



Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

Proceedings of the Third International Conference

**Computational
Linguistics in
Bulgaria**

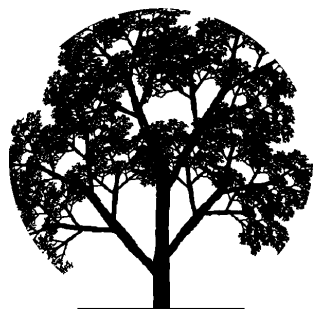


**27-29 May 2018
Sofia, Bulgaria**

The Third International Conference *Computational Linguistics in Bulgaria (CLIB 2018)* is organised with the support of the National Science Fund of the Republic of Bulgaria under the project *Towards a Semantic Network Enriched with a Variety of Relations*, Grant Agreement DN10/3/2016.




CLIB 2018 is organised by:



Department of
Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Third International Conference <i>Computational Linguistics in Bulgaria (CLIB 2018)</i>
ISSN:	2367 5675 (online)
Published and distributed by:	The Institute for Bulgarian Language Bulgarian Academy of Sciences
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria +3592/ 872 23 02
Copyright:	Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).  License details: http://creativecommons.org/licenses/by/4.0 Copyright © 2018

Proceedings of the
Third International Conference
*Computational Linguistics in
Bulgaria*



Sofia, Bulgaria
28-29 May 2018

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language (BAS)

Co-chairs:

Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology

Cvetana Krstev – University of Belgrade

Tania Avgustinova – Saarland University

Iana Atanassova – Centre Tesnière, Université de Franche-Comté, UFR SLHS

Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy

Mariana Damova – Mozaika, Bulgaria

Ivan Derzhanski – Institute of Mathematics and Informatics (BAS)

Radovan Garabík – L’udovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Stefan Gerdjikov – Sofia University

Filip Ginter – University of Turku

Kjetil Rå Hauge – University of Oslo

Zornitsa Kozareva – Google

Ivan Koychev – Sofia University

Hristo Krushkov – Plovdiv University

Eric Laporte – Université Paris-Est Marne-la-Vallée

Denis Maurel – University of Tours

Stoyan Mihov – Institute of Information and Communication Technologies

Ruslan Mitkov – University of Wolverhampton

Preslav Nakov – Qatar Computing Research Institute

Karel Oliva – Institute of the Czech Language, Academy of Sciences of the Czech Republic

Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences

Maciej Piasecki – Wrocław University of Technology

Dragomir Radev – Yale University

Marko Tadić – University of Zagreb

Hristo Tanev – Joint Research Centre of the European Commission

Tinko Tinchev – Sofia University

Dan Tufis – Research Institute for Artificial Intelligence, Romanian Academy

Duško Vitas – University of Belgrade

Radka Vlahova – Sofia University

Victoria Yaneva – University of Wolverhampton

ORGANISING COMMITTEE

Svetlozara Leseva – Institute for Bulgarian Language (BAS)

Iana Atanassova – Centre Tesnière, Université de Franche-Comté, UFR SLHS

Rositsa Dekova – Plovdiv University, Faculty of Philology, Department of English Studies

Tsvetana Dimitrova – Institute for Bulgarian Language (BAS)

Dimitar Hristov – Institute for Bulgarian Language (BAS)

Alexander Popov – Institute of Information and Communication Technologies

Borislav Rizov – Institute for Bulgarian Language (BAS)

Katya Saint-Amand – Senior Linguist, Chenope

Valentina Stefanova – Institute for Bulgarian Language (BAS)

Ivelina Stoyanova – Institute for Bulgarian Language (BAS)

Ekaterina Tarpomanova – Sofia University, Faculty of Slavic Studies

Maria Todorova – Institute for Bulgarian Language (BAS)

Martin Yalamov – Institute for Bulgarian Language (BAS)

Victoria Yaneva – University of Wolverhampton

PLENARY TALKS

LINGUISTIC INTELLIGENCE: COMPUTERS VS. HUMANS

Prof. Ruslan Mitkov (University of Wolverhampton)

Computers are ubiquitous – they are and are used everywhere. But how good are computers at understanding and producing natural languages (e.g. English or Bulgarian)? In other words, what is the level of their linguistic intelligence? This presentation will examine the linguistic intelligence of the computers and will look at the challenges ahead...

