'Ambito delle Digital Humanities: Esperimenti su Clavius*

In this work, we present an experiment in the modeling of a diachronic termino-ontological resource named CLAVIUS through both the N-ary relations model and the 4D-fluents approach. Some of the salient differences of these two models are discussed. The overall objective of this research is to illustrate the main advantages and disadvantages in the adoption of a given model to build diachronic resources.

Cognitive modeling of language processing and psycholinguistics

Poster n. 4: Francesca Guglielmi, Pierpaolo Basile, Antonietta Curci, Giovanni Semeraro *Sentiment Analysis: applicazione in un dominio psico-forense*

This study aims to apply the sentiment analysis to a psycho-forensic context through the analysis of narrative transcriptions related to the crimes committed by violent detainees. We investigate the presence of psychopathy through the Psychopathic Personality Inventory-Revised (PPI-R). Psychopathy is a personality disorder that is characterized by emotional detachment, the lack of empathy and it is often found in the prison population for the brutality of committed crimes. Our study explores possible associations between psychopathy and emotional content present in the narratives. Results show a neutral polarity for both psychopathic and not psychopaths offenders, however it is possible to identify significant emotional characteristics that diversify the narrative transcriptions.

Poster n. 5: Roberto Bottini, Daniel Casasanto, Andrea Nadalini, Davide Crepaldi *Stepping out of the Chinese Room: Word meaning with and without consciousness*

What is the role of consciousness in language processing? Unconscious priming experiments show that words can prime other words with related meanings (cat – dog), and these priming effects are assumed to reflect the activation of conceptual knowledge in semantic memory. Alternatively, however, unconscious priming effects could reflect predictive relationships between the words' forms, since words that are semantically related are also statistically related in language use. Therefore, unconscious “semantic” priming effects could be due to relationships between words' forms mimicking conceptual relationships, as in Searle's Chinese Room thought experiment. To distinguish wordform-based and semantics-based accounts of priming we conducted an experiment in which temporal words (e.g., earlier, later) were preceded by spatial words that were processed either consciously or unconsciously. Time is typically conceptualized as a spatial continuum extending along either the sagittal (front-back) or the lateral (left-right) axis, but only the sagittal space-time mapping is encoded in language (e.g. the future is ahead, not to the right). Results showed that temporal words were primed both by sagittal words (back, front) and lateral words (left, right) when primes were perceived consciously, as predicted by both wordform-based and semantics-based accounts. Yet, only sagittal words produced an unconscious priming effect, as predicted by the wordform-based account. Unconscious word processing appears to be limited to relationships between words' forms, and consciousness may be needed to activate words' meanings.

Morphology and Syntax Processing

Poster n. 6: Alberto Lavelli *Comparing State-of-the-art Dependency Parsers on the Italian Stanford Dependency Treebank*

In the last decade, many accurate dependency parsers have been made publicly available. It can be difficult for non-experts to select a good off-the-shelf parser among those available. This is even more true when working on languages different from English, because parsers have been tested mainly on English treebanks. Our analysis is focused on Italian and relies on the Italian Stanford Dependency Treebank (ISDT). This work is a contribution to help non-experts understand how difficult it is to apply a specific dependency parser to a new language/treebank and choose a parser that meets their needs.

Poster n. 7: Marco Passarotti, Marco Budassi *May the Goddess of Hope Help Us. Homonymy in Latin Lexicon and Onomasticon*

We present a study on the degree of homonymy between the lexicon of a morphological analyser for Latin and an Onomasticon. To understand the impact of homonymy, we discuss an experiment on four Latin texts of different era and genre.

NLP for Web and Social Media

Poster n. 8: Anita Alicante, Anna Corazza, Antonio Pironti *Twitter Sentiment Polarity Classification using Barrier Features*

A crucial point for the applicability of sentiment analysis over Twitter is represented by the degree of manual intervention necessary to adapt the approach to the considered domain. In this work we propose a new sentiment polarity classifier exploiting *barrier features*, originally introduced for the classification of textual data.

Poster n. 9: Jennifer-Carmen Frey, Aivars Glaznieks, Egon W. Stemle *The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts*

The DiDi corpus of South Tyrolean data of computer-mediated communication (CMC) is a multilingual sociolinguistic language corpus. It consists of around 600,000 tokens collected from 136 profiles of Facebook users residing in South Tyrol, Italy. In conformity with the multilingual situation of the territory, the main languages of the corpus are German and Italian (followed by English). The data has been manually anonymised and provides manually corrected part-of-speech tags for the Italian language texts and manually normalised data for German texts. Moreover, it is annotated with user-provided socio-demographic data (among others L1, gender, age, education, and internet communication habits) from a questionnaire, and linguistic annotations regarding CMC phenomena, languages and varieties. The anonymised corpus is freely available for research purposes.

Poster n. 10: Felicia Logozzo *Sequenze N+pN (nome comune + nome proprio): descrizione linguistica da un corpus dell'italiano*

This paper describes the most important N+pN (noun + proper noun) structures in Italian from the corpus of La Repubblica 2002-2005

Information Extraction, Entity Linking and (Linked) Open Data

Poster n. 11: Anita Alicante, Anna Corazza, Francesco Isgró, Stefano Silvestri *Relation mining from clinical records*

We propose a system to extract entities and relations from a set of clinical records in Italian based on two preceding works. This approach does not require annotated data and is based on existing domain lexical resources and unsupervised machine learning techniques.

Machine Learning for CL and NLP

Poster n. 12: Fabrizio Esposito, Anna Corazza, Francesco Cutugno *Topic Modelling with Word Embeddings*

This work aims at evaluating and comparing two different frameworks for the unsupervised topic modelling of the CompWHoB Corpus, namely our political-linguistic dataset. The first approach is represented by the application of the latent DirichLet Allocation (henceforth LDA), defining the evaluation of this model as baseline of comparison. The second framework employs Word2Vec technique to learn the word vector representations to be later used to topic-model our data. Compared to the previously defined LDA baseline, results show that the use of Word2Vec word embeddings significantly improves topic modelling performance but only when an accurate and task-oriented linguistic pre-processing step is carried out.