HrTAL 2016 Conference Chronicle
10th International Conference on Natural Language Processing

Organized by the Croatian Language Technologies Society and the Institute of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb

Dubrovnik Centre for Advanced Academic Studies of the University of Zagreb hosted the international linguistic conference **HrTAL2016**. The conference was held from 29th September to 1st October 2016. The chair of the conference was Marko Tadić, and the co–chair was Božo Bekavac, both from the Department of Linguistics of the Faculty of Humanities and Social Sciences, University of Zagreb as well as from the Croatian Language Technologies Society. The conference gathered about 30 scholars from Bulgaria, Croatia, France, Germany, Hungary, Italia, Poland, Romania, Serbia, Slovakia, Slovenia, Tunisia and so on, all wanting to discuss recent findings in linguistics. The official language was English. The weather was perfect, allowing the participants to enjoy Dubrovnik as well as Ston, where the conference dinner was held.

HrTAL2016 is the tenth in the series of TAL conferences, following PolTAL2014 (Warsaw, Poland), JapTAL2012 (Kanazawa, Japan), IceTAL 2010 (Reykjavik, Iceland), GoTAL 2008 (Gothenburg, Sweden), FinTAL 2006 (Turku, Finland), EsTAL 2004 (Alicante, Spain), PorTAL 2002 (Faro, Portugal), VexTAL 1999 (Venice, Italy) and FracTAL 1997 (Besançon, France).

The keynote speaker was **Marko Grobelnik** from the Jožef Štefan Institute in Ljubljana, Slovenia. The title of his talk was *Language as a Social Sensor for Knowledge and Reasoning*. He explained the philosophical approach to the interpretation of the role of language in general. He referred to the philosopher and linguist Ludwig Wittgenstein by quoting that the language itself is a social sensor as well as the tool for knowledge and reasoning. By means of the language the world is understood. At the same time, the scope of the language is to describe the nature, to express the uncertainty, to be efficient, to reflect changes, to transform the physical into the abstract world, which we call knowledge. Furthermore, written language nowadays could be analysed on lexical level, as well as linguistic and semantic level. Multilingual corpora published on the Internet as well as Wikipedia could explain and analyse the information, but manual effort is also required. We also need to distinguish shallow from deep knowledge. He concluded at the end that it was easier to analyse a huge corpus than to analyse one single document.

Matúš Pikuliak and **Marián Šimko** discussed linguistic relations in their talk, *Towards Relationship Extraction from Text Using Word Embeddings*. The talk focused on discussing the extraction of semantic relationships between lexical units from text data, using a deep learning technique of language modeling called word embeddings. The vector space was created using this technique, and the ability to preserve semantic information was assessed. A similarity relationship was extracted. The results of the experiment confirmed the feasibility of the approach.

Marijan Palmović, **Anita Peti–Stantić** and **Jana Willer–Gold** presented their talk, *No Case for Dropped Pronoun – An Eye Tracking Study*. Their paper addressed the issue of dropped pronouns (clitics) in Croatian as a model language. The structural and agreement conditions were offered for spelled–out or dropping of pronoun in addition to predicting their intended interpretation. The intuitions were explored and verified by eye–tracking studies combining auditory and visual stimulus presentation. In the investigation procedure, two sentences had been offered to the participants (30 students at the Department of Linguistics) requiring from them to answer the question *Who caught the fish ?* on the basis of the following sentences (1) Otac je išao pecati sa sinom. (2) On je ulovio ribu. The majority of the answers were *The* son *caught the fish.* The data obtained were integrated in the proposal to combine eye–tracking with corpus annotation methodology. The goal of the talk was to advance research in theoretical, experimental and computational linguistics as well as other research domains. The results of the investigation are expected to be very interesting and useful.

