BSNLP'2019

The 7th Workshop on Balto-Slavic Natural Language Processing

Proceedings of the Workshop

BSNLP'2019
August 2, 2019
Florence, Italy

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 978-1-950737-41-3

Preface

This volume contains the papers presented at BSNLP-2019: the Seventh 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 second BSNLP Challenge: a shared task on multilingual named entity recognition. We have built a new and significantly larger dataset than for the first shared task, organized in 2017. The data allows systems to be evaluated on recognizing mentions of named entities (NEs) in documents, lemmatization of NEs, and cross-lingual linking of NEs. This edition of the Challenge covers four Slavic languages—Bulgarian, Czech, Polish and Russian—and five named entity types, namely: persons, organizations, locations, events, and products.

We received 20 regular paper submissions, 11 of which were accepted for presentation.

The papers cover a range of topics. Two papers are related to lexical semantics, four to the development of linguistic resources, four to information filtering, information retrieval, and information extraction. Another group of four papers cover topics related to the processing of non-standard language or user-generated content. Finally, one paper describes the NE Challenge.

Sixteen teams expressed interest in participating in the NE Challenge, of which eight submitted results. Seven teams worked on NE recognition in all four languages. Five of the teams that participated in the shared task also submitted system description papers. They 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 nine Balto-Slavic languages: Bulgarian, Croatian, Czech, Polish, Russian, Slovak, Slovene, Serbian and 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.

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

The BSNLP'2019 Organizers

Organizers:

Tomaž Erjavec, Jožef Stefan Institute, Slovenia
Michał Marcińczuk, Wroclaw University of Science and Technology, Poland
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
Jan Šnajder, University of Zagreb, Croatia
Josef Steinberger, University of West Bohemia, Czech Republic
Roman Yangarber, University of Helsinki, Finland

Program Committee:

Željko Agić, Corti ApS, Copenhagen, Denmark
Senka Drobac, University of Helsinki, Finland
Tomaž Erjavec, Jožef Stefan Institute, Slovenia
Radovan Garabik, Comenius University in Bratislava, Slovakia
Goran Glavaš, University of Mannheim, Germany
Maxim Gubin, Facebook Inc., USA
Miloš Jakubíček, Masaryk University, Brno, Czech Republic / Lexical Computing, Brighton, UK
Tomas Krilavičius, Vytautas Magnus University, Kaunas, Lithuania
Cvetana Krstev, University of Belgrade, Serbia
Vladislav Kuboň, Charles University, Prague, Czech Republic
Nikola Ljubešić, Jožef Stefan Institute, Ljubljana, Slovenia
Olga Mitrofanova, St. Petersburg State University, Russia
Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar
Maciej Ogrodniczuk, Polish Academy of Sciences, Poland
Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
Rūta Petrauskaitė, Vytautas Magnus University, Lithuania
Maciej Piasecki, Wroclaw University of Science and Technology, Poland
Jakub Piskorski, Joint Research Centre, Ispra, Italy/PAS, Warsaw, Poland
Lidia Pivovarova, University of Helsinki, Finland
Maja Popovic, Dublin City University, Ireland
Alexandr Rosen, Charles University, Prague
Tanja Samardžić, University of Geneva, Switzerland
Agata Savary, University of Tours, France
Kiril Simov, Bulgarian Academy of Sciences, Bulgaria
Inguna Skadiņa, University of Latvia, Latvia
Jan Šnajder, University of Zagreb, Croatia
Marko Robnik Šikonja, University of Ljubljana, Slovenia
Serge Sharoff, University of Leeds, UK
Josef Steinberger, University of West Bohemia, Czech Republic
Pavel Stranak, Charles University, Czech Republic
Stan Szpakowicz, University of Ottawa, Canada
Hristo Tanev, Joint Research Centre, Italy
Irina Temnikova, Sofia University Bulgaria
Andrius Utka Vytautas, Magnus University, Lithuania

