論文内容の要旨

博士論文題目　Japanese Dependency Structure Analysis based on a Lexicalized Statistical Model
(語彙統計モデルに基づく日本語依存構造解析)

氏名　藤尾 正和

(論文内容の要旨)

この論文では，統語構造情報付きの日本語テキストのコーパスから，品詞タグ，主辞などの表層情報の統計を用いて，特定のアルゴリズムを用いることで，精度の高い係り受け解析が可能であることを示している．

自然言語による文の統語解析の分野では，統計的アプローチが一定の成果を納めている．これまでにもさまざまな統計的構文解析手法が提案されてきているが，実用的な自然言語アプリケーションに使うには，まだ十分な精度とは言えない．より高い解析精度を得るには，モデルをより複雑にするなどして今よりも多くの情報を使用する必要がある．しかし単純に使用情報を増やしただけでは，推定するべき統計パラメータが増加し，より多くの正解コーパスが必要となる (言い換えると，データの過疎性の問題がより深刻となる)．また，学習手法によっては，組み合わせの数が多すぎてコンピュータ資源 (メモリやディスクなど) の限界からパラメータを完全には推定できないことも多い．

本論文で提案する統計モデルでは，既存の統計的構文解析モデルのように複雑なモデルを追及する代わりに，単純なモデルを使用した．これは実装が容易で解析も効率よく行えるという利点もある．代わりに，日本語の解析で統語解析を難しくする主要な要因である従属節や並列構造に着目し，それぞれを考慮したモデル及び解析アルゴリズムを提案している．

また既存の統計モデルをそのまま使ってより高い精度を実現するために，統計的部分解析 (再現率を犠牲にして適合率を上げる手法) および統計的冗長解析手法 (適合率を犠牲にして再現率を上げる手法) を提案する．

2 章で関連研究について述べた後、3 章で日本語の統計的依存構造解析モデルを提案し、大規模な構文タグ付きコーパスを用いて評価を行っている。

4 章では、提案した統計的依存構造解析を実用的な応用システムに用いる場合の使用法として、再現率を重視する場合、および、適合率を重視する場合の 2 種を考慮し、それぞれの目的に応じた使用法と評価を行っている。

5 章では、日本語の従属節の係り受けに注目し、それを重点的に学習するモデルを提案し、さらに基本モデルと融合することによって、単独のシステム以上の解析精度を達成できることを示している。

6 章では、日本語の文構造の中で特に複雑と考えられている並列句の問題を取り扱うために基本システムの拡張について新しい方法を提案し、実験によりその有効性を確認した。

このように、本論文で行った一連の研究により、日本語文の係り受け構造の解析に対して統計的な手法がどの程度活用可能であるかを示すことができた。

氏 名	藤尾 正和

（論文審査結果の要旨）

平成11年12月22日に開催した公聴会の結果を参考に平成12年2月14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。

藤尾正和は、本博士論文において、語彙の間の共起確率に基づく統計学習によって日本語文の係り受け解析を行う手法を提案し、次のような観点から解析システムの性能評価および拡張を行っている。

1. 大規模な構文解析済コーパスを利用して、語あるいは品詞間の共起確率および係り受け距離確率を推定し、それに基づく係り受け解析がどの程度の解析精度を達成できるかを実験により確認した。
2. 情報検索やより詳細な統語解析など、いくつかの利用形態における係り受け解析の能力を考察するため、適合率を重視した使用法である部分解析、および、再現率を重視した使用法である冗長解析という手法を提案し、それぞれの使用法における統計的係り受け解析システムの挙動について実験を行い、適切な使用法を提案した。
3. 日本語の文の解析において困難な問題である従属節間の係り受けについて、それらに特化した学習を事前に行うことにより、文全体の係り受け解析の精度が向上可能であることを示した。
4. 長い日本語文においてよく見られる並列構造の解析を上記で提案した統計的係り受け解析と同時に行う方法を提案し、文全体の解析精度が向上させることが可能であることを実験により検証した。

語彙間の共起確率に基づく統計的な日本語係り受け解析法を提案した本研究は、独創性高く、しかも実用的であり、自然言語処理の分野において高い貢献があると評価する。