Mirjana Mladenović, **Cvetana Krstev**, **Jelena Mitrović** and **Ranka Stanković** from the University of Belgrade contributed with their talk, *Using WordNet Knowledge for Irony Classification.* This talk was very interesting and attractive for conference participants because it compared Serbian, Bosnian, Croatian and Montenegrin languages. The authors collected the tweets that had been manually annotated according to ironic and non–ironic features. They had used Serbian WordNet ontology (R), antonymous pairs in which one member has positive sentiment polarity (PPR), polarity of positive sentiment words (PSP), ordered sequence of sentiment tags (OSA), Part–of–Speech tags of words (POS) and irony markers (M). The best achieved accuracy of the developed classifier (acc = 86.1%) was achieved with the set of 5 features – (PPR, PSP, POS, OSA, M).

Rafał Jaworski, **Ivan Dunđer** and **Sanja Seljan** presented their talk, *Usability Analysis of the Concordia Tool Applying Novel Concordance Searching,* in which a new tool for concordance searching *Concordia* was described. It uses three data structures, i.e. hashed index, markers array and suffix array, which are loaded into memory to enable fast searches according to the fragments that cover a search pattern. The application of the new tool was analysed in detail according to the experiment of two English–Croatian human translation tasks. At the end the users' attitudes towards the usefulness and functionalities of *Concordia* were presented.

Balázs Indig and **Noémi Vadász** presented the talk *Window in Human Parsing – How Far Can a Preverb Go?* The target of their research was to establish a practical length of the necessary window in Hungarian for verb–preverb distances as well as to present the implementation of coupling the detached preverbs and verbs. They showed the initiative to create a psycholinguistically motivated parser by measuring the most important phenomena in sentence processing and taking into account the observations on human reading.

Karima Abidi and **Kamel Smaïli** distributed multilingual corpora into comparable and parallel corpora in their study *Measuring the Comparability of Multilingual Corpora Extracted from Twitter and Others.* They described the former as a set of texts in several languages dealing with analogous subjects, but these are not translations. The latter are translations of each other (parallel corpora). The new techniques had been applied, evaluated and compared on three different English–Arabic corpora: a corpus extracted from the social network Twitter, from Euronews and a parallel corpus extracted from newspapers.

Jaroslav Loebl and **Marian Šimko** in their talk *Syntactic Analysis in Real–World Corpora* claim that syntactic analysis is an important step in deeper understanding of the text. Natural language processing can influence machine translation, semantic analysis or information extraction. Great advances have been achieved in the last two decades on syntactic parsing. On the other hand, Slovak language hasn't been studied, but with the advent of the Slovak Dependency Treebank (SDT) the situation has changed considerably.

Iana Atanassova, **Marc Bertin**, **Sylviane Cardey** and **Peter Greenfield** presented their talk, *Interfacing the Domain of Global Security with Natural Language Processing: the Role of Language Modeling*. It was an interesting talk that focused on controlled languages and risk management in aviation transport. They discussed crisis and global security and the required language due to the prevailing circumstances for common safety of passengers in aviation and marine industry.

Daša Farkaš, **Matea Filko**, **Vanja Štefanec**, **Danijela Merkler** and **Marko Tadić** presented their talk, *Towards Automatic Semantic Role Labeling in Croatian*. They continued the research on the Croatian Dependency Treebank (HOBS, hobs.ffzg.hr). At the same time they presented the Croatian tagset for SRL and the set of manually annotated sentences which will be applicable both in NLP (natural language processing) and CALL (computer aided language learning). The sentences were annotated both at syntactic and semantic level. A new e–course for learning Croatian (www.hr4eu.eu) was presented.

Angelina Gašpar and **Sanja Seljan** presented the talk, *Terminological Consistency Measured by Herfindahl–Hirshman Index (HHI)*. The two scholars analysed the consistency of the translated terminology conducted on three types of legal domain subcorpora, i.e. Croatian–English parallel corpus (1991–2009), English and Croatian versions of the Code of Canon Law translated from Latin (1983) and English and Croatian versions of the EU legislation (2013). Herfindahl–Hirshman Index (HHI) was used as an indicator for evaluation. A contrastive analysis was made on the extracted terminology and verified in online terminology resources (IATE ad EuroVoc).

