Learning and recognizing noun phrases

András Hócza, Szabolcs Iván

University Of Szeged, Department of Artificial Intelligence
hocza@inf.u-szeged.hu, ajven@programozo.hu

Keywords: noun phrase recognition, rule based methods

Abstract Learning noun phrases is a very complex problem, therefore it can be divided into several sub-tasks. In the present paper, authors try to examine the type of sub-tasks there are and the way they can be solved in order to achieve the final aim: an efficient noun phrase recognition tool. Several different approaches exist, out of which we have chosen rule based methods due to some advantages they have over other approaches (e.g. rules are easily understood not only by programmers but also by linguistic experts; rules can be extended with expert knowledge). In our work, we used two different rule based learners: the first one is the well-known *C4.5* algorithm, the second one is the so-called *RGLearn*, developed by the authors of the present paper. *RGLearn* proved to have some advantages over *C4.5*, because it is simpler to build problem specific parts into it. As a result of the learning process, the learners produced context free and context dependent rule sets. The preprocessing step is a very important part of the learning procedure, where we have to define what the learning problem exactly is. We have to make sure that the method will really learn what we want it to learn, that the given information is enough for the learner, and that the conversion creates consistent training examples without redundancy.

In this paper, we demonstrate how manually annotated sentences can be transformed into learning problems. In the first step, we dismantle noun phrase structures into elementary tree building commands. Then we generate training examples from every word position and based on current context decide whether we have to use a tree building command.

Noun phrase recognition is done by a greedy, bottom-up algorithm, that builds up the noun phrase structure of a sentence. We compared the results of the automated noun phrase recognition with manually annotated example sentences. The comparison was performed on the Szeged Corpus, which is a manually annotated textual database containing approx. 1.2 million words. The context dependent rule set was found to be the best with 90% per word accuracy, then came the context free rules with 85%, and finally expert's rules only performed 65%. Considering precision, expert rules provided rather good results (95%-100%), therefore we chose them to preprocess a number noun phrases before manual annotation in order to help the work of the annotators.

Noun phrase recognition is an important part of *Information Extraction.* The aim of our research group on the long run is to develop a modular *ToolChain* for information extraction where one of the modules will be the described noun phrase recognizer. Here we have to note that some errors may come from previous phases of automated analysis conducted by the modules of the *ToolChain,* which can cause errors in noun phrase recognition as well. At the same time, the *ToolChain* provides the possibility to solve the problem of noun phrase recognition in another way: if the parser generates the more possible noun phrase structures, the following modules of the *ToolChain* (ontology, semantic frame recognition) can select the best one by using extra information.