BSNLP 2017

The 6th Workshop on Balto-Slavic Natural Language Processing

Proceedings of the Workshop

EACL 2017 Workshop
April 4, 2017
Valencia, Spain

Endorsed by the Special Interest Group on Slavic Natural Language Processing (SIGSLAV)

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 978-1-945626-45-6

Preface

This volume contains the papers presented at BSNLP-2017: the Sixth Workshop on Balto-Slavic Natural Language Processing. The Workshop is organized by SIGSLAV—Special Interest Group on NLP in Slavic Languages of the Association for Computational Linguistics.

The 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 the growing European consumer markets.

On the other hand, research on theoretical and applied NLP in some of these languages still lags behind the "major" languages, such as English and other West 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 construction of NLP tools, and require rich morphological and syntactic resources. This view is also reflected in Serge Sharoff's invited talk on "Pan-Slavic NLP." In the talk, he discusses an ambitious project on language adaptation—ways to adapt tools and resources among closely related languages, such as those in the Slavic group.

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

As a result of discussions at the previous BSNLP Workshops, to help catalyze collaboration, this year we have organized the first SIGSLAV Challenge: a shared task on multilingual named entity recognition. We have built a dataset, which allows systems to be evaluated on recognizing mentions of named entities in Web documents, their normalization/lemmatization, and cross-lingual matching. The Challenge initially covers seven Slavic languages, and it is intended as a first version of an evaluation standard to be expanded in the future.

We received 24 regular submissions, 14 of which were accepted for presentation.

The papers cover a wide range of topics. Two papers relate to lexical semantics, four to development of linguistic resources, and four to information filtering, information retrieval, and information extraction. Four papers cover topics related to processing of non-standard language or user-generated content. One paper describes the Challenge.

Additionally, 11 teams from 10 countries expressed interest in participating in the Named Entity Challenge, of which two teams have submitted results and system descriptions to date, and whose work is discussed during the session dedicated specifically to the Challenge.

Overall, this workshop's presentations cover at least 10 Balto-Slavic languages: Croatian, Lithuanian, Polish, Russian, Rusyn, Slovene, Serbian (via the regular Workshop papers), and additionally Czech, Slovak and Ukrainian (via the Shared Task Challenge).

This Workshop continues the proud tradition established by the earlier BSNLP Workshops, which were held in conjunction with:

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.

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

BSNLP 2017 Organizers

Organizers:

Tomaž Erjavec, Jožef Stefan Institute, Slovenia
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, University of Copenhagen, Denmark
Tomaž Erjavec, Jozef Stefan Institute, Slovenia
Katja Filippova, Google, Zurich, Switzerland
Darja Fišer, University of Ljubljana, 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
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, Qatar
Maciej Ogrodniczuk, Polish Academy of Sciences, Poland
Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
Maciej Piasecki, Wroclaw University of Technology, Poland
Jakub Piskorski, Joint Research Centre, Ispra, Italy/PAS, Warsaw, Poland
Lidia Pivovarova, University of Helsinki, Finland
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
Serge Sharoff, University of Leeds, UK
Josef Steinberger, University of West Bohemia, Czech Republic
Stan Szpakowicz, University of Ottawa, Canada
Hristo Tanev, Joint Research Centre, Italy
Irina Temnikova, Qatar Computing Research Institute, Qatar
Roman Yangarber, University of Helsinki, Finland
Marcin Woliński, Polish Academy of Sciences, Warsaw, Poland
Daniel Zeman, Charles University, Czech Republic

Invited Speaker:

Serge Sharoff, University of Leeds, UK

Table of Contents

Workshop Program

Tuesday, April 4, 2017

9:00–10:00 Opening Remarks and Invited Talk

9:10–10:00 *Toward Pan-Slavic NLP: Some Experiments with Language Adaptation*
Serge Sharoff

10:10–11:00 Session I: Lexical Semantics

10:10–10:35 *Clustering of Russian Adjective-Noun Constructions using Word Embeddings*
Andrey Kutuzov, Elizaveta Kuzmenko and Lidia Pivovarova

10:35–11:00 *A Preliminary Study of Croatian Lexical Substitution*
Domagoj Alagić and Jan Šnajder

11:00–11:30 Coffee Break

11:30–13:10 Session II: Development of Linguistic Resources

11:30–11:55 *Projecting Multiword Expression Resources on a Polish Treebank*
Agata Savary and Jakub Waszczuk

11:55–12:20 *Lexicon Induction for Spoken Rusyn – Challenges and Results*
Achim Rabus and Yves Scherrer

12:20–12:45 *The Universal Dependencies Treebank for Slovenian*
Kaja Dobrovoljc, Tomaž Erjavec and Simon Krek

12:45–13:10 *Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages*
Tanja Samardžić, Mirjana Starović, Željko Agić and Nikola Ljubešić

Tuesday, April 4, 2017 (continued)

13:10–14:30 Lunch

14:30–16:10 Session III: Processing Non-Standard Language and User-Generated Content

14:30–14:55 *Spelling Correction for Morphologically Rich Language: a Case Study of Russian*
Alexey Sorokin

14:55–15:20 *Debunking Sentiment Lexicons: A Case of Domain-Specific Sentiment Classification for Croatian*
Paula Gombar, Zoran Medić, Domagoj Alagić and Jan Šnajder

15:20–15:45 *Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text*
Nikola Ljubešić, Tomaž Erjavec and Darja Fišer

15:45–16:10 *Comparison of Short-Text Sentiment Analysis Methods for Croatian*
Leon Rotim and Jan Šnajder

16:10–16:30 Coffee Break

16:30–17:20 Session IV: Shared Task on Multilingual Named Entity Recognition

16:30–16:40 *The First Cross-Lingual Challenge on Recognition, Normalization, and Matching of Named Entities in Slavic Languages*
Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger and Roman Yangarber

16:40–16:50 *Liner2 — a Generic Framework for Named Entity Recognition*
Michał Marcińczuk, Jan Kocoń and Marcin Oleksy

16:50–17:00 *Language-Independent Named Entity Analysis Using Parallel Projection and Rule-Based Disambiguation*
James Mayfield, Paul McNamee and Cash Costello