I shall begin by a brief historical flashback. I shall plot the timeline of the linguistic intelligence of computers against that of humans. Natural Language Processing (NLP) advances in the last 20 years have made it possible for the linguistic intelligence of computers to increase significantly but they are still behind humans...

The presentation will explain why it is so difficult for computers to understand, generate and in general, to process natural language texts – it is a steep road/learning curve, it is long and winding road for both computers and researchers who seek to develop intelligent programs. The talk will also briefly present well-established NLP techniques computers follow when 'learning' to 'speak' our languages, including rule-based and knowledge-based methods initially and machine learning and deep learning methods more recently, the latter being regarded as highly promising. A selection of Natural Language Processing applications will be outlined next. Finally, a preview will be offered of selected slides from my plenary talk at CLIB'2018 (which will be given on the following day).

I am not a clairvoyant, but at some of my plenary talks I have been asked to predict how far will computers go... At the end of my presentation in Sofia I shall share with you what I predict for the future and in general, what my vision is.

WITH A LITTLE HELP FROM NLP: MY LANGUAGE TECHNOLOGY APPLICATIONS WITH IMPACT ON SOCIETY

Prof. Ruslan Mitkov (University of Wolverhampton)

The talk will present three original methodologies developed by the speaker, underpinning implemented Language Technology tools which are already having an impact on the following areas of society: e-learning, translation and interpreting and care for people with language disabilities.

The first part of the presentation will introduce an original methodology and tool for generating multiple-choice tests from electronic textbooks. The application draws on a variety of Natural Language Processing (NLP) techniques which include term extraction, semantic computing and sentence transformation. The presentation will include an evaluation of the tool which demonstrates that generation of multiple-choice tests items with the help of this tool is almost four times faster than manual construction and the quality of the test items is not compromised. This application benefits e-learning users (both teachers and students) and is an example of how NLP can have a positive societal impact, in which the speaker passionately believes.

The talk will go on to outline two other original recent projects which are also related to the application of NLP beyond academia. First, a project, whose objective is to develop next-generation translation memory tools for translators and, in the near future, for interpreters, will be briefly presented. Finally, an original methodology and system will be outlined which helps users with autism to read and better understand texts.

BUILDING CONVERSATIONAL ASSISTANTS USING DEEP LEARNING

Dr Zornitsa Kozareva (Google)

Over the years there has been a paradigm shift in how humans interact with machines. Today's users are no longer satisfied with seeing a list of relevant web pages, instead they want to complete tasks and take actions. This raises the questions: "How do we teach machines to become useful in a human-centered environment?" and "How do we build machines that help us organize our daily schedules, arrange our travel and be aware of our preferences and habits?". In this talk, I will describe these challenges in the context of conversational assistants. Then, I will delve into deep learning algorithms for entity extraction, user intent prediction and question answering. Finally, I will highlight findings on user intent prediction from shopping, movies, restaurant and sport domains.

NEURAL GRAPH LEARNING

Dr Sujith Ravi (Google)

Recent machine learning advances have enabled us to build intelligent systems that understand semantics from speech, natural language text and images. While great progress has been made in many AI fields, building scalable intelligent systems from "scratch" still remains a daunting challenge for many applications. To overcome this, we exploit the power of graph algorithms since they offer a simple elegant way to express different types of relationships observed in data and can concisely encode structure underlying a problem. In this talk I will focus on "How can we combine the flexibility of graphs with the power of machine learning?"

