ACL 2007

PRAGUE

ACL 2007

Proceedings of the Workshop on Deep Linguistic Processing

June 28, 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

This workshop was conceived with the aim of bringing together the different computational linguistic sub-communities which model language predominantly by way of theoretical syntax, either in the form of a particular theory (e.g. CCG, HPSG, LFG, TAG or the Prague School) or a more general framework which draws on theoretical and descriptive linguistics. We characterise this style of computational linguistic research as deep linguistic processing, due to it aspiring to model the complexity of natural language in rich linguistic representations. Aspects of this research have in the past had their own separate fora, such as the ACL 2005 workshop on deep lexical acquisition, as well as TAG+, Alpino, ParGram and DELPH-IN meetings. However, since the fundamental approach of building a linguistically-founded system, as well as many of the techniques used to engineer efficient systems, are common across these projects and independent of the specific grammar formalism chosen, we felt the need for a common meeting in which experiences could be shared among a wider community.

Deep linguistic processing has traditionally been concerned with grammar development for parsing and generation, with many deep processing systems using the same grammar for both directions. The linguistic precision and complexity of the grammars meant that they had to be manually developed and maintained, and were computationally expensive to run. With recent developments in computer hardware, parsing and generation algorithms and statistical learning theory, the way has been opened for deep linguistic processing to be successfully applied to an ever-growing range of languages, domains and applications.

The same trends that have made broad-coverage deep linguistic processing feasible have occurred at the same time as the rise of machine learning and statistical approaches to natural language processing. For a time, these two approaches were pursued separately, often without reference to advances in the other approach, even when the same problems were being addressed. In the past couple of years, this divide has begun to close from both sides. As witnessed by many of the papers in this workshop, many deep systems have statistical components to them (e.g., as pre- or post-processing to control ambiguity, as means of acquiring and extending lexical resources) or even use machine learning techniques to acquire deep grammars (semi-)automatically. From the other side of the divide, many of the largely statistical approaches are using progressively richer linguistic based features and are taking advantage of these deeper features to tackle problems traditionally reserved for deep systems, such as thematic role labelling.

The workshop has indeed brought together a range of theoretical perspectives, not just those originally foreseen. The papers presented cover current approaches to grammar development and issues of theoretical properties, as well as the application of deep linguistic techniques to large-scale applications such as question answering and dialogue systems. Having industrial-scale, efficient parsers and generators opens up new application domains for natural language processing, as well as interesting new ways in which to approach existing applications, e.g., by combining statistical and deep processing techniques in a triage process to process massive data quickly and accurately at a fine level of detail. Notably, several of the papers addressed the relationship of deep linguistic processing to topical statistical approaches, in particular in the area of parsing.

There were 45 submissions to the workshop, each of which was peer reviewed by three members of the international programme committee; at the end of the process 10 were accepted as papers to be presented orally and 10 as posters. We feel that such a large number of submissions for a one-day workshop reflects an increasing interest in deep linguistic processing, an interest which is buoyed by the realization that new, often hybrid, techniques combined with highly engineered parsers and generators and state-of-the-art machines open the way towards practical, real-world application of this research. We look forward to further opportunities for the different computational linguistic sub-communities who took part in this workshop, and others, to come together in the future.

We would like to thank all the authors who submitted papers, as well as the members of the programme committee for the time and effort they contributed in reviewing the papers, in some cases at very short notice. We should also like to thank Anette Frank for providing the perfect complement to the workshop with her invited talk.

The workshop received sponsorship from the Large Scale Syntactic Annotation of written Dutch (Lassy) project. The Lassy project is carried out within the STEVIN programme, which is funded by the Dutch and Flemish governments (`http://taalunieversum.org/taal/technologie/stevin/`).

Timothy Baldwin
Mark Dras
Julia Hockenmaier
Tracy Holloway King
Gertjan van Noord

Organizers

Chairs:

Timothy Baldwin (University of Melbourne)
Mark Dras (Macquarie University)
Julia Hockenmaier (University of Pennsylvania)
Tracy Holloway King (PARC)
Gertjan van Noord (University of Groningen)

Program Committee:

