Data-Driven Dependency Parsing across Languages and Domains: Perspectives from the CoNLL 2007 Shared Task

Joakim Nivre
Växjö University, School of Mathematics and Systems Engineering
Uppsala University, Department of Linguistics and Philology
E-mail: nivre@msi.vxu.se

Abstract

The Conference on Computational Natural Language Learning features a shared task, in which participants train and test their learning systems on the same data sets. In 2007, as in 2006, the shared task has been devoted to dependency parsing, this year with both a multilingual track and a domain adaptation track. In this paper, I summarize the main findings from the 2007 shared task and try to identify major challenges for the parsing community based on these findings.

1 Introduction

The annual Conference on Computational Natural Language Learning (CoNLL) has for the past nine years organized a *shared task*, where participants train and test their learning systems on the same data sets. In 2006, the shared task was multilingual dependency parsing, where participants had to train and test a parser on data from thirteen different languages (Buchholz and Marsi, 2006). In 2007, the task was extended by adding a second track for (monolingual) domain adaptation.

The CoNLL 2007 shared task on dependency parsing featured two tracks:

- In the *multilingual track*, the task was to train a parser using labeled data from Arabic, Basque, Catalan, Chinese, Czech, English, Greek, Hungarian, Italian, and Turkish.
- In the *domain adaptation track*, the task was to adapt a parser for English news text to other domains using unlabeled data from the target domains: biomedical and chemical abstracts, parent-child dialogues.[1] In the *closed class*, the base parser had to be trained using the English training set for the multilingual track and no external resources were allowed. In the *open class*, any base parser could be used and any external resources were allowed.

Both tracks used the same column-based format for labeled data with six input columns and two output columns for each word of a sentence:

- Input: word-id, word form, lemma, coarse part of speech, fine part-of-speech, morphosyntactic features.
- Output: head (word-id), dependency label.

The main evaluation metric for both tracks was the *labeled attachment score* (LAS), i.e., the percentage of words that have been assigned the correct head and dependency label. For more information about the setup, see Nivre et al. (2007)

In this paper, I will summarize the main findings from the CoNLL 2007 shared task, starting with a characterization of the different approaches used (section 2), and moving on to the most interesting results in the multilingual track (section 3) and the domain adaptation track (section 4). Finally, based on these findings, I will try to identify some important challenges for the wider parsing community (section 5).

[1] The biomedical domain was the development domain, which means that a small labeled development set was available for this domain. The final testing was only done on chemical abstracts and (optionally) parent-child dialogues.

Proceedings of the 10th Conference on Parsing Technologies, pages 168–170,
Prague, Czech Republic, June 2007.

2 Approaches

In total, test runs were submitted for twenty-three systems in the multilingual track, and ten systems in the domain adaptation track (six of which also participated in the multilingual track). The majority of these systems used models belonging to one of the two dominant approaches in data-driven dependency parsing in recent years (McDonald and Nivre, 2007):

- In *graph-based models*, every possible dependency graph for a given input sentence is given a score that decomposes into scores for the arcs of the graph. The optimal parse can be found using a spanning tree algorithm (Eisner, 1996; McDonald et al., 2005).
- In *transition-based models*, dependency graphs are modeled by sequences of parsing actions (or transitions) for building them. The search for an optimal parse is often deterministic and guided by classifiers (Yamada and Matsumoto, 2003; Nivre, 2003).

The majority of graph-based parsers in the shared task were based on what McDonald and Pereira (2006) call the first-order model, where the score of each arc is independent of every other arc, but there were also attempts at exploring higher-order models, either with exact inference limited to projective dependency graphs (Carreras, 2007), or with approximate inference (Nakagawa, 2007). Another innovation was the use of k-best spanning tree algorithms for inference with a non-projective first-order model (Hall et al., 2007b).

For transition-based parsers, the trend was clearly to move away from deterministic parsing by adding a probability model for scoring a set of candidate parses typically derived using a heuristic search strategy. The probability model may be either conditional (Duan et al., 2007) or generative (Titov and Henderson, 2007).

An interesting way of combining the two main approaches is to use a graph-based model to build an ensemble of transition-based parsers. This technique, first proposed by Sagae and Lavie (2006), was used in the highest scoring system in both the multilingual track (Hall et al., 2007a) and the domain adaptation track (Sagae and Tsujii, 2007).