I will describe how we address these challenges and design efficient algorithms by employing graph-based machine learning as a computing mechanism to solve real-world prediction tasks. Our graph-based machine learning framework can operate at large scale and easily handle massive graphs (containing billions of vertices and trillions of edges) and make predictions over billions of output labels while achieving $O(1)$ space complexity per vertex. In particular, we combine graph learning with deep neural networks to power a number of machine intelligence applications, including Smart Reply, image recognition and video summarization to tackle complex language understanding and computer vision problems. I will also introduce some of our latest research and share results on "neural graph learning", a new joint optimization framework for combining graph learning with deep neural network models.

Table of Contents

Ruslan Mitkov <i>With a little help from NLP: My Language Technology applications with impact on society</i>	1
Vito Pirrelli <i>NLP-based Assessment of Reading Efficiency in Early Grade Children</i>	5
Mila Vulchanova and Valentin Vulchanov <i>Figurative language processing: A developmental and NLP Perspective</i>	7
Nikola Taushanov, Ivan Koychev and Preslav Nakov <i>Abstractive Text Summarization with Application to Bulgarian News Articles</i>	15
Maria Gritz <i>Towards Lexical Meaning Formal Representation by virtue of the NL-DL Definition Transformation Method</i>	23
Junya Morita <i>Narrow Productivity, Competition, and Blocking in Word Formation</i>	34
Cvetana Krstev, Ranka Stanković and Duško Vitas <i>Knowledge and Rule-Based Diacritic Restoration in Serbian</i>	41
Anton Zinoviev <i>Perfect Bulgarian Hyphenation, or How not to Stutter at End-of-line</i>	52
Anna Roitberg and Denis Khachko <i>Russian Bridging Anaphora Corpus</i>	62
Ekaterina Tarpomanova <i>Aspectual and Temporal Characteristics of the Past Active Participles in Bulgarian – a Corpus-based Study</i>	69
Olena Siruk and Ivan Derzhanski <i>Unmatched Femininitives in a Corpus of Bulgarian and Ukrainian Parallel Texts</i>	77
Viktoriya Petrova <i>The Bulgarian Summaries Corpus</i>	85
Natalia Loukachevitch and Boris Dobrov <i>Ontologies for Natural Language Processing: Case of Russian</i>	93
Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas and Cvetana Krstev <i>Resource-based WordNet Augmentation and Enrichment</i>	104
Svetlozara Leseva, Ivelina Stoyanova and Maria Todorova <i>Classifying Verbs in WordNet by Harnessing Semantic Resources</i>	115

Maria Mitrofan, Verginica Barbu Mititelu and Grigorina Mitrofan <i>A Pilot Study for Enriching the Romanian WordNet with Medical Terms</i>	126
Ivelina Stoyanova <i>Factors and Features Determining the Inheritance of Semantic Primes between Verbs and Nouns within WordNet</i>	135
Borislav Rizov and Tsvetana Dimitrova <i>Online Editor for WordNets</i>	146
Amir Bakarov <i>The Effect of Unobserved Word-Context Co-occurrences on a Vector-Mixture Approach for Compositional Distributional Semantics</i>	153
Rositsa Dekova and Adelina Radeva <i>Introducing Computational Linguistics and NLP to High School Students</i>	162
Ivan Derzhanski and Milena Veneva <i>Linguistic Problems on Number Names</i>	169
Marina Dzhonova, Kjetil Rå Hauge and Yovka Tisheva <i>Parallel Web Display of Transcribed Spoken Bulgarian with its Normalised Version and an Indexed List of Lemmas</i>	177
Georgi Dzhumayov <i>Integrating Crowdsourcing in Language Learning</i>	185
Todor Lazarov <i>Bulgarian–English Parallel Corpus for the Purposes of Creating Statistical Translation Model of the Verb Forms. General Conception, Structure, Resources and Annotation</i>	193
Branislava Šandrih <i>Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting</i>	203