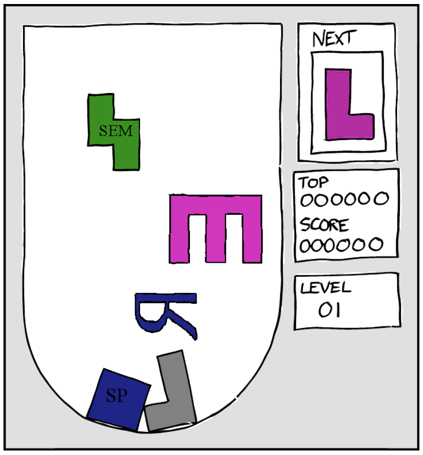
ACL 2012

50th Annual Meeting of the Association for Computational Linguistics



Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages

July 12, 2012
Jeju, Republic of Korea

- Endorsed by SIGPARSE, the ACL Special Interest Group on Natural Language Parsing,
- and SIGLEX, the ACL Special Interest Group on the Lexicon.
- Sponsored by the Pascal 2 Network, Network of Excellence funded by the European Union.

ISBN: 978-1-937284-30-5

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 to SP-SEM-MRL 2012

Morphologically Rich Languages (MRLs) are languages in which grammatical relations such as Subject, Predicate, and Object, are largely indicated morphologically (e.g., through inflection) instead of positionally. This poses serious challenges for current (English-centric) syntactic and semantic processing. Furthermore, since grammatical relations provide the interface to compositional semantics, morpho-syntactic phenomena may significantly complicate processing the syntax–semantics interface. In statistical parsing, English parsing performance has reached a high plateau in certain genres. Semantic processing of English has similarly seen much progress in recent years. MRL processing presents new challenges, such as optimal morphological representation, non position-centric algorithms, or different semantic distance measures.

These challenges lurk in areas where parses may be used as input, such as semantic role labeling, distributional semantics, paraphrasing and textual entailment; inadequate representation or pre-processing of morphological variation is likely to hurt parsing and semantic tasks alike.

This joint workshop aims to build upon the first and second SPMRL workshops (at NAACL-HLT 2010 and IWPT 2011, respectively) while extending the overall scope to include semantic processing. We aim to encourage cross-fertilization among researchers working on different languages and among those working on different levels of processing.

The syntax track received 11 papers of which 7 were accepted for publication. This year's collection of papers describe work on Korean, Basque, French, Spanish, Portuguese and Tamil (the latter three are a first for SPMRL), and encompass several different parsing approaches and combinations thereof, including dependency parsing, PCFG-LA parsing, rule-based parsing and precision-grammar-based parsing.

A trend of this year's papers is the problem of data sparsity in statistical parsing of MRLs: Candito et al. present a technique that involves the use of word clusters, lemmas and Wordnet synsets to alleviate the problem of OOV words in statistical parsing with the French Treebank; Silva and Branca investigate whether dependency information can be used to assign lexical types to OOV words in a HPSG precision grammar approach to Portuguese parsing; Le Roux et al. investigate the problem of data sparsity in the context of Spanish constituency parsing and show that optimising the processes of lemmatisation and part-of-speech tagging can lead to improved parsing performance; Green et al. tackle the problem of small training sets by applying ensemble parsing models trained on subsets of the entire training set. (They test their approach on the Tamil language but suggest that it is applicable to any language with minimal treebank resources).

We are also happy to present parsing papers that describe general parsing techniques that are applicable to any language, but which have been tested on MRLs: Goenaga et al. explore an approach which involves the combination of rule-based and data-driven parsing, and test this combined approach on the Basque language; Le Roux et al. present a reranking technique in which the n-best trees produced by a constituency parser are then converted to dependency trees and reranked using dependency information. (The approach is tested on a language with scant morphology, English, and a language with a richer inflectional system, French); Finally, Choi et al. present work which aims to reduce ambiguity in statistical parsing of Korean by transforming eojeol-based trees into entity-based trees. Their work is relevant to all languages where the word is not the natural unit of syntactic analysis.

