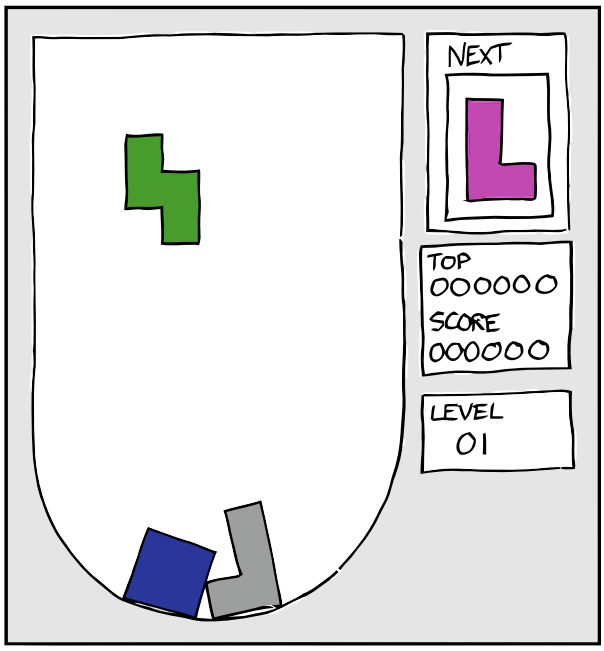
IWPT 2011

The Second Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2011)



Proceedings of SPMRL 2011

October 6, 2011
Dublin, Ireland

- Endorsed by SIGPARSE, the ACL Special Interest Group on Natural Language Parsing.
- Sponsored by the INRIA'S ALPAGE PROJECT.
- With the kind support of IWPT 2011's conference chairs, Dublin City University and Paris-Sorbonne University.

ISBN: 978-1-932432-73-2

Order copies of previous SPMRL 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

Forewords

Welcome to the second workshop on Statistical Parsing of Morphologically Rich Languages!
Following the warm reception of the first official SPMRL workshop at NAACL-HLT 2010, our aim with the second workshop is to build upon the success of the first and offer a platform to the growing community of people who are interested in developing tools and resources for parsing MRLs. We decided to collocate with the International Workshop on Parsing Technologies (IWPT), both because the themes of the two events are so closely related and because the seeds of the SPMRL workshop were planted during IWPT 2009 in Paris. The warm welcome and support of the IWPT community made it our unequivocal choice, and we are honored and pleased to collocate our second SPMRL workshop with this year's IWPT event

Fourteen papers were submitted in total to the workshop. After two withdrawals,we chose to accept four long papers and four short papers, giving an acceptance rate of 66%. Our goal during the selection process was to produce a varied, balanced and interesting program without compromising on quality, and we believe that we have achieved this goal. This year's papers cover a broad range of languages (Arabic, Basque, French, German, Hindi, Korean,Turkish) and are concerned with the most pressing issues (handling discontinuity, incorporating morphological information, the problems of real-world text) over a range of parsing approaches (discriminative and generative, constituency and dependency) We believe that they will result in a lively and productive workshop.

We are continuing the SPMRL tradition of ending the workshop with a panel discussion. We are very proud of this year's panel and are very grateful to our panellists (Josef van Genabith , James Henderson, Joakim Nivre, Slav Petrov and Yannick Versley) for agreeing to participate. This year's main theme for the panel will be the design of a shared task on parsing MRLs, in the face of various challenges that emerge when going beyond English and the WSJ Penn Treebank. A typical challenge when moving away from parsing English to other languages, is, for instance, the nature of the input, which consists of unanalyzed non-gold word tokens. Additional challenges have to do with the typological diversity of the syntactic structures and sound evaluation across experiments using different corpora. Our ultimate goal in this panel will be to tease apart the various issues we need to address and understand how we might organize a shared task that will bring our burgeoning field forward.

Finally, we would like to express our gratitude to the following people who helped us organise SPMRL 2011: the IWPT chairs, Joakim Nivre and his limitless availibility, Özlem Çetinoğlu and Harry Bunt, Alon Lavie and Kenji Sagae from SIGPARSE who continue to support this workshop, Josef van Genabith, Laurence Danlos and, last but not least, our very knowledgeable and hard-working review committee who did a sterling job during the off-peak review season. Thanks guys!

