INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

Alina Wróblewska

Polish Dependency Parser Trained on an Automatically Induced Dependency Bank

PhD Dissertation

Supervisor
dr hab. Adam Przepiórkowski, prof. IPI PAN

Warsaw 2014

Abstract

Dependency parsing has become important for various language processing tasks in recent years. The predicate-argument structure transparently encoded in dependency-based syntactic representations may support machine translation, question answering, information extraction, etc. Many contemporary dependency parsing systems are based on statistical methods. Using training data, parsers learn how to analyse sentences and predict dependency structures that are appropriate for these sentences.

Different statistical methods have been applied to data-driven dependency parsing. However, the best results so far are given by supervised methods. Supervised dependency parsers trained on correctly annotated data may have high parsing performance even for languages with relatively free word order, such as Czech or Bulgarian.

Nevertheless, supervised methods require manually annotated training data. The creation of such data is a very time-consuming and expensive process. Therefore, there is still a lot of languages without any manually annotated data and alternative methods of parser training or data gathering are needed. Since unsupervised training – with its low performance and high complexity – is often an infeasible solution, we study alternative methods of gathering training data. In this dissertation we address two research questions:

1. Is it possible to gather dependency trees automatically (or with a minimal human involvement)?
2. Is it possible to train a good quality supervised dependency parser on automatically or semi-automatically induced training data?

Due to the increasing interest in data-driven parsing, several shared tasks on multilingual dependency parsing were organised at the Conference on Computational Natural Language Learning (Buchholz and Marsi, 2006; Nivre et al., 2007). Different languages were represented in these tasks, including some Slavic languages such as Slovene, Bulgarian and Czech. Polish was not represented in any of these tasks, probably due to the lack of dependency-annotated training data for this language. Furthermore, dependency parsing is hardly represented in the Polish NLP community. We are aware of neither experiments with data-driven Polish dependency parsing nor existence of any publicly available Polish dependency parser. The only Polish dependency parser was developed by Tomasz Obrębski within his doctoral research (Obrębski, 2002, 2003). However, this rule-based parser founded on the syntactic description of Polish by Saloni and Świdziński (1989) was only tested against a small artificial test set and no wide-coverage grammar seems to accompany the work. Among many aspects of Obrębski's work, a particularly interesting element is a definition of Polish relation types.

The shared tasks mentioned above prompted the creation of many high quality dependency parsing systems. Even if there are some sophisticated systems that could be used to train dependency parsers for Polish, the lack of high quality training data is a major bottleneck in the development of such parsers. To overcome this problem we address various methods of data gathering in this dissertation.

Two ways of inducing dependency structures automatically are investigated here: constituency-to-dependency conversion and cross-lingual projection of dependency information. The conversion method has been successfully applied for languages such as English, German or Bulgarian. This method presupposes that a constituency treebank for a particular language is available. Since there is a publicly available Polish constituency treebank, the conversion technique may be adapted to Polish.

The second method builds on the assumption that a linguistic analysis of a sentence largely carries over to its translation in an aligned parallel corpus. Projected annotations can then be used to train natural language processing tools for the target language. The cross-lingual projection method has been successfully applied to various levels of linguistic analysis and corresponding natural language processing tasks, such as part-of-speech tagging or semantic role labelling, as well as dependency annotation and parser induction.

The doctoral research outlined in this dissertation contributes to the development of various aspects related to dependency parsing. First, a dependency annotation schema is designed to cover primary syntactic phenomena in Polish. Second, adaptation of a constituency-to-dependency conversion method contributes to the construction of a Polish dependency treebank annotated in accordance with the dependency annotation schema. Third, we propose a weighted induction method designed to acquire dependency

structures for Polish and possibly for other resource-poor languages. The weighted induction method is the principle scientific result of our doctoral research. Fourth, we conduct some experiments that consist in training and evaluation of dependency parsers on induced treebanks. In these experiments, we use publicly available dependency parsing systems. Fifth, we make results of our research, i.e., induced dependency treebanks and trained parsing models, publicly available in order to contribute to the further development of Polish dependency parsing and to serve as the basis for other NLP tasks.