Five papers, of seven submissions, were accepted for the Semantic Track of SP-SEM-MRL 2012. The

selected papers reflect a healthy diversity of semantic models and the fertile breadth of applications for semantics in morphologically rich languages: Versley applies supervised learning to the task of classifying German noun-verb semantic relations. The experiments evaluate a wide range of corpus- and lexicon-based features for representing the noun-verb pairs; Lorenzo and Cerisara present a Bayesian model for unsupervised Semantic Role Labeling for English and French, with promising results; Hawwari, Bar, and Diab propose a method for creating a resource of Arabic multi-word expressions. The method handles MWEs with gaps, which can be problematic for Arabic; Versley and Henrich describe an approach to word sense discrimination based on the hypothesis that an ambiguous word is unambiguous when embedded in the context of a compound word. Their findings support the utility of the hypothesis.
In research which combines both syntactic and semantic processing, Acedański, Slaski, and Przepiórkowski introduce a procedure for extracting dependency information from chunked data. Given the output of a chunker without prepositional phrase attachment information, their procedure is able to make attachment decisions using lexical, morphosyntactic, lexico-semantic, and association features.

It is our hope that the rich programme of SP-Sem-MRL 2012 will foster interactions and collaborations between the syntax and the semantics community on the topic of Morphologically Rich Languages processing. Our aim is to help to bring ideas (and solutions) to the fore and promote a more rapid advance of the state-of-the-art in the field.

We thank our authors and the Program Committee for making SP-Sem-MRL 2012 a success.

Marianna Apidianaki, Ido Daga, Katryn Erk, Jennifer Foster, Yuval Marton, Ines Rehbein, Djamé Seddah, Reut Tsarfaty and Peter Turney

General Chairs:

Marianna Apidianaki (LIMSI-CNRS, France)
Ido Dagan (Bar-Ilan University, Israel)
Jennifer Foster (Dublin City University, Ireland)
Yuval Marton (IBM Watson Research Center, USA)
Djamé Seddah (University of Paris Sorbonne, France)
Reut Tsarfaty (Uppsala University, Sweden)

Shared Session Chairs:

Katrin Erk (University of Texas at Austin, USA)
Ines Rehbein (University of Potsdam, Germany)
Peter Turney (National Research Council, Canada)
Yannick Versley (University of Tuebingen, Germany)

Invited Speakers:

Mark Steedman (University of Edinburgh, UK)
Ivan Titov (Saarland University, Germant)

Program Committee:

Ion Androutsopoulos (Athens Univ. of Economics and Business, Greece)
Mohammed Attia (Dublin City University, Ireland)
Adriane Boyd (Ohio State University, US)
Bernd Bohnet (University of Stuttgart, Germany)
Marie Candito (University of Paris 7, France)
Aoife Cahill (Educational Testing Service, US)
Gülşen Cebiroğlu Eryiğit (Istambul Technical University, Turkey)
Ozlem Cetinoglu (University of Stuttgart, Germany)
Jinho Choi (University of Colorado at Boulder, US)
Grzegorz Chrupala (Saarland University, Germany)
Benoit Crabbé (University of Paris 7, France)
Josef van Genabith (Dublin City University, Ireland)
Yoav Goldberg (Google Research NY, US)
Spence Green (Stanford University, US)