Jason Baldridge (University of Texas at Austin)
Emily Bender (University of Washington)
Raffaella Bernardi (University of Bolzano)
Francis Bond (NICT)
Gosse Bouma (University of Groningen)
Ted Briscoe (University of Cambridge)
Miriam Butt (University of Konstanz)
Aoife Cahill (Stuttgart University)
David Chiang (ISI)
Stephen Clark (Oxford University)
Ann Copestake (University of Cambridge)
James Curran (University of Sydney)
Stefanie Dipper (Potsdam University)
Katrin Erk (University of Texas at Austin)
Dominique Estival (Appen Pty Ltd)
Dan Flickinger (Stanford University)
Anette Frank (University of Heidelberg)
Josef van Genabith (Dublin City University)
John Hale (Michigan State University)
Ben Hutchinson (Google)
Mark Johnson (Brown University)
Aravind Joshi (University of Pennsylvania)
Laura Kallmeyer (Tübingen University)
Ron Kaplan (Powerset)
Martin Kay (Stanford University/Saarland University)
Valia Kordoni (Saarland University)
Anna Korhonen (University of Cambridge)
Jonas Kuhn (Potsdam University)
Rob Malouf (San Diego State University)
Ryan McDonald (Google)
Yusuke Miyao (University of Tokyo)
Diego Molla (Macquarie University)

Stefan Müller (Bremen University)
Joakim Nivre (Växjö University)
Stephan Oepen (University of Oslo and Stanford University)
Anoop Sarkar (Simon Fraser University)
David Schlangen (Potsdam University)
Mark Steedman (University of Edinburgh)
Beata Trawinski (Tübingen University)
Aline Villavicencio (Federal University of Rio Grande do Sul)
Tom Wasow (Stanford University)
Michael White (Ohio State University)
Shuly Wintner (University of Haifa)
Fei Xia (University of Washington)

Invited Speaker:

Anette Frank (University of Heidelberg)

Table of Contents

Conference Program

08:35–08:45 Opening Remarks

SESSION 1: PARSING

08:45–09:15 *Multi-Component Tree Adjoining Grammars, Dependency Graph Models, and Linguistic Analyses*
Joan Chen-Main and Aravind Joshi

09:15–09:45 *Perceptron Training for a Wide-Coverage Lexicalized-Grammar Parser*
Stephen Clark and James Curran

09:45–10:15 *Filling Statistics with Linguistics – Property Design for the Disambiguation of German LFG Parses*
Martin Forst

10:15–10:45 *Exploiting Semantic Information for HPSG Parse Selection*
Sanae Fujita, Francis Bond, Stephan Oepen and Takaaki Tanaka

10:45–11:15 COFFEE BREAK

SESSION 2: APPLICATIONS OF DEEP LINGUISTIC PROCESSING

11:15–11:45 *Deep Grammars in a Tree Labeling Approach to Syntax-based Statistical Machine Translation*
Mark Hopkins and Jonas Kuhn

11:45–12:15 *Question Answering based on Semantic Roles*
Michael Kaisser and Bonnie Webber

12:15–13:45 LUNCH

13:45–14:45 INVITED TALK

Across Languages and Grammar Paradigms – New Perspectives on Resource Acquisition, Grammar Engineering and Application
Anette Frank

Session 3: Posters

14:45–15:45 *Deep Linguistic Processing for Spoken Dialogue Systems*
James Allen, Myroslava Dzikovska, Mehdi Manshadi and Mary Swift

Self- or Pre-Tuning? Deep Linguistic Processing of Language Variants
Branco António and Costa Francisco

Pruning the Search Space of a Hand-Crafted Parsing System with a Probabilistic Parser
Aoife Cahill, Tracy Holloway King and John T. Maxwell III

Semantic Composition with (Robust) Minimal Recursion Semantics
Ann Copestake

A Task-based Comparison of Information Extraction Pattern Models
Mark Greenwood and Mark Stevenson

Creating a Systemic Functional Grammar Corpus from the Penn Treebank
Matthew Honnibal and James R. Curran

Verb Valency Semantic Representation for Deep Linguistic Processing
Aleš Horák, Karel Pala, Marie Duží and Pavel Materna

The Spanish Resource Grammar: Pre-processing Strategy and Lexical Acquisition
Montserrat Marimon, Nria Bel, Sergio Espeja and Natalia Seghezzi

Extracting a Verb Lexicon for Deep Parsing from FrameNet
Mark McConville and Myroslava O. Dzikovska

Fips, A "Deep" Linguistic Multilingual Parser
Eric Wehrli

Partial Parse Selection for Robust Deep Processing
Yi Zhang, Valia Kordoni and Erin Fitzgerald

15:45–16:15 Coffee Break

SESSION 4: GRAMMAR ENGINEERING