EACL 2021

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing,

Co-located with the 16th European Chapter of the Association for Computational Linguistics (EACL)

April 20, 2021

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-14-5

Preface

This volume contains the papers presented at BSNLP-2021: the eighth workshop on Balto-Slavic Natural Language Processing. The workshop is organized by ACL SIGSLAV, the Special Interest Group on NLP in Slavic Languages of the Association for Computational Linguistics.

The BSNLP workshops have been convening for over a decade, with a clear vision and purpose. On one hand, the languages from the Balto-Slavic group play an important role due to their widespread use and diverse cultural heritage. These languages are spoken by about one-third of all speakers of the official languages of the European Union, and by over 400 million speakers worldwide. The political and economic developments in Central and Eastern Europe place societies where Balto-Slavic languages are spoken at the center of rapid technological advancement and growing European consumer markets.

On the other hand, research on theoretical and applied NLP in some of these languages still lag behind the "major" languages, such as English and other Western European languages. In comparison to English, which has dominated the digital world since the advent of the Internet, many of these languages still lack resources, processing tools and applications—especially those with smaller speaker bases.

The Balto-Slavic languages pose a wealth of fascinating scientific challenges. The linguistic phenomena specific to the Balto-Slavic languages—complex morphology and free word order—present non-trivial problems for the construction of NLP tools, and require rich morphological and syntactic resources.

The BSNLP workshop aims to bring together researchers in NLP for Balto-Slavic languages from academia and industry. We aim to stimulate research, foster the creation of tools and dissemination of new results. The workshop serves as a forum for the exchange of ideas and experience and for discussing shared problems. One fascinating aspect of Slavic and Baltic languages is their structural similarity, as well as an easily recognizable lexical and inflectional inventory spanning the groups, which—despite the lack of mutual intelligibility—creates a special environment in which researchers can fully appreciate the shared problems and solutions.

In order to stimulate research and collaboration further, we have organized the third BSNLP Challenge: a shared task on multilingual named entity recognition. Due to rich inflection, free word order, derivation, and other phenomena present in the Slavic languages, work on named entities poses a challenging task. Fostering research and development on the problems of named entities — detecting mentions of names, lemmatization (normalization), classification, and cross-lingual matching — is crucial for cross-lingual information access and wider use of NLP in Slavic languages. This third edition of the shared task covered six languages: Bulgarian, Czech, Polish, Russian, Slovene, and Ukrainian. If further covered five types of named entities: persons, locations, organizations, events, and products.

We received 11 regular paper submissions, and we selected 8 of them for presentation. The papers cover topics ranging from pre-training BERT-style transformers for Slavic languages to paraphrasing, text simplification, sentiment analysis, abusive language recognition, and inappropriate language detection.

Ten teams registered to participate in the NE Challenge, of which eight submitted results, and six also submitted system description papers. These papers are included in this volume, and their work is discussed in the special session dedicated to the Challenge.

This workshop's presentations—the regular Workshop papers and the Shared Task Challenge—cover at least ten Balto-Slavic languages: Bosnian, Bulgarian, Croatian, Czech, Montenegrin, Polish, Russian, Serbian, Slovene, Ukrainian.

This workshop continues the proud tradition established by the earlier BSNLP workshops, which were held in conjunction with the following venues:

1. ACL 2007 Conference in Prague, Czech Republic;
2. IIS 2009: Intelligent Information Systems, in Kraków, Poland;
3. TSD 2011: 14th International Conference on Text, Speech and Dialogue in Plzeň, Czech Republic;
4. ACL 2013 Conference in Sofia, Bulgaria;
5. RANLP 2015 Conference in Hissar, Bulgaria;
6. EACL 2017 Conference in Valencia, Spain.
7. ACL 2019 Conference in Florence, Italy

We sincerely hope that this work will help stimulate further growth of our rich and exciting field.