Veronique Hoste (University College Ghent, Belgium)
Samar Husain (Potsdam University, Germany)
Sandra Kübler (Indiana University, US)
Jonas Kuhn (University of Stuttgart, Germany)
Mirella Lapata (University of Edinburgh, UK)
Alberto Lavelli (FBK-irst, Italy)
Alessandro Lenci (University of Pisa, Italy)
Joseph Le Roux (Université Paris-Nord, France)
Wolfgang Maier (University of Düsseldorf, Germany)
Nitin Madnani (Educational Testing Service, NJ)
Takuya Matsuzaki (University of Tokyo, Japan)
Aurélien Max (LIMSI-CNRS, France)
Yusuke Miyao (University of Tokyo, Japan)
Preslav Nakov (Qatar Computing Research Institute, Qatar)
Roberto Navigli (Sapienza University of Rome, Italy)
Kemal Oflazer (Carnegie Mellon University, Qatar)
Sebastian Pado (University of Heidelberg, Germany)
Patrick Pantel (Microsoft Research, US)
Sameer Pradhan (BBN Technologies, US)
Benoit Sagot (INRIA Rocquencourt, France)
Kenji Sagae (University of Southern California, US)
Idan Szpektor (Bar-Ilan University, Israel)
Lamia Tounsi (Dublin City University, Ireland)
Tim Van de Cruys (University of Cambridge, UK)
Stephen Wan (CSIRO ICT Centre, Australia)
Deniz Yuret (Koc University Istanbul, Turkey)
Zdenek Zabokrtsky (Charles University, Czech Republic)
Wajdi Zaghouani (Université de Montréal, Canada)
Shiqi Zhao (Baidu Inc., China)

Table of Contents

Conference Program

Thursday, July 12, 2012

Session 1: (08:50-10:05) Opening Session

08:50-09:05 Statistical Parsing and Semantic Processing of MRLs: Overview of the workshop by Reut Tsarfaty

09:05-10:05 Invited Talk (I) by Ivan Titov

Session 2: (10:05-10:30) Syntactic Parsing of MRLs (I)

10:05–10:30 *Probabilistic Lexical Generalization for French Dependency Parsing*
Enrique Henestroza Anguiano and Marie Candito

10:30-11:00 Coffee Break

Session 3: (11:00-12:25) Semantic Processing of MRLs

11:00–11:25 *Supervised Learning of German Qualia Relations*
Yannick Versley

11:25–11:40 *Building an Arabic Multiword Expressions Repository*
Abdelati Hawwari, Kfir Bar and Mona Diab

11:40–11:55 *Unsupervised frame based Semantic Role Induction: application to French and English*
Alejandra Lorenzo and Christophe Cerisara

11:55–12:10 *Using Synthetic Compounds for Word Sense Discrimination*
Yannick Versley and Verena Henrich

12:10–12:25 *Machine Learning of Syntactic Attachment from Morphosyntactic and Semantic Co-occurrence Statistics*
Szymon Acedański, Adam Slaski and Adam Przepiórkowski

12:30-14:00 Lunch Break

Thursday, July 12, 2012 (continued)

Session 4: (14:00-15:30) Syntactic Parsing of MRLs (II)

14:00-15:00 Invited Talk (II) by Mark Steedman

15:00–15:15 *Combining Rule-Based and Statistical Syntactic Analyzers*
Iakes Goenaga, Koldobika Gojenola, María Jesús Aranzabe, Arantza Díaz de Ilarraza and Kepa Bengoetxea

15:15–15:30 *Statistical Parsing of Spanish and Data Driven Lemmatization*
Joseph Le Roux, Benoit Sagot and Djamé Seddah

15:30-16:00 Coffee Break

Session 5: (16:00-17:30) Syntactic Parsing of MRLs (III)

16:00–16:25 *Assigning Deep Lexical Types Using Structured Classifier Features for Grammatical Dependencies*
João Silva and António Branco

16:25–16:40 *Using an SVM Ensemble System for Improved Tamil Dependency Parsing*
Nathan Green, Loganathan Ramasamy and Zdeněk Žabokrtský

16:40–17:05 *Korean Treebank Transformation for Parser Training*
DongHyun Choi, Jungyeul Park and Key-Sun Choi

17:05–17:30 *Generative Constituent Parsing and Discriminative Dependency Reranking: Experiments on English and French*
Joseph Le Roux, Benoit Favre, Alexis Nasr and Seyed Abolghasem Mirroshandel

17:30-17:40 Short Break

Thursday, July 12, 2012 (continued)

Session 6: (17:40-18:20) Closing Session

17:40-18:10	Panel: Disclosing the SPMRL 2013 Shared Task
18:10-18:20	Concluding Remarks by Reut Tsarfaty