Automated Syntactic Analysis of Natural Language

Dominika ČERVEŇOVÁ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
cervenova.dominika@gmail.com

Natural language as one of the most common means of expression is also used for storing information on the Web. To work with them effectively and fast, we need to process text in a way understandable to computers. Natural language processing is, however, a difficult and problematic process, because of the informality and not very good structuring of the natural language. Syntactic analysis, as a part of the natural language processing, discovers formal relations between syntagms in a sentence and assigns them syntactic roles. That can be very helpful during later semantics acquisition and information extraction.

There are many approaches to automatic syntactic analysis and their accuracy depends on complexity of a chosen language. For example, Slavic languages are very difficult to parse, as they are flective, with free words order in sentences. There are many researches focused on parsing Slavic languages, like Czech or Russian, however parsing of Slovak language still falls behind. One of the few existing solutions for parsing Slovak, presented by Čižmár et. al [1], uses a rule-based approach. Their parser can recognize five syntactic roles in a sentence with 72-98 % accuracy. A high precision is an advantage, but it is important to recognize also relations between syntagms. To recognize whole sentence structure we need more complex approach. Machine learning appears to be useful in this domain. With enough training data – e.g. corpus of pre-annotated sentences for specific language – it is possible to train a parser to recognize syntactic structure with state-of-the-art accuracy. With this approach we can also use one parser to parse any language we have a corpus for, like Zeman et al. [2]. However, every language has its own specific features that need to be taken into account to reach the best accuracy.

We propose a hybrid method for Slovak language parsing. It consists of two classifiers, one existing parser, that uses machine learning and one additional classifier, called oracle. A schema of our method can be seen in the Figure 1.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 87–88.

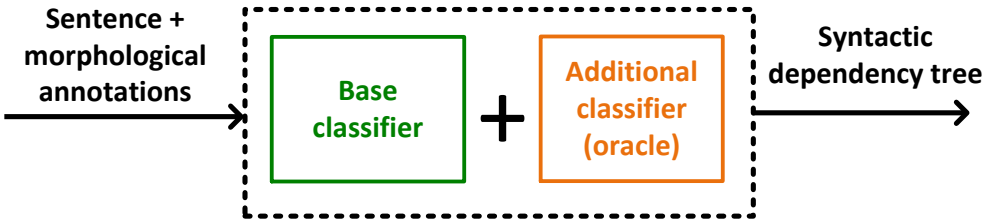

Figure 1. A schema of hybrid method for Slovak language parsing.

Both classifiers predict syntactic roles and relations based on morphological information (such as POS-tags or lemmas). The base classifier parses all sentences given. The additional classifier, oracle, is developed and trained specially for Slovak language problematic cases. By experimenting with the base parser and data from Slovak dependency corpus, we managed to find sentences, that are problematic for the base parser in Slovak (the base parser classifies their syntactic features wrong). By training our oracle on this dataset we are able to recognize critical sentence and then correct the classification from the base parser. So, in the parsing process, after every sentence goes through the base parser, our oracle looks at it and predicts whether the base classification was wrong or correct. If it was wrongly classified, the oracle does another prediction, based on decision tree model, trained on the "problematic" dataset, and changes the types of syntactic relations between tokens. This helps to improve classifications from the base parser and to achieve higher accuracy of Slovak language parsing.

To evaluate our method, we use manually annotated data of Slovak dependency corpus created at Ľudovít Štúr Institute of Linguistics. A part of this dataset will be used for training the base parser as well as our method in a form of a software prototype. Another part of the corpus data will be used as a testing set and we will compare results from our solution and the results from the base parser alone.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Čižmár, A., Juhár, J., Ondáš, S. (2010): Extracting sentence elements for the natural language understanding based on slovak national corpus. In Proc. of the Int. Conf. on Analysis of Verbal and Nonverbal Communication and Enactment, LNCS Vol. 6800, Springer, 2010 pp. 171-177.

[2] Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J. (2014): HamleDT: Harmonized Multi-Language Dependency Treebank. In: Language Resources and Evaluation, ISSN 1574-020X, vol. 48, no. 4, pp. 601–637. Springer Netherlands.