論 文 内 容 の 要 旨

博士論文題目　Supervised Learning of Syntactic Structure

氏　名　Wide R. Hogenhout

（論文内容の要旨）

Some of the most important issues concerning automatic processing of syntax are the discovery of syntactic structure and automatic disambiguation of syntactic structure. Besides being important to almost all major applications of Natural Language Processing such as Machine Translation or Dialogue Systems, they are also related to research questions in other fields, such as how a computer system that combines syntactic analysis with higher level understanding could be realized, the question of how the human brain processes language, or the design of a system that also includes speech recognition.

In this thesis we investigate supervised learning of syntactic structure as a goal on itself, and supervised learning with the goal of disambiguating syntactic structure. In both cases we look at two approaches; a grammar based method and a method that is not based on a grammar. This results in the following four approaches.

- Parsing on the basis of a grammar is the traditional approach, although it is not easy to realize. A stochastic grammar is a technique that is often applied in this case, but at the same time it is often criticized for being fundamentally inadequate as it lacks crucial information, namely about the content of the phrases it handles. We suggest a method for extending stochastic grammars with this sort of information.

- Learning a grammar from a corpus is a very challenging problem which is still in its infant stages. Practically all approaches that have been suggested suffer from intractability—learning a full-scale grammar with them would require enormous computing power. We show how one such method can be made more efficient.

- Parsing without a grammar has been developed recently and may well prove to be faster and more robust than grammar based parsing. We investigate the problem of parsing under special conditions. We look at the problem of incremental parsing, i.e., parsing from left to right as humans do, or as one would expect in a speech recognition system.

- Learning syntactic structure without learning a grammar is possible by focusing on words rather that on sentences. This is a rather new direction and we take a first step in this direction by showing how clusters of words can be constructed on the basis of their syntactic behavior.

The connecting thread between these four and the results obtained with them reflect the unavoidable tradeoff between speed versus fundamental shortcomings of certain methods, hand labor versus machine learning, and loss of data versus sparse data problems. We have found that we can improve accuracy of a stochastic grammar, speed up the induction of a grammar, make a robust parsing system and form clusters of words reflecting syntactic usage, every time finding new trade-offs between these factors.

As investigations to both grammar based methods and non-grammar based methods are part of this thesis, the question whether the advantages of grammars outweigh their disadvantages is constantly present on the background. This is however a question we cannot answer, especially since the concept of what a grammar is and the methods of its application are constantly adapted.

All of the investigations we have done are aimed at supervised learning, which has in fact become a standard approach in this field since hand-parsed corpora became available. This shows a new relation to linguistics: instead of using a (always changing and controversial) linguistical theory, we use large samples of data annotated with noncontroversial information. This has not only to do with the learning process, it also defines our goal and as such is indispensable.

氏 名	Wide R. Hogenhout

（論文審査結果の要旨）

平成 9 年 12 月 26 日に開催した公聴会の結果を参考に平成 10 年 2 月 17 日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

Hogenhout 君は、本博士論文において、構文解析済の言語データを用いて文法や言語解析システムを自動学習するための研究を行った。本研究の成果は次のようにまとめることができる。

- 文法に基づく言語解析の分野では確率的な文法を言語データから学習する研究が行われているが、充分な成果が挙げられていなかった。本研究では、各構文解析木にその木の主辞となる単語の情報をに持たせ、その情報を含めた確率文法を定義することにより精度の高い解析システムを自動学習する方法をした。この分野における独創的な研究であり、高い解析精度が達成できることを実験により示した。

- 文法規則を言語データから学習する研究が行われているが、計算量の問題により、大規模な文法学習が困難であった。本研究では、従来の方法に比べ、同様の文法規則を遥かに短い計算時間で学習する方法を提案した。

- 一方、学習により実用的な文法を構築することは今だ困難であり、それを人手で作るにしても膨大な手間が必要である。音声認識や形態素解析で用いられている隠れマルコフモデルの手法を用いて、構文解析済みの言語データから構文構造の大部分を復元するための逐次処理システムが構築可能であることを示した。

- 言語現象の多様性のため、言語データからの学習については、常にデータの過疎性が問題になる。本研究では、統語的な振舞いによる類似性を定義することにより、統語的に似た語を自動クラスタリングする方法を提案した。この手法を用いることにより、より少量のデータからより精度の高い学習が可能となる。

以上のように、本研究は、解析済み言語データを利用することによって、言語処理に必要な文法や解析システムの構築を自動化する様々な方法を提案している。それぞれの研究は独創性も高く、しかも実際的なデータを解析し得るシステムとして実現されている。自然言語の構文解析における重要な成果を挙げていると考えられ、情報科学、言語処理の分野における学術的貢献として高く評価できる。