ACL 2007

Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007 Special Theme: Information Extraction and Enabling Technologies

June 29, 2007
Prague, Czech Republic

The Association for Computational Linguistics

Production and Manufacturing by
Omnipress
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`

Preface

There are over 400 million speakers of **Balto-Slavonic languages** world-wide (synonymously used: Balto-Slavic). As of 2007, almost a third of the 23 official European Union languages are Balto-Slavonic, i.e. Bulgarian, Czech, Latvian, Lithuanian, Polish, Slovak and Slovene. The two most recent rounds of the EU Enlargement fundamentally raised the interest in these languages: translators and interpreters for new language pairs need to be found, the interest in Machine (Aided) Translation systems has risen and tools that help language specialists and information-seeking individuals are now highly sought after.

For some of the countries speaking Balto-Slavonic languages, there is a **rich linguistic heritage**, and computational linguistics research and development is rather advanced. For others, however, there has been little development. This is due to the often small number of speakers of that language (Latvian, for instance, is spoken natively by around 1.5 Million people) combined with a lack of access to basic resources needed for Natural Language Processing such as machine-readable corpora and dictionaries, morphological analysers, part-of-speech taggers, parsers, etc. This leads to a linguistic **brain drain** as some of the best computational linguistics students go abroad or – even when staying in their mother country – work on developing new systems for English, French or German because resources for these languages are readily available.

Even when linguistic resources and tools are available to the scientific community, methods that have been successfully applied to Germanic and Romance languages cannot simply be ported to languages from the Balto-Slavonic group. The most well-known **linguistic phenomena** making Balto-Slavonic text analysis harder are the highly inflectional character and the related phenomenon of free word order in these languages. The invited speaker at the workshop, Adam Przepiórkowski from the Polish Academy of Sciences, explains these and a number of further specific linguistic phenomena typical for this language group. Interestingly, he points out that the differences may not always make text analysis tasks harder, but that they can also make some tasks easier.

When proposing this workshop to the Association for Computational Linguistics, we knew that Language Technology for some of the languages is not very advanced and that, to date, not much work has been carried out in the area of Information Extraction. The **objective of this workshop** was thus to promote the work on Balto-Slavonic languages by helping scientists to describe and share their resources and to describe their efforts, hoping that the experiences of a few will be useful for many others. We did not expect to receive papers presenting highly novel approaches to Information Extraction, but rather an adaptation of known methods to new languages and solutions for specific challenges regarding the Balto-Slavonic group. We found, however, that the current work on applying known approaches to a different language type did produce some very interesting work.

Contents of these proceedings

We received in total 20 submissions, of which only eight were specifically about **Information Extraction**, i.e. named entity recognition or the extraction of definitions. One of these papers targeted a more high-level Information Extraction task: an application combining more than one entity to fill a scenario template. The other 12 papers fall under the category Enabling Technology, describing mostly morphological tools and resources, taggers, corpora, WordNet developments and topic segmentation of texts.

Each of the submissions was reviewed by at least three peers. Finally, we have accepted 9 papers for oral presentations (**acceptance rate of 45%**) and selected a further three for poster presentations. We want to use this opportunity to thank the Programme Committee for their thorough reviews and mostly extensive and useful comments, as well as for keeping to the deadlines.

We do not want to claim that the proceedings of this workshop exhaustively reflect existing work on Information Extraction for the targeted language group. The submissions we received nevertheless gave us an overview of the **current state-of-the-art** for the various languages. Most of the submitted papers covered work on the languages Polish, Czech and Bulgarian. Only individual papers concerned Lithuanian, Croatian, Serbian, Ukrainian and Russian. When selecting papers for acceptance, an important criterion for us was to cover various languages.

The **final workshop program** includes 6 papers specifically addressing Information Extraction while the other six discuss Enabling Technologies. All of the papers clearly show how Natural Language Processing differs for Balto-Slavonic languages, compared to Germanic, Romance or other languages.

Our appeal

During the organisation of this workshop, we saw that some very good text analysis work has been carried out for some of the Balto-Slavonic languages. However, we were also reminded that many good scientists cannot work on interesting and promising research subjects and high-level applications because they first need to create the necessary linguistic resources. We therefore want to appeal to all researchers in the community to make as many resources and tools available to their peers as they can. This will in the end benefit all the scientists and – in the long run – the country as a whole. In this spirit, the Joint Research Centre has compiled – and distributes for free – a paragraph-aligned and subject domain-classified parallel corpus in 22 languages (the *JRC-Acquis*). May this resource be useful for the communities working on less-widely spoken languages.

Ispra, Italy, May 2007

Jakub Piskorski
Hristo Tanev
Bruno Pouliquen
Ralf Steinberger

European Commission – Joint Research Centre

Organizers

Chairs:

Jakub Piskorski, Joint Research Centre, IPSC
Bruno Pouliquen, Joint Research Centre, IPSC
Ralf Steinberger, Joint Research Centre, IPSC
Hristo Tanev, Joint Research Centre, IPSC

Program Committee:

Kalina Bontcheva, University of Sheffield
Tomaž Erjavec, Jožef Stefan Institute
Vladislav Kuboň, Charles Univeristy Prague
Anna Kupść, Université Paris 3
Rūta Marcinkevičienė, Vytautas Magnus University, Kaunas
Agnieszka Mykowiecka, Polish Academy of Sciences
Jakub Piskorski, Joint Research Centre, IPSC
Bruno Pouliquen, Joint Research Centre, IPSC
Hristo Tanev, Joint Research Centre, IPSC
Marko Tadić, University of Zagreb
Agata Savary, University of Tours
Kiril Simov, Bulgarian Academy of Sciences
Wojciech Skut, Google Inc.
Ralf Steinberger, Joint Research Centre, IPSC
Duško Vitas, University of Beograd
Roman Yangarber, Univeristy of Helsinki

Program Committee Co-Chairs:

Jakub Piskorski, Joint Research Centre, IPSC
Hristo Tanev, Joint Research Centre, IPSC

Additional Reviewers:

Niraj Aswani
Jan Daciuk
Camelia Ignat
Mladen Kolar
Domen Marinčič
Diana Maynard

Invited Speaker:

Adam Przepiórkowski, Polish Academy of Sciences

Table of Contents

Conference Program

Friday, June 29, 2005

9:00–9:20 Opening Remarks

Invited Talk:

9:20–10:20 *Slavic Information Extraction and Partial Parsing*
Adam Przepiórkowski

Session 1: Information Extraction

10:20–10:45 *Implementation of Croatian NERC System*
Božo Bekavac and Marko Tadić

10:45–11:15 Morning Cofee Break

11:15–11:40 *A Language Independent Approach for Name Categorization and Discrimination*
Zornitsa Kozareva, Sonia Vázquez and Andrés Montoyo

11:40–12:05 *Lemmatization of Polish Person Names*
Jakub Piskorski, Marcin Sydow and Anna Kupść

12:05–12:30 *Automatic Processing of Diabetic Patients' Hospital Documentation*
Małgorzata Marciniak and Agnieszka Mykowiecka

12:30–14:30 Lunch

Friday, June 29, 2005 (continued)

Session 2: Information Extraction and Enabling Technologies

Session 3: Enabling Technologies

Session 4: Poster Session