Natalia Grabar and **Iris Eshkol** (over Skype) explained the phenomenon of reformulation in the talk *Why do we Reformulate? Automatic Prediction of Pragmatic Functions*. The focus of the talk was the French language and its three spontaneous reformulations introduced by three markers (*c'est–á–dire*, *je veux dire*, *disons*). The data were provided from spoken corpora and a corpus containing forum discussions on health issues. The supervised categorization algorithms were exploited as well as several features such as syntactic, formal, semantic and discursive with the objective to predict reformulation categories. The results of the research showed that it was easier to predict the types of functions at the general level than at the level of individual categories.

Antonio Sammartino and **Marko Tadić** described their efforts to collect the treasure of a valuable variety of the Croatian language, which is in use in Molise region in Italy. The title of the presentation *Building the Corpus* *of Molise Croatian* provided information regarding the initial steps to collect digital texts – the corpus of the dialect. They started with the historical background of the migrations of Croats into South Italy five centuries ago. The speakers showed great enthusiasm in their work, wanting to preserve interesting linguistic features of the dialect. The dictionary and the grammar already exist, but the data of the investigation of the future corpus might result in several papers in frequency linguistics, statistics, contrastive analysis and contact linguistics. The dialect named Slavomolisano is under special UNESCO protection as an endangered language.

Maciej Ogrodniczuk and **Magdalena Zawisławska** presented the talk *Referential Relations in Polish*: *a Classification Attempt.* The paper attempted to present a common taxonomy starting from nominal co–reference through various types of anaphora to bridging relations in Polish. Their approach introduced the concept of facets signaling relation incompleteness (such as dissimilation of referent, uncertainty or subjectivity of speaker) and a means of providing textual evidence proving or disproving co–referential character of a relation.

Rahma Boujelbane, **Ines Zribi**, **Syrine Kharroubi** and **Mariem Ellouze** (over Skype) discussed the problem of standardization of spelling systems in their presentation on *An Automatic Process for Tunisian Arabic Orthography Normalization*. Since there is no standard dialectal spelling system of Arabic dialects, they displayed varieties of orthographic forms on Tunisian Arabic (TA) and compared it with conventional orthography CODA–TA. The statistical component was added as well.

Ivelina Stoyanova, **Svetla Koeva** and **Svetlozara Leseva** contributed with their talk, *Clause Splitting for Bulgarian.* The talk comprised the discussion on clause splitting method based on a set of patterns for recognition of predicates and clause linking elements and an algorithm for clause boundaries detection. They eliminated syntactic parsing. Bulgarian was studied and the results showed F1 score of 0.942 for the recognition of clause opening and clause closing elements and 0.893 for entire clauses.

Verginica Barbu Mittelu, **Radu Ion**, **Radu Simionescu**, **Elena Irmia**, **Cenel–Augusto Perez** presented their talk, *The Romanian Treebank Annotated According to Universal Dependencies.* The talk presented the Romanian treebank according to the Universal Dependency Project annotation principles and set of syntactic relations. Inconsistencies as well as ambiguities were discussed. At the same time, the applicability of the Universal Dependency principles and set of relations was represented on the Romance family language.

HrTAL proceedings are expected to appear in the Springer LNCS/LNAI series as in the case of previous TAL conferences. 31 program committee members from Canada, Croatia, Czech Republic, Denmark, Estonia, Finland, Germany, Hungary, Ireland, Poland, Romania, Slovakia and Sweden contributed to the conference papers. Local organizing committee involved 7 members from the University of Zagreb: Božo Bekavac, Daša Farkaš, Matea Filko, Diana Hriberski, Jurica Polančec, Krešimir Šojat and Marko Tadić.

Congratulations!

Lia Dragojević