FOURTH
INTERNATIONAL
CONFERENCE

clib '20

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2020

25 – 26 June **2020**

Sofia, Bulgaria

PROCEEDINGS

The Special Session on Wordnets and Ontologies at the Fourth International Conference *Computational Linguistics in Bulgaria (CLIB 2020)* 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.

NATIONAL
SCIENCE
FUND
Ministry of Education and Science

CLIB 2020 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 Fourth International Conference *Computational Linguistics in Bulgaria* (CLIB 2020)
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

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 © 2020

Proceedings of the

Fourth International Conference

Computational Linguistics in Bulgaria

Computational Linguistics in Bulgaria ‘20

25 – 26 June 2020
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

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

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

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

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

Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology

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

Stefan Gerdjikov – Sofia University, Faculty of Mathematics and Informatics

Kjetil Rå Hauge – University of Oslo, Department of Literature, Area Studies and European Languages, ILOS

Ivan Koychev – Sofia University, Faculty of Mathematics and Informatics

Zornitsa Kozareva – Google

Cvetana Krstev – University of Belgrade, Faculty of Philology

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

Bernardo Magnini – Bruno Kessler Center in Information and Communication Technology

Ruslan Mitkov – University of Wolverhampton

Preslav Nakov – Qatar Computing Research Institute, HBKU

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

Kemal Oflazer – Carnegie Mellon University in Qatar

Maciej Piasecki – Wrocław University of Technology

Vito Pirrelli – Institute for Computational Linguistics, ILC-CNR

Stan Szpakowicz – University of Ottawa

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

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

Hristo Tanev

Irina Temnikova – Mitra Translations

Tinko Tinchev – Sofia University, Faculty of Mathematics and Informatics

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

Dan Tufis – Research Institute for Artificial Intelligence, Romanian Academy

Cristina Vertan – University of Hamburg

Victoria Yaneva – University of Wolverhampton

Katerina Zdravkova – University St Cyril and Methodius in Skopje

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

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

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

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

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

SPECIAL SESSION ON WORDNETS AND ONTOLOGIES 136

PLENARY TALKS

TAG SENSE DISAMBIGUATION IN LARGE IMAGE COLLECTIONS: IS IT POSSIBLE?

Prof. D.Sc. Galia Angelova (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences)

Automatic identification of intended tag meanings is a challenge in large annotated image collections where human authors assign tags inspired by emotional or professional motivations. This task can be viewed as part of the AI-complete problem to integrate language and vision. Algorithms for automatic Tag Sense Disambiguation (TSD) need "golden" collections of manually created tags to establish baselines for accuracy assessment. In this talk the TSD task will be presented with its background, complexity and possible solutions. An approach to use WordNet senses and Lesk algorithm proves to be successful but the evaluation was done manually for a small number of tags. Another experiment with the MIRFLICKR-25000 image collection will be presented as well. Word embeddings create a specific baseline so the results can be compared. The accuracy achieved in this exercise is 78.6%.

By improving TSD and obtaining high quality synsets for the image tags, we are actually supporting the machine translation of the large annotated image collections to languages other than English.

CLINICAL NATURAL LANGUAGE PROCESSING IN BULGARIAN

Assoc. Prof. Svetla Boytcheva (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences)

Healthcare is a data intense domain. A large amount of patient data is generated daily. However, more than 80% of this information is stored in an unstructured format – as clinical texts. Usually, clinical narratives contain a description with telegraph-style sentences, ambiguous abbreviations, many typographical errors, lack of punctuation, concatenated words, and etc. Especially in the Bulgarian context – medical texts contain terminology both in Bulgarian, Latin and transliterated Latin terminology in Cyrillic, that makes the task for text analytics more challenging. Recently, with the improvement of the quality of natural language processing (NLP), it is increasingly recognized as the most useful tool for extracting clinical information from free text in scientific medical publications and clinical records. Natural language processing (NLP) of non-English clinical text is quite a challenge because of the lack of resources and NLP tools. International medical ontologies such as SNOMED, MeSH (Medical Subject Headings), and the UMLS (Unified Medical Languages System) are not yet available in most languages. This necessitates the development of new methods for processing clinical information and for semi-automatically generating medical language resources. This is not an easy task because of the lack of a sufficiently accessible repositories with medical records, due to the specific nature of the content, which contains a lot of personal data and specific regulations for their access.

In this talk will be discussed the multilingual aspects of automation Extract text from clinical narratives in the Bulgarian language. This is very important task for medical informatics, because it allows the automatic structuring of patient information and the generation of databases that can be further investigated by retrieving data to search for complex relationships. The results can help improve clinical decision support, diagnosis and treatment support systems.

DETECTING THE FAKE NEWS AT ITS SOURCE, MEDIA LITERACY, AND REGULATORY COMPLIANCE

Dr. Preslav Nakov (Qatar Computing Research Institute, Hamad Bin Khalifa University)

Given the recent proliferation of disinformation online, there has been also growing research interest in automatically debunking rumors, false claims, and “fake news”. A number of fact-checking initiatives have been launched so far, both manual and automatic, but the whole enterprise remains in a state of crisis: by the time a claim is finally fact-checked, it could have reached millions of users, and the harm caused could hardly be undone. An arguably more promising direction is to focus on fact-checking entire news outlets, which can be done in advance. Then, we could fact-check the news before they were even written: by checking how trustworthy the outlets that published them are.

We will show how we do this in the Tanbih news aggregator (`http://www.tanbih.org/`), which aims to limit the effect of “fake news”, propaganda and media bias by making users aware of what they are reading. The project’s primary aim is to promote media literacy and critical thinking, which are arguably the best way to address disinformation and “fake news” in the long run. In particular, we develop media profiles that show the general factuality of reporting, the degree of propagandistic content, hyper-partisanship, leading political ideology, general frame of reporting, stance with respect to various claims and topics, as well as audience reach and audience bias in social media. We further offer explainability by automatically detecting and highlighting the instances of use of specific propaganda techniques in the news (`https://www.tanbih.org/propaganda`).

Finally, we will show how this research can support broadcasters and content owners with their regulatory measures and compliance processes. This is a direction we recently explored as part of our TM Forum & IBC 2019 award-winning Media-Telecom Catalyst project on AI Indexing for Regulatory Compliance, which QCRI developed in partnership with Al Jazeera, Associated Press, RTE Ireland, Tech Mahindra, V-Nova, and Metaliquid.

DEMONSTRATION OF THE EUROPEAN LANGUAGE GRID

Dr. Georg Rehm (Speech and Language Technology Lab, German Research Center for Artificial Intelligence

With 24 official EU and many additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by hundreds of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented – by nation states, languages, verticals and sectors, significantly holding back its impact. The European Language Grid (ELG) project addresses this fragmentation by establishing the ELG as the primary platform for LT in Europe. The ELG is a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data sets and resources. Once fully operational, it will enable the commercial and non-commercial European LT community to deposit and upload their technologies and data sets into the ELG, to deploy them through the grid, and to connect with other resources. The ELG will boost the Multilingual Digital Single Market towards a thriving European LT community, creating new jobs and opportunities. Furthermore, the ELG project organises two open calls for up to 20 pilot projects, one of which was recently closed. The presentation will give an overview of the European Language Grid project and it will also contain a demonstration of the emerging ELG technology platform.