Coling 2008

# 22nd International Conference on Computational Linguistics

# Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation

Workshop chairs:
Johan Bos, Edward Briscoe, Aoife Cahill, John Carroll, Stephen Clark, Ann Copestake, Dan Flickinger, Josef van Genabith, Julia Hockenmaier, Aravind Joshi, Ronald Kaplan, Tracy Holloway King, Sandra Kübler, Dekang Lin, Jan Tore Lønning, Christopher Manning, Yusuke Miyao, Joakim Nivre, Stephan Oepen, Kenji Sagae, Nianwen Xue, and Yi Zhang

23 August 2008
Manchester, UK

Order copies of this and other Coling 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

ISBN 978-1-905593-50-7

*Design by Chimney Design, Brighton, UK*
*Production and manufacture by One Digital, Brighton, UK*

# Introduction

Broad-coverage parsing has come to a point where distinct approaches can offer (seemingly) comparable performance: statistical parsers acquired from the Penn Treebank (PTB); data-driven dependency parsers; 'deep' parsers trained off enriched treebanks (in linguistic frameworks like CCG, HPSG, or LFG); and hybrid 'deep' parsers, employing hand-built grammars in, for example, HPSG, LFG, or LTAG. Evaluation against trees in the Wall Street Journal (WSJ) section of the PTB has helped advance parsing research over the course of the past decade. Despite some scepticism, the crisp and, over time, stable task of maximizing ParsEval metrics (i.e. constituent labeling precision and recall) over PTB trees has served as a dominating benchmark. However, modern treebank parsers still restrict themselves to only a subset of PTB annotation; there is reason to worry about the idiosyncrasies of this particular corpus; it remains unknown how much the ParsEval metric (or any intrinsic evaluation) can inform NLP application developers; and PTB-style analyses leave a lot to be desired in terms of linguistic information.

The Grammatical Relations (GR) scheme, inspired by Dependency Grammar, offers a level of abstraction over specific syntactic analyses. It aims to capture the 'gist' of grammatical relations in a fashion that avoids reference to a token linguistic theory. GR has recently been applied successfully in a series of cross-framework parser evaluation studies. At the same time, rather little GR gold standard data is available, and the GR scheme has been questioned for some of its design decisions. More specifically, GR builds on a combination of syntactic and, albeit very limited, some semantic information. Existing studies suggest that the GR gold standard can be both overly rich and overly shallow in some respects. Furthermore, the mapping of 'native' parser outputs into GR introduces noise, and it raises a number of theoretical and practical questions.

Gold standard representations at the level of propositional semantics have at times been proposed for cross-framework parser evaluation, specifically where the parsing task is broadly construed as a tool towards 'text understanding', i.e. where the parser is to provide all information that is grammaticalized and contributing to interpretation. PropBank would seem a candidate gold standard, but to date very few studies exist that report on the use of PropBank for parser evaluation. The reasons might be that (at least some) parser developers believe that PropBank goes too far beyond the grammatical level to serve for parser evaluation, and that starting from PTB structures may have led to some questionable annotation decisions.

Finally, a complementary topic to cross-framework evaluation is the increasing demand for cross-domain parser evaluation. At conferences in 2007, concerns were expressed about results that might rely on particular properties of the WSJ PTB, and over idiosyncrasies of this specific sample of natural language. For example, it remains a largely open question to what degree progress made in PTB parsing can carry over to other genres and domains; a related question is on the fitness of some specific approach (when measured in parser evaluation metrics) for actual NLP applications. In summary, it may be necessary that the WSJ- and PTB-derived parser benchmarks be complemented by other gold standards, both in terms of the selection of texts and target representations. And to further the adaptation of parser evaluation to more languages, it will be important to carefully distill community experience from ParsEval and GR evaluations.

This workshop aims to bring together developers of broad-coverage parsers who are interested in questions of target representations and cross-framework and cross-domain evaluation and benchmarking. From informal discussions that the co-organizers had among themselves and with colleagues, it seems evident that there is comparatively broad awareness of current issues in parser evaluation, and a lively interest in detailed exchange of experience (and beliefs). Specifically, the organizers have tried to attract representatives from diverse parsing approaches and frameworks, ranging from 'traditional' treebank parsing, over data-driven dependency parsing, to parsing in specific linguistic frameworks. For the latter class of parsers, in many frameworks there is a further sub-division into groups pursuing 'classic' grammar engineering vs. ones who rely on grammar acquisition from annotated corpora.

