ACL 2013

The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing

Workshop Proceedings

August 8-9, 2013
Sofia, Bulgaria

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

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 978-1-937284-59-6

Preface

This volume contains the papers presented at BSNLP-2013: the Fourth in a series of Workshops on Balto-Slavic Natural Language Processing.

The motivation for convening the Workshops is clear. On one hand, the languages from the Balto-Slavic group play an important role due to their widespread use and diverse cultural heritage. The languages are spoken by over 400 million speakers worldwide. The recent political and economic developments in Central and Eastern Europe bring Balto-Slavic societies and their languages into new focus in terms of rapid technological advancement and expanding consumer markets. In the context of the European Union, the Balto-Slavic group today constitutes about one third of its 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. The advent of the Internet in the 1990's established the dominant role of English in science, popular culture, and other areas of on-line activity, which further weakened the presence of Balto-Slavic languages. Consequently, in comparison to English, there is a dire lack of resources, processing tools and applications for most of these languages, especially the smaller ones.

Despite this "minority" status, the Balto-Slavic languages offer a wealth of fascinating scientific and technical challenges for researchers 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 problems for construction of NLP tools for these languages, and require richer morphological and syntactic resources to be exploited. In this direction, the invited talk by Kiril Simov on *"Ontologies and Linked Open Data for Acquisition and Exploitation of Language Resources"* presents methods for acquisition of language resources from different types of on-line and off-line data sources.

The goal of this Workshop was to bring together academic researchers and industry practitioners who work on NLP for Balto-Slavic languages. It is our hope that the Workshop would further stimulate research on NLP for these languages and foster the creation of tools for them. The Workshop gives the researchers a forum for exchange of ideas and experience, for discussion difficult-to-tackle problems, and for making new resources more widely-known. One fascinating aspect of this sub-family of languages is their structural similarity, as well as an easily recognisable 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, and communicate in a natural way.

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

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

This year we received 31 submissions—a 50% increase over the First BSNLP Workshop in 2007. Of these, 16 were accepted for presentation (resulting in an acceptance rate of 51%). Compared to the previous BSNLP workshops, this year we have more papers about higher-level tasks, such as information extraction and sentiment analysis. This hopefully shows a trend towards building user-oriented applications for Balto-Slavic languages, in addition to working on lower-level NLP tools.

Three papers discuss approaches to sentiment analysis and opinion mining. Five are about classic information extraction tasks: three on named entity recognition and two on event extraction. Two papers are about WordNets for different Slavic languages. Two papers are about morphological processing and parsing. The rest of the papers cover different topics, including acquisition of resources, keyword extraction, and lexicon analysis.

The papers together cover nine different languages: 5 on Croatian, 4 on Russian, 2 each on Bulgarian, Polish and Slovene, and one each on Czech, Lithuanian and Serbian. We also accepted an interesting paper about named-entity recognition for Estonian—which, although it does not belong to the Balto-Slavic group, does belong to the Baltic area, and has morphological complexity at least matching that of the Baltic and Slavic languages.

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

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

Organizers:

Jakub Piskorski, Polish Academy of Sciences
Lidia Pivovarova, University of Helsinki, and St.Petersburg State University, Russia
Hristo Tanev, Joint Research Centre of the European Commission, Ispra, Italy
Roman Yangarber, University of Helsinki, Finland

Program Committee:

Tania Avgustinova, University of Saarland, Germany
Kalina Bontcheva, University of Sheffield, UK
Pavel Braslavski, Kontur labs, and Ural Federal University, Russia
Boris Dobrov, Moscow State University, Russia
Tomaž Erjavec, Jozef Stefan Institute, Ljubljana, Slovenia
Katja Filippova, Google, Zurich, Switzerland
Radovan Garabik, Comenius University in Bratislava
Tatiana Gornostay, Tilde, Riga, Latvia
Maxim Gubin, Facebook Inc., Menlo Park CA
Mikhail Kopotev, University of Helsinki, Finalnd
Vladislav Kuboň, Charles University, Prague, Czech Republic
Olga Mitrofanova, St.Petersburg State University, Russia
Karel Pala, Masaryk University, Brno, Czech Republic
Maciej Piasecki, Wrocław University of Technology, Poland
Jakub Piskorski, Polish Academy of Sciences, Warsaw, Poland
Lidia Pivovarova, Univeristy of Helsinki/St.Petersburg State University, Russia
Adam Przepiórkowski, Polish Academy of Sciences, Warsaw, Poland
Agata Savary, Université François Rabelais, Tours, France
Kiril Simov, Bulgarian Academy of Sciences, Bulgaria
Joseph Steinberger, University of West Bohemia, Plzeň, Czech Republic
Pavel Straňák, Charles University in Prague, Czech Republic
Stan Szpakowicz, University of Ottawa, Canada
Marko Tadic, University of Zagreb, Croatia
Hristo Tanev, Joint Research Centre, Ispra, Italy
Roman Yangarber, University of Helsinki, Finalnd

Invited Speaker:

Kiril Simov, Bulgarian Academy of Sciences, Bulgaria

Table of Contents

Workshop Program

Thursday, August 8, 2013

9:00–9:15	Welcome Remarks
9:15–10:30	Invited Talk: *Ontologies and Linked Open Data for Acquisition and Exploitation of Language Resources* Kiril Simov
10:30–11:00	Coffee Break
	Session I: Opinion Mining and Sentiment Analysis
11:00–11:25	*A Comparison of Approaches for Sentiment Classification on Lithuanian Internet Comments* Jurgita Kapočiutė-Dzikienė, Algis Krupavičius and Tomas Krilavičius
11:25–11:45	*Evaluating Sentiment Analysis Systems in Russian* Ilia Chetviorkin and Natalia Loukachevitch
11:45–12:05	*Aspect-Oriented Opinion Mining from User Reviews in Croatian* Goran Glavaš, Damir Korenčić and Jan Šnajder
	Session II: Morphology, Syntax and Semantics
12:05–12:30	*Frequently Asked Questions Retrieval for Croatian Based on Semantic Textual Similarity* Mladen Karan, Lovro Žmak and Jan Šnajder
12:30–14:00	Lunch
14:00–14:25	*Parsing Russian: a hybrid approach* Dan Skatov, Sergey Liverko, Vladimir Okatiev and Dmitry Strebkov
14:25–14:45	*GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts* Marko Bekavac and Jan Šnajder
14:45–15:10	*Lemmatization and Morphosyntactic Tagging of Croatian and Serbian* Željko Agić, Nikola Ljubešić and Danijela Merkler

Thursday, August 8, 2013 (continued)

Session III: Cross-lingual methods and Machine Translation

15:10–15:30 *Modernizing historical Slovene words with character-based SMT*
Yves Scherrer and Tomaž Erjavec

15:30–16:00 Coffee Break

16:00–16:25 *Improving English-Russian sentence alignment through POS tagging and Damerau-Levenshtein distance*
Andrey Kutuzov

16:25–16:50 *Identifying false friends between closely related languages*
Nikola Ljubešić and Darja Fišer

16:50–17:15 Discussion: Establishing BSNLP SIG

Friday, August 9, 2013

Session IV: Information Extraction

09:20–09:45 *Named Entity Recognition in Estonian*
Alexander Tkachenko, Timo Petmanson and Sven Laur

09:45–10:10 *On Named Entity Recognition in Targeted Twitter Streams in Polish.*
Jakub Piskorski and Maud Ehrmann

10:10–10:30 *Recognition of Named Entities Boundaries in Polish Texts*
Michał Marcińczuk and Jan Kocoń

10:30–11:00 Coffee Break

11:00–11:25 *Adapting the PULS event extraction framework to analyze Russian text*
Lidia Pivovarova, Mian Du and Roman Yangarber

11:25–11:50 *Semi-automatic Acquisition of Lexical Resources and Grammars for Event Extraction in Bulgarian and Czech*
Hristo Tanev and Josef Steinberger

Friday, August 9, 2013 (continued)

11:50–12:15	*Wordnet-Based Cross-Language Identification of Semantic Relations* Ivelina Stoyanova, Svetla Koeva and Svetlozara Leseva
12:15–12:30	Concluding remarks