BSNLP 2015

The 5th Workshop on Balto-Slavic Natural Language Processing

Sponsored by SIGSLAV

Proceedings of the Workshop

associated with

The 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015)

10–11 September 2015
Hissar, Bulgaria

The 5th Workshop on
Balto-Slavic Natural Language Processing
associated with THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING 2015

PROCEEDINGS

Hissar, Bulgaria
10–11 September 2015

ISBN 978-954-452-033-5

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

This volume contains the papers presented at BSNLP 2015: the fifth in a series of Workshops on Balto-Slavic Natural Language Processing. This BSNLP Workshop is the first endorsed by SIGSLAV—the newly established ACL Special Interest Group on Natural Language Processing in Slavic Languages.[1]

The driving motivation behind convening the BSNLP Workshops is twofold. On one hand, the languages from the Balto-Slavic group are important for NLP due to their widespread use and diverse cultural heritage. They are spoken by over 400 million speakers worldwide. Due to the recent political and economic developments in Central and Eastern Europe, the countries where Balto-Slavic languages are spoken were brought into new focus in terms of rapid technological advancement and rapidly expanding consumer markets. In the context of the European Union, the Balto-Slavic group today covers about one third of all speakers of EU's official languages.

On the other hand, research on theoretical and applied NLP in many of the Balto-Slavic languages is still in its early stages, although it is continually progressing. The advent of the Internet over twenty years ago established the dominant role of English in a broad range of on-line activities, which further weakened the position of other languages, including the Balto-Slavic group. Consequently, as compared to English, there is still a lack of resources, processing tools and applications for most of these languages, especially ones with smaller speaker bases.

Despite this "minority" status, the Balto-Slavic languages offer a wealth of fascinating scientific and technical challenges for researchers and practitioners to work on. The linguistic phenomena specific to Balto-Slavic languages—such as rich morphological inflection and relatively free word order—present highly intriguing and non-trivial challenges to building NLP tools, and require richer morphological and syntactic resources. Related to this theme, the invited talk by Tanja Samardžic, titled "A computational cross-linguistic approach to Slavic verb aspect" discusses challenges encountered in the computational treatment of the complex phenomena related to verbal aspect in Slavic languages. The talk presents how fine-grained aspectual classes can be automatically extracted using parallel corpora, and then used in temporal classification of events across languages. In the second invited talk, titled "Challenges in launching an NLP start-up company: Research meets the Real World," Josef Steinberger discusses his experience in transferring research results related to Slavic languages into commercial products.

The main goal of the BSNLP 2015 Workshop is to bring together all related stakeholders, including academic researchers and industry practitioners who are involved in work on NLP for Balto-Slavic languages. The Workshop aims to further stimulate research on NLP for these languages and to foster the creation and dissemination of relevant tools and resources. The Workshop serves as an interactive platform for researchers to exchange ideas and experiences, discuss difficult and shared problems, and to facilitate making new resources more widely-known.

This Workshop continues the proud tradition established by the previous BSNLP Workshops:

1. the First BSNLP Workshop, held in conjunction with ACL 2007 Conference in Prague, Czech Republic;

2. the Second BSNLP Workshop, held in conjunction with IIS 2009: Intelligent Information Systems, in Kraków, Poland;

3. the Third BSNLP Workshop, held in conjunction with TSD 2011, 14th International Conference on Text, Speech and Dialogue in Plzeň, Czech Republic;

[1] http://sigslav.cs.helsinki.fi/

4. the Fourth BSNLP Workshop, held in conjunction with ACL 2007 Conference in Sofia, Bulgaria.

This year we received 29 submissions, out of which 16 were accepted for presentation: 13 as regular papers and three as interactive presentations (resulting in an overall acceptance rate of 55%). Compared to previous BSNLP workshops, this year we have a mixed balance of papers on enabling technologies and higher-level tasks, such as information extraction, sentiment analysis and text classification. This shows the ongoing trend towards building user-oriented applications for Balto-Slavic languages, in addition to working on lower-level NLP tools.

The papers directly deal with at least seven Balto-Slavic languages: Bulgarian, Croatian, Czech, Lithuanian, Polish, Russian, and Serbian. Three of the papers discuss approaches to syntactic and semantic analysis. Three papers are about information extraction. Three papers cover sentiment analysis and text classification. Other papers address a broad range of topics, including word-sense disambiguation, corpus analysis, text and author modeling, and linguistic resources.

It is our sincere hope that this work will help to further strengthen the community and stimulate the growth of research in this rich and exciting field.