3 Multilingual Parsing

The ten languages involved in the multilingual track can be grouped into three classes with respect to the best parsing accuracy achieved:

- Low (LAS = 76.3–76.9):
  Arabic, Basque, Greek
- Medium (LAS = 79.2–80.2):
  Czech, Hungarian, Turkish
- High (LAS = 84.4–89.6):
  Catalan, Chinese, English, Italian

To a large extent, these classes appear to be definable from typological properties. The class with the highest top scores contains languages with a rather impoverished morphology. Medium scores are reached by the two agglutinative languages, Hungarian and Turkish, as well as by Czech. The most difficult languages are those that combine a relatively free word order with a high degree of inflection. Based on these characteristics, one would expect to find Czech in the last class. However, the Czech training set is four times the size of the training set for Arabic, which is the language with the largest training set of the difficult languages. On the whole, however, training set size alone is a poor predictor of parsing accuracy, which can be seen from the fact that the Italian training set is only about half the size of the Arabic one and only one sixth of Czech one. Thus, there seems to be a need for parsing methods that can cope better with richly inflected languages.

4 Domain Adaptation

One result from the domain adaptation track that may seem surprising at first was the fact that the best closed class systems outperformed the best open class systems on the official test set containing chemical abstracts. To some extent, this may be explained by the greater number of participants in the closed class (eight vs. four). However, it also seems that the major problem in adapting existing, often grammar-based, parsers to the new domain was not the domain as such but the mapping from the native output of the parser to the kind of annotation provided in the shared task data sets. In this respect, the closed class systems had an advantage by having been trained on exactly this kind of annotation. This result serves to highlight the fact that domain adaptation, as well as the integration of grammar-based and data-driven methods, often involves transformations between different kinds of linguistic representations.

The best performing (closed class) system in the domain adaptation track used a combination of co-learning and active learning by training two different parsers on the labeled training data, parsing the unlabeled domain data with both parsers, and adding parsed sentences to the training data only if the two parsers agreed on their analysis (Sagae and Tsujii, 2007). This resulted in a LAS of 81.1 on the test set of chemical abstracts, to be compared with 89.0 for the English test set in the multilingual track.

5 Conclusion

Based on the results from the CoNLL 2007 shared task, it is clear that we need to improve our methods for parsing richly inflected languages. We also need to find better ways of integrating parsers developed within different frameworks, so that they can be reused effectively for, among other things, domain adaptation. More generally, we need to increase our knowledge of the multi-causal relationship between language characteristics, syntactic representations, and parsing and learning methods. In order to do this, perhaps we also need a shared task at the International Conference on Parsing Technologies.

Acknowledgments

I want to thank my fellow organizers of the shared task, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret, who are also co-authors of the longer paper on which this paper is partly based (Nivre et al., 2007). I am also indebted to all the people who have contributed to the shared task by providing data or participating.

References

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of CoNLL*, 149–164.

X. Carreras. 2007. Experiments with a high-order projective dependency parser. In *Proc. of EMNLP-CoNLL (Shared Task)*.

X. Duan, J. Zhao, and B. Xu. 2007. Probabilistic parsing action models for multi-lingual dependency parsing. In *Proc. of EMNLP-CoNLL (Shared Task)*.

J. M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proc. of COLING*, 340–345.

J. Hall, J. Nilsson, J. Nivre, G. Eryigit, B. Megyesi, M. Nilsson, and M. Saers. 2007a. Single malt or blended? A study in multilingual parser optimization. In *Proc. of EMNLP-CoNLL (Shared Task)*.

K. Hall, J. Havelka, and D. Smith. 2007b. Log-linear models of non-projective trees, k-best MST parsing and tree-ranking. In *Proc. of EMNLP-CoNLL (Shared Task)*.

R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proc. of EMNLP-CoNLL*.

R. McDonald and F. Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proc. of EACL*, 81–88.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proc. of HLT/EMNLP*, 523–530.

T. Nakagawa. 2007. Multilingual dependency parsing using gibbs sampling. In *Proc. of EMNLP-CoNLL (Shared Task)*.

J. Nivre and J. Nilsson. 2005. Pseudo-projective dependency parsing. In *Proc. of ACL*, 99–106.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. of EMNLP-CoNLL (Shared Task)*.

J. Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proc. of IWPT*, 149–160.

K. Sagae and A. Lavie. 2006. Parser combination by reparsing. In *Proc. of HLT-NAACL (Short Papers)*, 129–132.

K. Sagae and J. Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proc. of EMNLP-CoNLL (Shared Task)*.

I. Titov and J. Henderson. 2007. Fast and robust multilingual dependency parsing with a generative latent variable model. In *Proc. of EMNLP-CoNLL (Shared Task)*.

H. Yamada and Y. Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proc. of IWPT*, 195–206.