Computational
Linguistics in
Bulgaria

clib '16

SECOND INTERNATIONAL CONFERENCE

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2016

9 September **2016**

Sofia, Bulgaria

Department of
Computational
Linguistics

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

The Second International Conference *Computational Linguistics in Bulgaria (CLIB 2016)* is organised within the Operation for Support for International Scientific Conferences Held in Bulgaria of the National Science Fund Grant № ДПМНФ 01/9 of 11 Aug 2016.

National Science Fund

CLIB 2016 is organised by:

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

PUBLICATION AND CATALOGUING INFORMATION

Title:	**Proceedings of the Second International Conference *Computational Linguistics in Bulgaria (CLIB 2016)***
ISSN:	**2367-5675 (online)**
Published and distributed by:	**The Institute for Bulgarian Language *Prof. Lyubomir Andreychin*, Bulgarian Academy of Sciences**
Editorial address:	**Institute for Bulgarian Language *Prof. Lyubomir Andreychin*, Bulgarian Academy of Sciences** **52 Shipchenski prohod blvd., bldg. 17** **Sofia 1113, Bulgaria** **+359 2/ 872 23 02**

Proceedings of the
Second International Conference
Computational Linguistics in Bulgaria

9 September 2016
Sofia, Bulgaria

PREFACE

We are excited to welcome you to the second edition of the International Conference *Computational Linguistics in Bulgaria* (CLIB 2016) in Sofia, Bulgaria!

CLIB aspires to foster the NLP community in Bulgaria and further the cooperation among researchers working in NLP for Bulgarian around the world. The need for a conference dedicated to NLP research dealing with or applicable to Bulgarian has been felt for quite some time. We believe that building a strong community of researchers and teams who have chosen to work on Bulgarian is a key factor to meeting the challenges and requirements posed to computational linguistics and NLP in Bulgaria. We share the hope that CLIB will establish itself as an international forum for sharing high-quality scientific work in all areas of computational linguistics and NLP and will grow in scope and scale with each new edition. The CLIB community will be dedicated to supporting the creation and improvement of advanced NLP resources, tools and technologies for mono- and multilingual language processing, machine translation and translation aids, content creation, localisation and personalisation, speech recognition and generation, information retrieval and information extraction. The Conference was made possible due to the hard work of many people.

We would like to thank the authors who trusted us and submitted their contributions to CLIB 2016. Their efforts and high-quality research are the chief factor that enabled us to create an interesting and solid scientific programme. We would also like to thank our industrial participants for sharing their insights, ideas and know-how with the research community.

We would like to express our sincere gratitude to the members of the Programme Committee, who accepted to join us and invested a lot of expertise to provide valuable feedback to the authors. Special thanks are due to Prof. Svetla Kœva, who is the person behind the whole CLIB concept. We hope that CLIB 2016 will be a useful and productive experience that we all will enjoy!

CLIB 2016 Organising Committee

PROGRAMME COMMITTEE

Svetla Koeva (Chair) – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Cvetana Krstev – University of Belgrade
Denis Maurel – François-Rabelais University of Tours
Dragomir Radev – University of Michigan, Department of Electrical Engineering and Computer Science
Duško Vitas – University of Belgrade
Éric Laporte – University of Paris-Est Marne-la-Vallée
Hristo Krushkov – Plovdiv University *Paisii Hilendarski*
Hristo Tanev – Joint Research Centre of the European Commission, Ispra, Italy
Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Ivelina Nikolova – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Jan Šnajder – University of Zagreb, TakeLab
Karel Oliva – Institute of the Czech Language, Academy of Sciences of the Czech Republic
Kjetil Rå Hauge – University of Oslo
Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences
Maciej Piasecki – Wrocław University of Technology
Mariana Damova – *Mozaika* Ltd., Bulgaria
Marko Tadić – University of Zagreb
Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology
Nikolay Vazov – University of Oslo
Preslav Nakov – Qatar Computing Research Institute, Hamad bin Khalifa University
Radka Vlahova – Sofia University *St. Kliment Ohridski*
Radovan Garabík – *Ľudovít Štúr* Institute of Linguistics, Slovak Academy of Sciences
Ruslan Mitkov – University of Wolverhampton
Stoyan Mihov – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Tania Avgustinova – Saarland University
Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy
Zornitsa Kozareva – Yahoo! Labs

ORGANISING COMMITTEE

Svetlozara Leseva – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Tsvetana Dimitrova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Ivelina Stoyanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Maria Todorova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Valentina Stefanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Borislav Rizov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Dimitar Hristov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Martin Yalamov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Ekaterina Tarpomanova – Sofia University *St. Kliment Ohridski*
Rositsa Dekova – Plovdiv University *Paisii Hilendarski*

INVITED TALK

Dr. Preslav Nakov
(Qatar Computing Research Institute, HBKU)

Exposing Paid Opinion Manipulation Trolls in News Community Forums

The practice of using opinion manipulation trolls has been reality since the rise of Internet and community forums. It has been shown that user opinions about products, companies and politics can be influenced by posts by other users in online forums and social networks. This makes it easy for companies and political parties to gain popularity by paying for "reputation management" to people or companies that write in discussion forums and social networks fake opinions from fake profiles.

During the 2013-2014 Bulgarian protests against the Oresharski cabinet, social networks and news community forums became the main "battle grounds" between supporters and opponents of the government. In that period, there was a very notable presence and activity of government supporters in Web forums. In series of leaked documents in the independent Bulgarian media Bivol, it was alleged that the ruling Socialist party was paying Internet trolls with EU Parliament money. Allegedly, these trolls were hired by a PR agency and were given specific instructions what to write.

A natural question is whether such trolls can be found and exposed automatically. This is a very hard task, as there is no enough data to train a classifier; yet, it is possible to obtain some test data, as these trolls are sometimes caught and widely exposed (e.g., by Bivol). Yet, one still needs training data. We solve the problem by assuming that a user who is called a troll by several different people is likely to be one, and one who has never been called a troll is unlikely to be such. We compare the profiles of (i) paid trolls vs. (ii) "mentioned" trolls vs. (iii) non-trolls, and we further show that a classifier trained to distinguish (ii) from (iii) does quite well also at telling apart (i) from (iii).

KEYNOTE TALK

Prof. Dragomir Radev
(Department of Electrical Engineering and Computer Science, University of Michigan)

Natural Language Processing for Collective Discourse

Natural Language Processing (NLP) has become very popular in recent years thanks to new technologies like IBM's Watson, Apple's Siri, Google Translate, and Yahoo's text summarization system. One of the fundamental challenges in NLP is to automatically recognize similar words and sentences. I will talk about research done in the Computational Linguistics And Information Retrieval lab (CLAIR) on graph-based methods for similarity recognition and its applications to NLP tasks. These projects are related to Collective Discourse (text collections produced by large numbers of users) and its inherent properties such as centrality and diversity. In the first project we team up with the New Yorker magazine. Each week a captionless cartoon is published in the magazine and thousands of readers try to come up with funny captions for it. In our work, we try to uncover the topics of the jokes in the submitted captions. The second project is about analysing a corpus of word clues used in New York Times crossword puzzles. We compare different clustering methods for word sense disambiguation using these crossword clues. The third project is about the automatic generation of citation-based summaries of research articles. These summaries describe what readers of the papers find most important in the cited papers. If there is time, I will also briefly mention some applications to bioinformatics, political science, and social network analysis.

Table of Contents