BSNLP Organizers:
Jakub Piskorski (Polish Academy of Sciences)
Lidia Pivovarova (University of Helsinki)
Jan Šnajder (University of Zagreb)
Hristo Tanev (Joint Research Centre)
Roman Yangarber (University of Helsinki)

Organizers:

Jakub Piskorski, Polish Academy of Sciences, Poland
Lidia Pivovarova, University of Helsinki, Finland
Jan Šnajder, University of Zagreb, Croatia
Hristo Tanev, Joint Research Centre of the European Commission, Ispra, Italy
Roman Yangarber, University of Helsinki, Finland

Program Committee:

Željko Agić, University of Copenhagen, Denmark
Tomaž Erjavec, Jožef Stefan Institute, Slovenia
Darja Fišer, University of Ljubljana, Slovenia
Radovan Garabik, Comenius University in Bratislava, Slovakia
Maxim Gubin, Facebook Inc., Menlo Park CA, USA
Tomas Krilavičius, Vytautas Magnus University, Kaunas, Lithuania
Vladislav Kuboň, Charles University, Prague, Czech Republic
Natalia Loukachevitch, Moscow State University, Russia
Preslav Nakov, Qatar Computing Research Institute, Qatar
Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
Karel Pala, Masaryk University, Brno, Czech Republic
Maciej Piasecki, Wrocław University of Technology, Poland
Jakub Piskorski, Polish Academy of Sciences, Warsaw, Poland
Lidia Pivovarova, University of Helsinki, Finland
Tanja Samardžić, University of Zurich, Switzerland
Agata Savary, University of Tours, France
Kiril Simov, Bulgarian Academy of Sciences, Bulgaria
Inguna Skadina, University of Latvia, Latvia
Jan Šnajder, University of Zagreb, Croatia
Josef Steinberger, University of West Bohemia, Czech Republic
Stan Szpakowicz, University of Ottawa, Canada
Marko Tadić, University of Zagreb, Croatia
Hristo Tanev, Joint Research Centre, Italy
Irina Temnikova, Qatar Computing Research Institute, Qatar
Marcin Woliński, Polish Academy of Sciences, Warsaw, Poland
Roman Yangarber, University of Helsinki, Finland

Invited Speakers:

Tanja Samardžić, University of Zurich, Switzerland
Josef Steinberger, University of West Bohemia, Czech Republic

Table of Contents

Workshop Program

Thursday, September 10, 2015

09:00–09:10 Welcoming Remarks: BSNLP Organizers

09:10–10:00 Invited Talk

A Computational Cross-Linguistic Approach to Slavic Verb Aspect
Tanja Samardžić

10:10–11:00 Session I: Syntax

10:10–10:35 *Universal Dependencies for Croatian (that work for Serbian, too)*
Željko Agić and Nikola Ljubešić

10:35–11:00 *Analytic Morphology – Merging the Paradigmatic and Syntagmatic Perspective in a Treebank*
Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová and Přemysl Vítovec

11:00–11:30 Coffee Break

11:30–12:35 Session II: Information Extraction

11:30–11:50 *Resolving Entity Coreference in Croatian with a Constrained Mention-Pair Model*
Goran Glavaš and Jan Šnajder

11:50–12:15 *Evaluation of Coreference Resolution Tools for Polish from the Information Extraction Perspective*
Adam Kaczmarek and Michał Marcińczuk

12:15–12:35 *Open Relation Extraction for Polish: Preliminary Experiments*
Jakub Piskorski

Thursday, September 10, 2015 (continued)

12:35–14:00 Lunch

14:00–14:50 Invited Talk

Challenges in Launching an NLP Start-up Company: Research Meets the Real World
Josef Steinberger

15:00–15:30 Interactive Session

15:00–15:10 *Regional Linguistic Data Initiative (ReLDI)*
Tanja Samardžić, Nikola Ljubešić and Maja Miličević

15:10–15:20 *Online Extraction of Russian Multiword Expressions*
Mikhail Kopotev, Llorenç Escoter, Daria Kormacheva, Matthew Pierce, Lidia Pivovarova and Roman Yangarber

15:20–15:30 *E-law Module Supporting Lawyers in the Process of Knowledge Discovery from Legal Documents*
Marek Kozłowski, Maciej Kowalski and Maciej Kazula

15:30–16:00 Coffee Break

16:00–17:00 Discussion on BSNLP/SIGSLAV Activities: Shared NLP Task

Friday, September 11, 2015

09:00–09:45 Session III: Semantics

09:00–09:25 *Experiments on Active Learning for Croatian Word Sense Disambiguation*
Domagoj Alagić and Jan Šnajder

09:25–09:45 *Automatic Classification of WordNet Morphosemantic Relations*
Svetlozara Leseva, Ivelina Stoyanova, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov and Svetla Koeva

09:55–11:00 Session IV: Corpus Analysis and Resources

09:55–10:20 *Applying Multi-Dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres*
Anisya Katinskaya and Serge Sharoff

10:20–10:40 *Distinctive Similarity of Clausal Coordinate Ellipsis in Russian Compared to Dutch, Estonian, German, and Hungarian*
Karin Harbusch and Denis Krusko

10:40–11:00 *Universalizing BulTreeBank: a Linguistic Tale about Glocalization*
Petya Osenova and Kiril Simov

11:00–11:30 Coffee Break

Friday, September 11, 2015 (continued)

11:30–12:35 Session V: Sentiment Analysis and Text Classification

11:30–11:50 *Types of Aspect Terms in Aspect-Oriented Sentiment Labeling*
Natalia Loukachevitch, Evgeniy Kotelnikov and Pavel Blinov

11:50–12:15 *Authorship Attribution and Author Profiling of Lithuanian Literary Texts*
Jurgita Kapočiūtė-Dzikienė, Andrius Utka and Ligita Šarkutė

12:15–12:35 *Classification of Short Legal Lithuanian Texts*
Vytautas Mickevičius, Tomas Krilavičius and Vaidas Morkevičius

12:35–12:40 Closing Remarks

12:40–14:00 Lunch