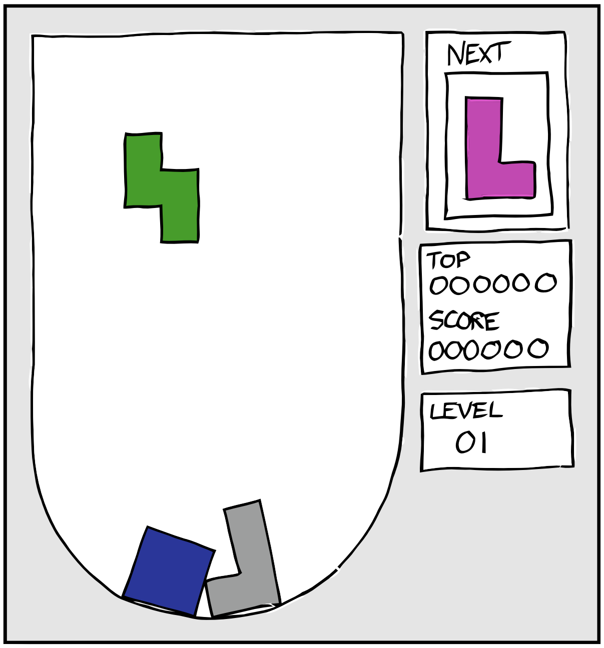
NAACL HLT 2010

The First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)



Proceedings of the Workshop

June 5, 2010
Los Angeles, California

USB memory sticks produced by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

- Endorsed by SIGPARSE, the ACL Special Interest Group on Natural Language Parsing.
- Sponsored by the INRIA's ALPAGE PROJECT.

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

Foreword

The idea of organizing this workshop was sparked following very interesting discussions that occurred during EACL09 among various researchers working on statistical parsing of different types of languages. Indeed, an opportunity to discuss the issues that we were all experiencing was much needed, and it seemed such a good idea that we decided to take advantage of IWPT'09, which was held that year in Paris, to organize a panel on this topic. We planned to have presentations on the various issues faced by this small emerging community, which would allow us to share our sometimes similar solutions for parsing different languages.

Inspired by the idea of organizing such a meeting, but without knowing quite yet if there was any sense in comparing, for example, Modern Hebrew and French parsing issues, Deirdre Hogan (Dublin City University) suggested that - should the panel be successful - we ought to organize a real workshop. She was right. We had an extremely successful and animated panel discussion. We were surprised to see the extent to which the IWPT'09 audience chose to contribute to the discussion instead of taking a break from the long presentation sessions. This encouraged us to pursue these attempts at providing a forum for discussing such matters even further, and to create a new community of shared interests. This workshop is the result of our common will to do so.

We believe that the issues faced by researchers involved in statistical parsing of morphologically rich languages are not always well known outside of this small community, and that the kind of challenges that we all face require a more thorough introduction than we could possibly provide in this foreword. Therefore, we decided to include here an elaborated preface which presents the current state-of-affairs with respect to parsing MRLs and frames the various contributions to our workshop in relation to it. The overview should act as a primer for those who are not experienced in the subject and yet wish to participate in the discussion. All in all, we are proud to have 11 very nice papers presented in our proceedings that will help advance the state of the art in parsing MRLs. In order to obtain sufficient presentation slots, we asked our authors to choose between different modes of presentation, we are glad the authors involved in 3 papers accepted to present them as posters.

Finally, we would like to express our gratitude to the many people who encouraged us on this journey: Harry Bunt, Alon Lavie and Kenji Sagae from SIGPARSE which fully endorses this project; Joakim Nivre who heartily encouraged us to launch our workshop, Eric de la Clergerie who agreed to give us a slot at IWPT'09 and Josef van Genabith who very kindly chaired our first panel, all of whom constantly advised us during this year - this was precious to us. More than 20 very busy researchers agreed to review for our workshop - without their commitment this would have been plainly impossible. We further wish to thank Kevin Knight who kindly agreed to give a talk on a pressing topic, morphology in SMT, in this workshop, and Dan Bikel, Julia Hockenmaier, Slav Petrov and Owen Rambow, who willingly agreed to engage in our panel discussion. Last but not least, we want to thank Laurence Danlos - whose team, the Alpage project, is funding our workshop - for believing in our project from the start.