The BSNLP'2021 Organizers: Bogdan Babych, Olga Kanishcheva, Preslav Nakov, Jakub Piskorski, Lidia Pivovarova, Vasyl Starko, Josef Steinberger, Roman Yangarber, Michał Marcińczuk, Senja Pollak, Pavel Přibáň, Marko Robnik-Šikonja

Organizers:

Bogdan Babych (Heidelberg University, Germany)
Olga Kanishcheva (Kharkiv Polytechnic Institute, Ukraine)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Jakub Piskorski (Joint Research Centre of the European Commission, Ispra, Italy)
Lidia Pivovarova (University of Helsinki, Finland)
Vasyl Starko (Ukrainian Catholic University, Ukraine)
Josef Steinberger (University of West Bohemia, Czech Republic)
Roman Yangarber (University of Helsinki, Finland)
Michał Marcińczuk (Wrocław University of Technology, Poland)
Senja Pollak (Jožef Stefan Institute, Ljubljana, Slovenia)
Pavel Přibáň (University of West Bohemia, Czech Republic)
Marko Robnik-Šikonja (University of Ljubljana, Slovenia)

Program Committee:

Željko Agić (Corti ApS, Copenhagen, Denmark)
Radovan Garabik (Comenius University in Bratislava, Slovakia)
Tomas Krilavičius (Vytautas Magnus University, Kaunas, Lithuania)
Cvetana Krstev (University of Belgrade, Serbia)
Vladislav Kubon (Charles University, Prague, Czech Republic)
Nikola Ljubešić (Jožef Stefan Institute, Ljubljana, Slovenia)
Natalia Loukachevitch (Lomonosov Moscow State University Moscow, Russia)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Maciej Ogrodniczuk (Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Alexander Panchenko (Skoltech, Moscow, Russia)
Jakub Piskorski (Joint Research Centre, Ispra, Italy/PAS, Warsaw, Poland)
Lidia Pivovarova (University of Helsinki, Finland)
Senja Pollak (Jožef Stefan Institute, Ljubljana, Slovenia)
Marko Robnik-Šikonja (University of Ljubljana, Slovenia)
Alexandr Rosen (Charles University, Prague)
Tanja Samardžić (University of Geneva, Switzerland)
Agata Savary (University of Tours, France)
Serge Sharoff (University of Leeds, UK)
Stan Szpakowicz (University of Ottawa, Canada)
Irina Temnikova (Sofia University, Bulgaria)
Ivan Vulić (University of Cambridge, UK)
Miloš Jakubíček (Lexical Computing, Czech Republic UK)
Roman Yangarber (University of Helsinki, Finland)
Daniel Zeman (Charles University, Czech Republic)

Table of Contents

Workshop Program

HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish
Robert Mroczkowski, Piotr Rybak, Alina Wróblewska and Ireneusz Gawlik

Russian Paraphrasers: Paraphrase with Transformers
Alena Fenogenova

Abusive Language Recognition in Russian
Kamil Saitov and Leon Derczynski

Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company's Reputation
Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov and Alexander Panchenko

BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian
Nikola Ljubešić and Davor Lauc

RuSentEval: Linguistic Source, Encoder Force!
Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel and Ekaterina Artemova

Exploratory Analysis of News Sentiment Using Subgroup Discovery
Anita Valmarska, Luis Cabrera-Diego, Elvys Linhares Pontes and Senja Pollak

Creating an Aligned Russian Text Simplification Dataset from Language Learner Data
Anna Dmitrieva and Jörg Tiedemann

Multilingual Named Entity Recognition and Matching Using BERT and Dedupe for Slavic Languages
Marko Prelevikj and Slavko Zitnik

Priberam Labs at the 3rd Shared Task on SlavNER
Pedro Ferreira, Ruben Cardoso and Afonso Mendes

Multilingual Slavic Named Entity Recognition
Rinalds Vīksna and Inguna Skadina

Using a Frustratingly Easy Domain and Tagset Adaptation for Creating Slavic Named Entity Recognition Systems
Luis Adrián Cabrera-Diego, Jose G. Moreno and Antoine Doucet

No Day Set (continued)