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 them more effective and faster, we need to process text in natural language in such a way, that computers could understand.

Natural language processing is, however a difficult and problematic process, because of the informality and not very good structuring of the natural language. The processing typically consists of the sequential application of different analysis components, which try to solve several problems such as phonological, syntactic, or context ambiguity, homonymy, polysemy, etc. 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 help make natural language and information stored in it more machine-processable.

Our goal is to analyze possibilities of maximizing the automation of this process and to minimize human manual work. We are working on a method that will be able to automate the syntactic text analysis process as much as possible. Currently, we focus on analyzing existing tools for various languages. There already are some parsers that can perform syntactic analysis in languages that are more simple and easier to formalize (like English, for instance), but we are also exploring options for Slavic languages (e.g. Russian, Czech or Slovak language) where automated syntagmas recognition is a nontrivial problem.

In general, there are many approaches to automated syntactic analysis. Machine learning, for example, appears to be very useful in this domain. With enough training data - e.g. corpus of annotated sentences for specific language - it is possible to train a parser to recognize syntagms with state-of-the-art accuracy. One of the greatest advantages of this approach is that we can use one parser to analyze any language we a corpus for [1]. Having enough pre-annotated data there is no need to have special linguistic skills to make parser work for any language. However, the accuracy of this

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 21–22.

type of parser varies depending on amount of specific features and rules of the language and it also depends on a quality of the annotations used for training.

Another successful approach is a rule-based parsing. This approach was also used for Slovak language by Čižmár et. al [2]. Their parser, based on rules created using *Pravidlá slovenského pravopisu* and Slovak national corpus, can recognize 98% of predicates and subjects, objects, adverbials and attributes were recognized correctly in 72 - 85% of all cases. These results are good but it is important to recognize not only syntagms alone, but also relations between them. Moreover, especially for adverbials and attributes, the recognition should be more accurate.

Creating automatic parser for Slovak language is a difficult task and there is currently no tool or approach that would provide such accuracy for Slovak as there are e.g., for English. Our aim is to create a method that will be able to recognise syntagms and relations between them automatically with as much accuracy as possible. We plan to use approach similar to machine learning, probably with some extensions. As a training data syntactic annotations made be people at the Slovak National Corpus at Ľudovít Štúr Institute of Linguistics will be used.

To create a method, applicable on a syntactic layer of a language, we have to analyse a morphological layer first. To syntagms recognition, we need to know for instance lemma of every word. Morphological information are also included in the annotations of Slovak National Corpus, however they are not always complete. As a first step we need to fill the missing values. We plan to use some of the existing tools here. Unlike syntactic analysis, the morphological has been more successful in Slovak language. Even at our faculty there has been made a research in this field before.

Before we create our own parser, we will try to use data from corpus to train existing parsers, originally made for other, but similar languages, e.g., Czech, Slovenian or Russian. This could help us to identify many problems, connected with language differences, we should be aware of, by creating our own method.

We plan to evaluate our method using a software prototype and as a golden standard a part of syntactic annotations of the Slovak National Corpus will be used.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Buchholz, S., Marsi, E. 2006, Conllx shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. association for Computational Linguistics.

[2] Čižmár, A., Juhár, J., Ondáš, S. (2010): Extracting sentence elements for the natural language understanding based on slovak national corpus. In *Proceedings of the International Conference on Analysis of Verbal and Nonverbal Communication and Enactment, COST'10*, LNCS Vol. 6800, Springer, 2010 pp. 171-177.