Adventures in Multilingual Parsing

Joakim Nivre
Uppsala university
Department of Linguistics and Philology
Uppsala, Sweden

1 Introduction

The typological diversity of the world's languages poses important challenges for the techniques used in machine translation, syntactic parsing and other areas of natural language processing. Statistical models developed and tuned for English do not necessarily perform well for richly inflected languages, where larger morphological paradigms and more flexible word order gives rise to data sparseness. Since paradigms can easily be captured in rule-based descriptions, this suggests that hybrid approaches combining statistical modeling with linguistic descriptions might be more effective. However, in order to gain more insight into the benefits of different techniques from a typological perspective, we also need linguistic resources that are comparable across languages, something that is currently lacking to a large extent.

In this talk, I will report on two ongoing projects that tackle these issues in different ways. In the first part, I will describe techniques for joint morphological and syntactic parsing that combines statistical dependency parsing and rule-based morphological analysis, specifically targeting the challenges posed by richly inflected languages. In the second part, I will present the Universal Dependency Treebank Project, a recent initiative seeking to create multilingual corpora with morphosyntactic annotation that is consistent across languages.

2 Morphological and Syntactic Parsing

In Bohnet et al. (2013), the goal is to improve parsing accuracy for morphologically rich languages by performing morphological and syntactic analysis jointly instead of in a pipeline. In this way, we can ideally make use of syntactic information to disambiguate morphology, and not just vice versa. We use a transition-based framework for dependency parsing, and explore different ways of integrating morphological features into the model. Furthermore, we investigate the use of rule-based morphological analyzers to provide hard or soft constraints in order to tackle the sparsity of lexical features. Evaluation on five morphologically rich languages (Czech, Finnish, German, Hungarian, and Russian) shows consistent improvements in both morphological and syntactic accuracy for joint prediction over a pipeline model, with further improvements thanks to the morphological analyzers. The final results improve the state of the art in dependency parsing for all languages.

3 Treebanks for Multilingual Parsing

In McDonald et al. (2013), we present a new collection of treebanks with homogeneous syntactic annotation for six languages: German, English, Swedish, Spanish, French and Korean. The annotation is based on the Google universal part-of-speech tags (Petrov et al., 2012) and the Stanford dependencies (de Marneffe et al., 2006), adapted and harmonized across languages. To show the usefulness of such a resource, we also present a case study of cross-lingual transfer parsing with more reliable evaluation than has been possible before. The 'universal' treebank is made freely available in order to facilitate research on multilingual dependency parsing.[1] A second release including eleven languages is planned for the spring of 2014.

4 Conclusion

Although both projects reviewed in the talk may contribute to a better understanding of how natural language processing techniques are affected by linguistic diversity, there are still important gaps that need to be filled. For instance, the universal treebank annotation still fails to capture most of the morphological categories used by the parser. In the final part of the talk, I will try to outline some of the challenges that lie ahead of us.

[1] Downloadable at https://code.google.com/p/uni-dep-tb/.

References

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.