Quite likely for the first time in the history of these approaches, there now exist large, broad-coverage parsing systems representing diverse traditions that can be applied to running text, often producing comparable representations. In our view, these recent developments present a new opportunity for re-energizing parser evaluation research. We sincerely wish this workshop will provide participants with the opportunity for in-depth and cross-framework exchange of expertise and discussion of future directions in parser evaluation.

A specific sub-goal of the workshop is to establish an improved shared knowledge among participants of the strengths and weaknesses of extant annotation and evaluation schemes. In order to create a joint focus for detailed discussion, the workshop preparation included a 'lightweight' shared task. For a selection of 50 sentences (of which ten were considered obligatory, the rest optional) for which PTB, GR, and PropBank (and other) annotations are available, contributors were invited to scrutinize existing gold-standard representations contrastively, identify perceived deficiencies, and sketch what can be done to address these. As an optional component, participants in the shared task were welcome to include 'native', framework-specific output representations and actual results for a parsing system of their choice (be it their own or not) in the contrastive study. In either case, submissions to the shared task reflect on the nature of different representations, highlight which additional distinctions are made in either scheme, and argue why these are useful (for some task) or unmotivated (in general). Of the eight papers selected for presentation at the workshop, the following three were submissions to the shared task, viz. those by Flickinger (page 17), Tateisi (page 24), and McConville and Dzikovska (page 51). For further information on the workshop as a whole, its shared task, and some specific datasets used, please see:

```
http://lingo.stanford.edu/events/08/pe/
```

**Organizers:**

Johan Bos, University of Rome ‘La Sapienza’ (Italy)
Edward Briscoe, University of Cambridge (UK)
Aoife Cahill, University of Stuttgart (Germany)
John Carroll, University of Sussex (UK)
Stephen Clark, Oxford University (UK)
Ann Copestake, University of Cambridge (UK)
Dan Flickinger, Stanford University (USA)
Josef van Genabith, Dublin City University (Ireland)
Julia Hockenmaier, University of Illinois at Urbana-Champaign (USA)
Aravind Joshi, University of Pennsylvania (USA)
Ronald Kaplan, Powerset, Inc. (USA)
Tracy Holloway King, PARC (USA)
Sandra Kübler, Indiana University (USA)
Dekang Lin, Google Inc. (USA)
Jan Tore Lønning, University of Oslo (Norway)
Christopher Manning, Stanford University (USA)
Yusuke Miyao, University of Tokyo (Japan)
Joakim Nivre, Växjö and Uppsala Universities (Sweden)
Stephan Oepen, University of Oslo (Norway) and CSLI Stanford (USA)
Kenji Sagae, University of Southern California (USA)
Nianwen Xue, University of Colorado (USA)
Yi Zhang, DFKI GmbH and Saarland University (Germany)

# Table of Contents

# Conference Programme

**Saturday, August 23, 2008**

| | |
|---|---|
| 9:00–9:30 | Workshop Motivation and Overview (Cahill, Oepen, et al.) |
| 9:30–10:00 | *The Stanford Typed Dependencies Representation*<br>Marie-Catherine de Marneffe and Christopher D. Manning |
| 10:00–10:30 | *Exploring an Auxiliary Distribution Based Approach to Domain Adaptation of a Syntactic Disambiguation Model*<br>Barbara Plank and Gertjan van Noord |
| 10:30–11:00 | Coffee Break |
| 11:00–11:30 | *Toward an Underspecifiable Corpus Annotation Scheme*<br>Yuka Tateisi |
| 11:30–12:00 | *Toward a Cross-Framework Parser Annotation Standard*<br>Dan Flickinger |
| 12:00–12:30 | Discussion |
| 12:30–14:00 | Lunch Break |
| 14:00–14:30 | Summary of CoNLL 2008 Shared Task (Nivre) |
| 14:30–15:00 | *Parser Evaluation Across Frameworks without Format Conversion*<br>Wai Lok Tam, Yo Sato, Yusuke Miyao and Junichi Tsujii |
| 15:00–15:30 | *Large Scale Production of Syntactic Annotations to Move Forward*<br>Anne Vilnat, Gil Francopoulo, Olivier Hamon, Sylvain Loiseau, Patrick Paroubek, and Eric Villemonte de la Clergerie |
| 15:30–16:00 | Coffee Break |
| 16:00–16:30 | *Constructing a Parser Evaluation Scheme*<br>Laura Rimell and Stephen Clark |
| 16:30–17:00 | *'Deep' Grammatical Relations for Semantic Interpretation*<br>Mark McConville and Myroslava O. Dzikovska |
| 17:00–17:30 | Discussion |