FIFTH
INTERNATIONAL
CONFERENCE

clib '22

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2022

8 – 9 September **2022**
Sofia, Bulgaria

Organiser:

1869

Department of Computational Linguistics
Institute for Bulgarian Language
Institute of Information and Communication Technologies
BULGARIAN ACADEMY OF SCIENCES

PROCEEDINGS

The Fifth International Conference *Computational Linguistics in Bulgaria* (CLIB 2022)is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. KP-06-MNF/7 of 20.07.2022.

The National Science Fund does not take responsibility for the contents of the papers presented at the Conference or for any of the Conference materials.

CLIB 2022 is organised by:

Department of Computational Linguistics
Institute for Bulgarian Language

Institute for Information and Communication Technologies

Bulgarian Academy of Sciences

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Fifth International Conference *Computational Linguistics in Bulgaria* (CLIB 2022)
ISSN:	2367 5675 (online)
Published and distributed:	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

Proceedings of the

Fifth International Conference

Computational Linguistics in Bulgaria

8 – 9 September 2022
Sofia, Bulgaria

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language, Bulgarian Academy of Sciences

Co-chair:

Petya Osenova – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences / Sofia University, Faculty of Slavic Studies

Iana Atanassova – University of Burgundy, Centre for Interdisciplinary and Transcultural Research, France

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

Svetla Boytcheva – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Khalid Choukri – Evaluations and Language Resources Distribution Agency, France

Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Tsvetana Dimitrova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

A. Seza Doğruöz – Ghent University, Belgium

Radovan Garabík – Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Maria Gavrilidou – Institute for Language and Speech Processing, Natural Language Processing and Knowledge Extraction Department, Greece

Stefan Gerdjikov – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

Voula Giouli – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

Ivan Koychev – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

Cvetana Krstev – University of Belgrade, Faculty of Philology, Serbia

Eric Laporte – University of Paris-Est Marne-la-Vallée, France

Natalia Loukachevitch – Research Computing Center of Moscow State University, Russia

John P. McCrae – National University of Ireland, Galway, Ireland

Preslav Nakov – Qatar Computing Research Institute, HBKU, Qatar

Maciej Piasecki – Wrocław University of Technology, Poland

Vito Pirrelli – Institute for Computational Linguistics, ILC-CNR, Italy

Ewa Rudnicka – Wrocław University of Technology, Poland

Ivelina Stoyanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Stan Szpakowicz – University of Ottawa, Canada

Marko Tadić – University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Croatia

Hristo Tanev – Joint Research Centre of the European Commission, Italy

Irina Temnikova – Big Data for Smart Society Institute (GATE), Bulgaria

Tinko Tinchev – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

Maria Todorova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Cristina Vertan – University of Hamburg, Germany

Katerina Zdravkova – University St Cyril and Methodius in Skopje, North Macedonia

ORGANISING COMMITTEE

Chair:

Svetlozara Leseva – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

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

Dimitar Hristov – Cleversoft, Bulgaria

Georgi Iliev – Milestone Systems, Bulgaria

Hristina Kukova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Todor Lazarov – New Bulgarian University

Valentina Stefanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Ekaterina Tarpomanova – Sofia University, Faculty of Slavic Studies

Table of Contents

PLENARY TALKS

The Hebrew Essay Corpus

Prof. Shuly Wintner (University of Haifa, Israel)

The Hebrew Essay Corpus is an annotated corpus of Hebrew language argumentative essays authored by prospective higher-education students. The corpus includes both essays by native speakers, written as part of the psychometric exam that is used to assess their future success in academic studies; and essays authored by non-native speakers, with three different native languages, that were written as part of a language aptitude test. The corpus is uniformly encoded and stored. The non-native essays were annotated with target hypotheses whose main goal is to make the texts amenable to automatic processing (morphological and syntactic analysis).

I will describe the corpus and the error correction and annotation schemes used in its analysis. In addition, I will discuss some of the challenges involved in identifying and analyzing non-native language use in general, and propose various ways for dealing with these challenges. Then, I will present classifiers that can accurately distinguish between native and non-native authors; determine the mother tongue of the non-natives; and predict the proficiency level of non-native Hebrew learners. This is important for practical (mainly educational) applications, but the endeavor also sheds light on the features that support the classification, thereby improving our understanding of learner language in general, and transfer effects from Arabic, French, and Russian on nonnative Hebrew in particular.

Detect – Verify – Communicate: Combating Misinformation with More Realistic NLP

Prof. Iryna Gurevych (Technical University of Darmstadt, Germany)

Dealing with misinformation is a grand challenge of the information society directed at equipping the computer users with effective tools for identifying and debunking misinformation. Current Natural Language Processing (NLP) including its fact-checking research fails to meet the expectations of real-life scenarios. In this talk, we show why the past work on fact-checking has not yet led to truly useful tools for managing misinformation, and discuss our ongoing work on more realistic solutions. NLP systems are expensive in terms of financial cost, computation, and manpower needed to create data for the learning process. With that in mind, we are pursuing research on detection of emerging misinformation topics to focus human attention on the most harmful, novel examples. Automatic methods for claim verification rely on large, high-quality datasets. To this end, we have constructed two corpora for fact checking, considering larger evidence documents and pushing the state of the art closer to the reality of combating misinformation. We further compare the capabilities of automatic, NLP-based approaches to what human fact checkers actually do, uncovering critical research directions for the future. To edify false beliefs, we are collaborating with cognitive scientists and psychologists to automatically detect and respond to attitudes of vaccine hesitancy, encouraging anti-vaxxers to change their minds with effective communication strategies.

Lexical Conceptual Resources in the Era of Neural Language Models

Prof. Bolette Sandford Pedersen (Copenhagen University, Denmark)

Lexical conceptual resources in terms of e.g. wordnets, framenets, terminologies and ontologies have been compiled for many languages during the last decades in order to provide NLP systems with formally expressed information about the semantics of words and phrases, and about how they refer to the world. In most recent years, neural language models have become a game-changer in the NLP field – based, as they are, solely on text from large corpora. It is time we ask ourselves: What is the role of lexical conceptual resources in the era of neural language models? The claim of my talk is that they still play a crucial role since NLP systems based on textual distribution alone will always to some extent be insufficient and biased. Through my own work, which has over the years taken place in close collaboration with leading lexicographers in Denmark, I will illustrate how such conceptual resources can be compiled based on existing high-quality and continuously updated lexicographical resources and how they can be further curated by examining the distributional patterns captured in word embeddings.

Towards AI that Reasons with Scientific Text and Images

Jose Manuel Gomez-Perez (Expert.ai)

Reading a textbook in a particular discipline and being able to answer the questions at the end of each chapter is one of the grand challenges of artificial intelligence, which requires advances in language, vision, problem-solving, and learning theory. Such challenges are best illustrated in the scientific domain, where complex information is presented over a variety of modalities involving not only language but also visual information, like diagrams and figures.

In this talk, we will analyze the specific challenges entailed in understanding scientific documents and share some of the recent advances in the area that enable the development of AI systems capable to answer scientific questions. In addition, we will reflect on what new developments will be required to address the next grand challenge: to create an AI system that can make major scientific discoveries by itself.