The SPMRL 2011 Program Committee

Program Chairs:

Djamé Seddah, INRIA/University of Paris-Sorbonne (France)
Reut Tsarfaty, Uppsala University (Sweden)
Jennifer Foster, NCLT, Dublin City University (Ireland)

Program Committee:

Marie Candito, INRIA/University Paris 7 (France)
Yoav Goldberg, Ben Gurion University of the Negev (Israel)
Ines Rehbein, Universität Saarbrücken (Germany)
Lamia Tounsi, NCLT, Dublin City University (Ireland)
Yannick Versley, Universität Tübingen (Germany)

Review Committee:

Mohammed Attia (Dublin City University, Ireland)
Bernd Bohnet (University of Stuttgart, Germany)
Adriane Boyd (Ohio State University, USA)
Özlem Çetinoğlu (Dublin City University, Ireland)
Grzegorz Chrupała (Saarland University, Germany)
Benoit Crabbé (University of Paris 7, France)
Josef van Genabith (Dublin City University, Ireland)
Spence Green (Stanford University, US)
Deirdre Hogan (Dublin City University, Ireland)
Samar Husain (Inter. Institute of Information Technology, India)
Sandra Kuebler (Indiana University, US)
Jonas Kuhn (University of Stuttgart, Germany)
Alberto Lavelli (FBK-irst, Italy)
Joseph Le Roux (Dublin City University, Ireland)
Wolfgang Maier (University of Tüebingen, Germany)
Yuval Marton (IBM Watson Resarch Center, US)
Takuya Matsuzaki (University of Toyko, Japan)
Yusuke Miyao (University of Toyko, Japan)
Joakim Nivre (Uppsala University, Sweden)
Ines Rehbein (Saarland University, Germany)
Owen Rambow (Columbia University, US)
Kenji Sagae (University of Southern California, USA)
Benoit Sagot (Inria Rocquencourt, France)

Panelists:

Josef van Genabith (Dublin City University, Ireland)
James Henderson (Université de Geneve, Switzerland)
Joakim Nivre (Uppsala University, Sweden)
Slav Petrov (Google Research NY, USA)
Reut Tsarfaty (Chair, Uppsala University, Sweden)
Yannick Versley (University of Tübingen, Germany)

Table of Contents

Workshop Program

Thursday October 6, 2011

13:30–13:35 **Opening Remarks**

13:35-14:45 **Session 1** (Chair: Djamé Seddah)

Statistical Dependency Parsing in Korean: From Corpus Generation To Automatic Parsing
Jinho D. Choi and Martha Palmer

Morphological Features for Parsing Morphologically-rich Languages: A Case of Arabic
Jon Dehdari, Lamia Tounsi and Josef van Genabith

French parsing enhanced with a word clustering method based on a syntactic lexicon
Anthony Sigogne, Matthieu Constant and Eric Laporte

Testing the Effect of Morphological Disambiguation in Dependency Parsing of Basque
Kepa Bengoetxea, Arantza Casillas and Koldo Gojenola

14:45–15:05 **Break**

15:05-16:15 **Session 2** (Chair: Jennifer Foster)

Discontinuous Data-Oriented Parsing: A mildly context-sensitive all-fragments grammar
Andreas van Cranenburgh, Remko Scha and Federico Sangati

Multiword Expressions in Statistical Dependency Parsing
Gulsen Erygit, Tugay Ilbay and Ozan Arkan Can

Linguistically Rich Graph Based Data Driven Parsing For Hindi
Samar Husain, Raghu Pujitha Gade and Rajeev Sangal

Data point selection for self-training
Ines Rehbein

Thursday October 6, 2011 (continued)

16:15-16:25 **Short Break**

16:25-17:25 **Discussion Panel: Josef van Genabith, James Hendersen, Joakim Nivre, Slav Petrov, Yannick Versley** (Chair: Reut Tsarfaty)

17:25-17:30 **Concluding Remarks**