Learning and recognizing full syntax of sentence

András Hócza

University of Szeged, Department of Artificial Intelligence
6720 Szeged, Árpád tér 2.
hocza@inf.u-szeged.hu
http://www.inf.u-szeged.hu

Keywords: full syntax, machine learning, rule based methods

Full syntax recognition is the process of determining whether sequences of words can be grouped together as noun, adjective, verb, etc. phrases. This information is essential in machine understanding of a sentence from natural language. This means that each word of a sentence and possible word groups must be identified. Phrases often contain another phrases, therefore the phrase structure of a sentence is a tree. An additional feature of syntax a tree that it is coherent, fully cover the sentence and it has only one root which traditionally labeled by S.

Hungarian is an agglutinated language with a rich morphology and relatively free word order, whose properties add difficulties to the full analysis of the Hungarian language compared to Indo-European languages. These difficulties mean that the automatic syntax recognition of Hungarian language is too complicated to solve using experts' rules only. An efficient solution for this problem might be the application of machine learning methods, but it requires a large number of training and test examples of annotated sentences. Since the Szeged Corpus[10] became available, new methods have begun to be developed for syntactically parsing Hungarian sentences. The corpus contains texts from five different topic areas and is currently comprised of about 1.2 million word entries, 145 thousand different word forms, and an additional 225 thousand punctuation marks.

After the completion of the annotation work the Szeged Corpus was then used for training and testing machine learning algorithms to retrieve syntax recognition rules. This paper introduces an application of the RGLearn algorithm that was used to learn syntax tree patterns described by regular expressions. The tree patterns are completed with probability values using error statistics. The syntax parser uses this grammar to build up the best syntax trees of a sentence by backtracking. The results look fairly promising after comparing them to related works. This method was developed as a part of a system which extracts information from short business news texts written in the Hungarian language.

[10] The different versions of the Szeged Corpus are available at http://www.inf.u-szeged.hu/hlt.