The Impact of Deep Linguistic Processing on Parsing Technology

Timothy Baldwin
University of Melbourne
tim@csse.unimelb.edu.au

Mark Dras
Macquarie University
madras@ics.mq.edu.au

Julia Hockenmaier
University of Pennsylvania
juliahr@cis.upenn.edu

Tracy Holloway King
PARC
thking@parc.com

Gertjan van Noord
University of Groningen
vannoord@let.rug.nl

Abstract

As the organizers of the ACL 2007 Deep Linguistic Processing workshop (Baldwin et al., 2007), we were asked to discuss our perspectives on the role of current trends in deep linguistic processing for parsing technology. We are particularly interested in the ways in which efficient, broad coverage parsing systems for linguistically expressive grammars can be built and integrated into applications which require richer syntactic structures than shallow approaches can provide. This often requires hybrid technologies which use shallow or statistical methods for pre- or post-processing, to extend coverage, or to disambiguate the output.

1 Introduction

Our talk will provide a view on the relevance of deep linguistic processing for parsing technologies from the perspective of the organizers of the ACL 2007 Workshop on Deep Linguistic Processing (Baldwin et al., 2007). The workshop was conceived with the broader 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, the Prague School) or a more general framework which draws on theoretical and descriptive linguistics. These "deep linguistic processing" approaches differ from shallower methods in that they yield richer, more expressive, structural representations which capture long-distance dependencies or the underlying predicate-argument structure directly.

Aspects of this research have often had their own separate fora, such as the ACL 2005 workshop on deep lexical acquisition (Baldwin et al., 2005), as well as the TAG+ (Kallmeyer and Becker, 2006), Alpino (van der Beek et al., 2005), ParGram (Butt et al., 2002) and DELPH-IN (Oepen et al., 2002) projects and meetings. However, the fundamental approaches to building a linguistically-founded system and many of the techniques used to engineer efficient systems are common across these projects and independent of the specific grammar formalism chosen. As such, we felt the need for a common meeting in which experiences could be shared among a wider community, similar to the role played by recent meetings on grammar engineering (Wintner, 2006; Bender and King, 2007).

2 The promise of deep parsing

Deep linguistic processing has traditionally been concerned with grammar development (for use in both parsing and generation). However, the linguistic precision and complexity of the grammars meant that they had to be manually developed and maintained, and were computationally expensive to run.

In recent years, machine learning approaches have fundamentally altered the field of natural language processing. The availability of large, manually annotated, treebanks (which typically take years of prior linguistic groundwork to produce) enabled the rapid creation of robust, wide-coverage parsers. However, the standard evaluation metrics for which such parsers have been optimized generally ignore

Proceedings of the 10th Conference on Parsing Technologies, pages 36–38,
Prague, Czech Republic, June 2007.

much of the rich linguistic information in the original treebanks. It is therefore perhaps only natural that deep processing methods, which often require substantial amounts of manual labor, have received considerably less attention during this period.

But even if further work is required for deep processing techniques to fully mature, we believe that applications that require natural language understanding or inference, among others, will ultimately need detailed syntactic representations (capturing, e.g., bounded and unbounded long-range dependencies) from which semantic interpretations can easily be built. There is already some evidence that our current deep techniques can, in some cases, outperform shallow approaches. There has been work demonstrating this in question answering, targeted information extraction and the recent textual entailment recognition task, and perhaps most notably in machine translation: in this latter field, after a period of little use of linguistic knowledge, deeper techniques are beginning to lead to better performance, e.g. by redefining phrases by syntactic "treelets" rather than contiguous word sequences, or by explicitly including a syntactic component in the probability model, or by syntactic preprocessing of the data.

3 Closing the divide

In the past few years, the divide between "deep", rule-based, methods and "shallow", statistical, approaches, has begun to close from both sides. Recent advances in using the same treebanks that have advanced shallow techniques to extract more expressive grammars or to train statistical disambiguators for them, and in developing framework-specific treebanks, have made it possible to obtain similar coverage, robustness, and disambiguation accuracy for parsers that use richer structural representations. As witnessed by many of the papers in our workshop (Baldwin et al., 2007), a large proportion of current 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 automatically. From the other side of the divide, many of the purely statistical approaches are using progressively richer linguistic features and are taking advantage of these more expressive features to tackle problems that were traditionally thought to require deep systems, such as the recovery of traces or semantic roles.

4 The continued need for research on deep processing

Although statistical techniques are becoming commonplace even for systems built around hand-written grammars, there is still a need for further linguistic research and manual grammar development. For example, supervised machine-learning approaches rely on large amounts of manually annotated data. Where such data are available, developers of deep parsers and grammars can exploit them to determine frequency of certain constructions, to bootstrap gold standards for their systems, and to provide training data for the statistical components of their systems such as parse disambiguators. But for the majority of the world's languages, and even for many languages with large numbers of speakers, such corpora are unavailable. Under these circumstances, manual grammar development is unavoidable, and recent progress has allowed the underlying systems to become increasingly better engineered, allowing for more rapid development of any given grammar, as well as for overlay grammars that adapt to particular domains and applications and for porting of grammars from one language to another.

Despite recent work on (mostly dependency grammar-based) multilingual parsing, it is still the case that most research on statistical parsing is done on English, a fixed word-order language where simple context-free approximations are often sufficient. It is unclear whether our current models and algorithms carry over to morphologically richer languages with more flexible word order, and it is possible that the more complex structural representations allowed by expressive formalisms will cease to remain a luxury.

Further research is required on all aspects of deep linguistic processing, including novel linguistic analyses and implementations for different languages, formal comparisons of different frameworks, efficient parse and learning algorithms, better statistical models, innovative uses of existing data resources, and new evaluation tools and methodologies. We were fortunate to receive so many highquality submissions on all of these topics for our workshop.

5 Conclusion and outlook

Deep linguistic processing brings together a range of perspectives. It covers current approaches to grammar development and issues of theoretical linguistic and algorithmic properties, as well as the application of deep linguistic techniques to large-scale applications such as question answering and dialog 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 is 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 opens the way towards practical, real-world application of this research. We look forward to further opportunities for the different computational linguistic subcommunities who took part in this workshop, and others, to continue to come together in the future.

References

Timothy Baldwin, Anna Korhonen, and Aline Villavicencio, editors. 2005. *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*. Ann Arbor, USA.

Timothy Baldwin, Mark Dras, Julia Hockenmaier, Tracy Holloway King, and Gertjan van Noord, editors. 2007. *Proceedings of the ACL Workshop on Deep Linguistic Processing*, Prague, Czech Republic.

Emily Bender and Tracy Holloway King, editors. 2007. *Grammar Engineering Across Frameworks*, Stanford University. CSLI On-line Publications. to appear.

Miriam Butt, Helge Dyvik, T. H. King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *COLING Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

Laura Kallmeyer and Tilman Becker, editors. 2006. *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+)*, Sydney, Australia.

Stephan Oepen, Dan Flickinger, J. Tsujii, and Hand Uszkoreit, editors. 2002. *Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing*. CSLI Publications.

Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Mark-Jan Nederhof, Gertjan van Noord, Robbert Prins, and Bego na Villada Moirón. 2005. Algorithms for linguistic processing. NWO Pionier final report. Technical report, University of Groningen.

Shuly Wintner. 2006. Large-scale grammar development and grammar engineering. Research workshop of the Israel Science Foundation.