Best regards,

The SPRML2010 extended Program Committee

Organizers:

Djamé Seddah, INRIA/University of Paris-Sorbonne (France)
Sandra Kübler, Indiana University (USA)
Reut Tsarfaty, Uppsala University (Sweden)

Program Committee:

Marie Candito, INRIA/University Paris 7 (France)
Jennifer Foster, NCLT, Dublin City University (Ireland)
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:

Mohamed Attia (Dublin City University, Ireland)
Adriane Boyd (Ohio State University, USA)
Aoife Cahill (University of Stuttgart, Germany)
Grzegorz Chrupała (Saarland University, Germany)
Benoit Crabbé (University of Paris 7, France)
Michael Elhadad (Ben Gurion University, Israel)
Emar Mohamed (Indiana University, USA)
Josef van Genabith (Dublin City University, Ireland)
Julia Hockenmaier (University of Illinois, USA)
Deirdre Hogan (Dublin City University, Ireland)
Alberto Lavelli (FBK-irst, Italy)
Joseph Le Roux (Dublin City University, Ireland)
Wolfgang Maier (University of Tüebingen, Germany)
Takuya Matsuzaki (University of Toyko, Japan)
Yusuke Miyao (University of Toyko, Japan)
Joakim Nivre (Uppsala University, Sweden)
Ines Rehbein (Saarland University, Germany)
Kenji Sagae (University of Southern California, USA)
Benoit Sagot (Inria Rocquencourt, France)
Khalil Sima'an (University of Amsterdam, The Netherlands)
Nicolas Stroppa (Google Research Zurich, Switzerland)

Invited Speaker:

Kevin Knight, University of Southern California/Information Sciences Institute

Panelists:

Dan Bikel, Google Research NY (USA)
Julia Hockenmaier, University of Illinois at Urbana-Champaign (USA)
Slav Petrov, Google Research NY (USA)
Owen Rambow, Columbia University (USA)

Table of Contents

Workshop Program

Saturday, June 5, 2010

09:00-09:15 **Opening Remarks**

Statistical Parsing of MRLs: Workshop Preface and Overview

Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither
Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein and Lamia Tounsi

09:15-10:30 **Dependency-based parsing of MRLs** (Chair: Reut Tsarfaty)

Improving Arabic Dependency Parsing with Lexical and Inflectional Morphological Features
Yuval Marton, Nizar Habash and Owen Rambow

Two Methods to Incorporate 'Local Morphosyntactic' Features in Hindi Dependency Parsing
Bharat Ram Ambati, Samar Husain, Sambhav Jain, Dipti Misra Sharma and Rajeev Sangal

Application of Different Techniques to Dependency Parsing of Basque
Kepa Bengoetxea and Koldo Gojenola

10:30-11:00 **Break**

11:00-12:15 **Constituency-based parsing of MRLs** (Chair: Djamé Seddah)

Modeling Morphosyntactic Agreement in Constituency-Based Parsing of Modern Hebrew
Reut Tsarfaty and Khalil Sima'an

Factors Affecting the Accuracy of Korean Parsing
Tagyoung Chung, Matt Post and Daniel Gildea

Direct Parsing of Discontinuous Constituents in German
Wolfgang Maier

12:15-1:40 **Lunch Break**

Saturday, June 5, 2010 (continued)

1:40-2:30 **Invited Talk** (Chair: Reut Tsarfaty)

Morphology in Statistical Machine Translation: Integrate-in or Tack-on?
Kevin Knight

2:30-3:00 **Improved Estimation for parsing MRLs** (Chair: Yoav Goldberg)

Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French
Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi and Josef van Genabith

Parsing Word Clusters
Marie Candito and Djamé Seddah

3:00-3:30 **Break**

3:30-4:45 **Rich Morphology and Lemmatisation: Short Papers and Posters** (Chair: Jennifer Foster)

Lemmatization and Lexicalized Statistical Parsing of Morphologically-Rich Languages: the Case of French
Djamé Seddah, Grzegorz Chrupała, Ozlem Cetinoglu, Josef van Genabith and Marie Candito

On the Role of Morphosyntactic Features in Hindi Dependency Parsing
Bharat Ram Ambati, Samar Husain, Joakim Nivre and Rajeev Sangal

Easy-First Dependency Parsing of Modern Hebrew
Yoav Goldberg and Michael Elhadad

4:45-5:45 **Discussion Panel: Dan Bikel, Julia Hockenmaier, Slav Petrov and Owen Rambow** (Chair: Sandra Kübler)

5:45-6:00 **Concluding remarks**