論文の内容の要旨

論文題目　Efficient Domain Adaptation and Detailed Error Analysis of a Deep Parser
(深い構文解析器の効率的な分野適応および詳細なエラー分析)

氏名　原　　忠　義

Improvement of the performance of a parser is crucial in diverse applications of natural language processing such as machine translation, information extraction, automatic summarization, etc. In this thesis, we propose two approaches to this problem. One is to adapt the statistical model of a parser to a specific domain, while the other is to analyze errors systematically and thus avoid time consuming trial and error cycles involved in parser improvement. These two approaches are complementary to each other, and we also discuss in the thesis how to integrate these two approaches.

Thanks to recent progresses in statistical modeling, the performances of parsers have improved significantly. However, when one applies a parser to text in real application, one notices that the performance improvement in recent years is fairly elusive. Firstly, the performances reported in papers are based on experiments using specific corpora, typically articles in newswires such as Wall Street Journal. When applied to text in different domains such as papers in the biomedical domain, the performance of a parser tends to degrade significantly. Secondly, although the common metrics for evaluation of parser performance such as the F-values of labeled or unlabeled phrase structure trees are useful for comparison of performances of different parsers, they do not give any useful clues as to how to improve the performance. In other words, there are no systematic ways of improving the performance of a parser, given text in an application domain to be analyzed.

In this thesis, we are interested in performance improvement of a deep parser based on a linguistic formalism called HPSG (Head Driven Phrase Structure Grammar). The parser produces predicate-argument structures (PAS) in the form of DAG (directed acyclic graph) as output, much more expressive than a simple tree which is produced by an ordinary shallow parser. PAS is more semantic oriented as representation and is proven to lead to improved performances in application systems. However, due to the inherent complexity of a deep grammar, performance improvement of a deep parser itself becomes more difficult and time consuming.

Both of the two approaches in this thesis use annotated corpora of given application domains. The first method is based on a statistical model which combines a newly learned model from a domain corpus with an existing model. The method is both efficient and effective. It is highly efficient in the sense that the time required for learning is 7.5 times less than the conventional methods. The second method is to systematize improvement of a parser by showing which part of a parsing model should be improved. It exploits an annotated domain corpus not only to categorize errors into linguistically meaningful types but also to capture their mutual relationships.

The effectiveness of our approaches is shown in parsing sentences in the biomedical domains.