Enabling Adaptation of Lexicalised Grammars to New Domains

Valia Kordoni & Yi Zhang
DFKI GmbH & Dept. of Computational Linguistics, Saarland University
PO Box 15 11 50, 66041 Saarbrücken, GERMANY
kordoni,yzhang@coli.uni-sb.de

Abstract

This extended abstract focuses on the main points we will be touching upon during our talk, the aim of which is to present in a concise manner our group's work on enhancing robustness of lexicalised grammars for real-life applications and thus also on enabling their adaptation to new domains in its entirety.

1 Introduction

At present, various wide-coverage symbolic parsing systems for different languages exist and have been integrated into real-world NLP applications, such as IE, QA, grammar checking, MT and intelligent IR. This integration, though, has reminded us of the shortcomings of symbolic systems, in particular lack of coverage, one consequence of which relates to enormous and sometimes insurmountable difficulties with porting and re-using such systems to new domains. When the hand-crafted grammars which usually lie at the heart of symbolic parsing systems are applied to naturally occurring text, we often find that they are underperforming. Typical sources of coverage deficiency include unknown words, words for which the dictionary did not contain the relevant category, Multiword Expressions (MWEs), but also more general grammatical knowledge, such as grammar rules and word ordering constraints. Currently, grammars and their accompanying lexica often need to be extended manually.

In this talk, we offer the overview to a range of machine learning-based methods which enable us to derive linguistic knowledge from corpora, for instance, in order to solve problems of coverage and efficiency deficiency of large-scale lexicalised grammars, ensuring this way their portability and re-usability and aiming at domain-independent linguistic processing. In particular, we illustrate and underline the importance of making detailed linguistic information a central part of the process of automatic acquisition of large-scale lexicons as a means for enhancing robustness and ensuring maintainability and re-usability of lexicalised grammars.

To this effect, we focus on enhancing robustness and ensuring maintainability and re-usability for two large-scale "deep" grammars, one of English (ERG; [3]) and one of German (GG; [4]), developed in the framework of Head-driven Phrase Structure Grammar (HPSG). Specifically, we show that the incorporation of detailed linguistic information into the process of automatic extension of the lexicon of such language resources enhances their performance and provides linguistically sound and more informative predictions which bring a bigger benefit for the grammars when employed in practical real-life applications.

2 Main Focus Points

In recent years, various techniques and resources have been developed in order to improve robustness of deep grammars for real-life applications in various domains. Nevertheless, low coverage of such grammars remains the main hindrance to their employment in open domain natural language processing. [2], as well as [6] and [7] have clearly shown that the majority of parsing failures with large-scale deep grammars are caused by missing or wrong entries in the lexica accompanying grammars like the aforementioned ones. Based on these findings, it has become clear that it is crucial to explore and come up with efficient methods for automated (Deep) Lexical Acquisition (henceforward (D)LA), the process of automatically recovering missing entries in the lexicons of deep grammars.

Recently, various high-quality DLA approaches have been proposed. [1], as well as [7] and [5] describe efficient methods towards the task of lexicon acquisition for large-scale deep grammars for English and Dutch. They treat DLA as a classification task and make use of various robust and efficient machine learning techniques to perform the acquisition process.

We use the ERG and GG grammars for the work we present in this talk, for the ERG is one of the biggest deep linguistic resources to date which has been being used in many (industrial) applications, and GG is to our knowledge one of the most thorough deep grammars of German, a language with rich morphology and free word order, which exhibits a range of interesting linguistic phenomena. Thus, the aforementioned grammars are valuable linguistic resources since they provide linguistically sound and detailed analyses of phenomena in English and German. Apart from the interesting syntactic structures, though, the lexical entries in the lexicons of the aforementioned grammars also exhibit a rich and complicated structure and contain various important linguistic constraints. Based on our claim above, in this talk we show how the information these constraints provide can be captured and used in linguistically-motivated DLA methods which we have developed. It has been shown that, comparing

Workshop Adaptation of Language Resources and Technology to New Domains 2009 - Borovets, Bulgaria, pages 48–49

to statistically treebank-based parsers, parsers based on these hand-written linguistic grammars have more consistent performance over changing of domains [8]. In this we prove our assumption that the linguistic information we incorporate into our DLA methods is vital for the good performance of the acquisition process and for the maintainability and re-usability of the grammars domain-independently, as well for their successful practical application.

References

[1] T. Baldwin. Bootstrapping deep lexical resources: Resources for courses. In *Proceedings of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 67–76, Ann Arbor, USA, 2005.

[2] T. Baldwin, E. M. Bender, D. Flickinger, A. Kim, and S. Oepen. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth Internation Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.

[3] A. Copestake and D. Flickinger. An open-sourse grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000.

[4] B. Crysmann. On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, pages 112–116, Borovets, Bulgaria, 2003.

[5] T. V. de Cruys. Automatically extending the lexicon for parsing. In *Proceedings of the Student Session of the European Summer School in Logic, Language and Information (ESSLLI)*, pages 180–191, Malaga, Spain, 2006.

[6] G. van Noord. Error mining for wide coverage grammar engineering. In *Proceedings of the 42nd Meeting of the Assiciation for Computational Linguistics (ACL'04), Main Volume*, pages 446–453, Barcelona, Spain, 2004.

[7] Y. Zhang and V. Kordoni. Automated deep lexical acquisition for robust open text processing. In *Proceedings of the Fifth International Conference on Language Resourses and Evaluation (LREC 2006)*, Genoa, Italy, 2006.

[8] Y. Zhang and R. Wang. Cross-domain dependency parsing using a deep linguistic grammar. In *Proceedings of ACL-IJCNLP 2009*, Singapore, 2009.