Shallow Parsing for Information Extraction

András Hócza

Current paper presents a new approach to shallow parsing of natural language texts based on machine learning methods. Shallow Parsing (SP) is a complicated task in natural language processing: it requires sequences of words to be grouped together and to be classified. Full parsing builds complete syntactic tree of a sentence, as opposed to SP that provides identified groups of words for other natural language tasks. The linguistic information identified by SP is rich enough to support a number of large-scale natural language processing applications including information ex-traction, phrase identification in information retrieval, named entity identification , and a variety of text-mining operations. In addition, partial parsers are typically very fast when compared to full parsers.

Information Extraction (IE) is a form of text processing, which locates the relevant information in a text document. The architecture of our IE system is a pipeline with the following steps: sentence and word segmentation, morpho-syntactic analysis, part of speech tagging, SP, ontological analysis, and semantic frame recognition. The main task of each step is to provide the best information for the following step with the final aim of improving the results of semantic frame recognition.

An important question in the SP phase is how to recognize noun phrases properly for IE? We believe that our Method - recognition of complete noun phrase trees with regular expressions - is an adequate solution for this problem. The generation of NP recognition rule set contains the following steps:

- Taking complete noun phrase trees (trees under most outer NPs) from training examples
- Making rules with regular expressions by most general unification of trees
- Giving probability values for rules by evaluating training tests

The main task of the NP recognition parser is to find a most possible coverage by backtracking. Rule fitting is processed by a top-down method avoids overgeneration, because it always manages existing trees. Another important advantage is finding the boundaries of most outer NPs with good accuracy. Therefore this NP recognition parser is useful method for IE.

We tested the IE system (including NP recognition) on Hungarian business news. The result of NP recognition was between 80-90% depending on the type of texts. This result can be considered good for Hungarian considering that it is an agglutinating language with very rich morphology and relatively free word-order, which makes the full analysis of the language difficult, compared to other languages, e.g. English.