Barbora Vidová Hladká, Charles University, Czech Republic
Roman Yangarber, University of Helsinki, Finland
Daniel Zeman, Charles University, Czech Republic

Invited Speaker:

Ivan Vulić, University of Cambridge, UK

Table of Contents

Workshop Program

Thursday, August 2, 2019

8:50–9:00 Opening Remarks

9:00–10:30 Session I: Morphology

9:00–9:25 *Unsupervised Induction of Ukrainian Morphological Paradigms for the New Lexicon: Extending Coverage for Named Entities and Neologisms using Inflection Tables and Unannotated Corpora*
Bogdan Babych

9:25–9:50 *Multiple Admissibility: Judging Grammaticality using Unlabeled Data in Language Learning*
Anisia Katinskaia and Sardana Ivanova

9:50–10:10 *Numbers Normalisation in the Inflected Languages: a Case Study of Polish*
Rafał Poświata and Michał Perełkiewicz

10:10–10:30 *What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian*
Nikola Ljubešić and Kaja Dobrovoljc

10:30–11:00 Coffee Break

11:00–12:25 Session II: Development of Linguistic Resources

11:00–11:25 *AGRR 2019: Corpus for Gapping Resolution in Russian*
Maria Ponomareva, Kira Droganova, Ivan Smurov and Tatiana Shavrina

11:25–11:45 *Creating a Corpus for Russian Data-to-Text Generation Using Neural Machine Translation and Post-Editing*
Anastasia Shimorina, Elena Khasanova and Claire Gardent

11:45–12:05 *Data Set for Stance and Sentiment Analysis from User Comments on Croatian News*
Mihaela Bošnjak and Mladen Karan

12:05–12:25 *A Dataset for Noun Compositionality Detection for a Slavic Language*
Dmitry Puzyrev, Artem Shelmanov, Alexander Panchenko and Ekaterina Artemova

Thursday, August 2, 2019 (continued)

12:25–13:30 Lunch

13:30–14:30 Session III: Keynote

13:30–14:30 *Cross-Lingual Word Embeddings in (Less Than) 60 Minutes*
Ivan Vulić

14:30–15:30 Session IV: Shared Task – Part I

14:30–14:45 *The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages*
Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger and Roman Yangarber

14:45–15:00 *BSNLP2019 Shared Task Submission: Multisource Neural NER Transfer*
Tatiana Tsygankova, Stephen Mayhew and Dan Roth

15:00–15:15 *TLR at BSNLP2019: A Multilingual Named Entity Recognition System*
Jose G. Moreno, Elvys Linhares Pontes, Mickael Coustaty and Antoine Doucet

15:15–15:30 *Tuning Multilingual Transformers for Language-Specific Named Entity Recognition*
Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov and Alexey Sorokin

15:30–16:00 Coffee Break

Thursday, August 2, 2019 (continued)

16:00–16:30 Session V: Shared Task – Part II

16:00–16:15 *Multilingual Named Entity Recognition Using Pretrained Embeddings, Attention Mechanism and NCRF*
Anton Emelyanov and Ekaterina Artemova

16:15–16:30 *JRC TMA-CC: Slavic Named Entity Recognition and Linking. Participation in the BSNLP-2019 shared task*
Guillaume Jacquet, Jakub Piskorski, Hristo Tanev and Ralf Steinberger

16:30–17:45 Session IV: Sentiment Analysis and Recommendation

16:30–16:55 *Building English-to-Serbian Machine Translation System for IMDb Movie Reviews*
Pintu Lohar, Maja Popović and Andy Way

16:55–17:15 *Improving Sentiment Classification in Slovak Language*
Samuel Pecar, Marian Simko and Maria Bielikova

17:15–17:35 *Sentiment Analysis for Multilingual Corpora*
Svitlana Galeshchuk, Ju Qiu and Julien Jourdan

17:35–17:45 Closing Remarks