論文内容の要旨

博士論文題目　Partial Language Analysis using Support Vector Learning
(Support Vector 学習を用いた部分言語解析)

氏名　　山田 寛康

(論文内容の要旨)

電子化された大量の文書の中から必要な情報にアクセスするために、キーワードマッチを基本とした情報検索が多く行われている。しかし、ユーザの高度な要求に対し適切な情報を検索するためには、自然言語処理技術を用いて文書処理を行う必要がある。

一般に自然言語は多くの曖昧性を含んでおり、それらすべてを解消し、解析結果を一意に決定することは難しく多大なコストを要する。このため情報検索などの応用分野では、名詞句や動詞句など、曖昧性の少ない部分的な情報を確実に解析できる部分言語解析が重要である。そこで本論文では「日本語固有表現抽出」および「英語部分構文解析」と呼ばれる2つの部分言語解析に焦点をあてて研究を行った。

固有表現抽出は、固有名詞、数量表現などに注目した名詞句同定と呼ばれる部分言語解析の一つで、人名、地名、組織名、日付表現などを自動的に抽出し分類する技術である。固有表現抽出は情報検索の前処理として使用され、とくに質問応答システムでは、ユーザの「いつ」,「だれ」,「どこ」などの質問に的確に答えるため、必須の技術とされている。

英語部分構文解析は、名詞句、動詞句などの文の構文構造を解析する自然言語処理の基礎技術の一つであり、情報検索や情報抽出だけでなく、機械翻訳など幅広い応用に使用される。

近年、これら2つの解析技術に対し、機械学習を用いた研究が多く行われている。これは、予め人手によって作成された解析済みデータから、機械学習を用いて解析規則を自動学習する方法である。部分解析を含め自然言語処理の解析規則の学習には、その手がかりとなる素性として、出現した単語や品詞などを使用する必要がある。そのため素性空間は非常に高次元な空間となり、一般に解析精度を低下させる過学習に陥る危険性が高くなる。機械学習アルゴリズム Support Vector Machine は高次元素性空間においても過学習を軽減する理論的な裏付けがあり、近年、自然言語処理の様々な解析において高い精度が報告されている。

本論文では、Support Vector Machine を使用し、2つの部分言語解析「日本語固有表現抽出」および「英語部分構文解析」のそれぞれに対し、解析規則を自動学習する手法を提案する。そして提案手法を大規模テキストコーパスに対して適用し、高い解析精度を達成した。

氏 名	山田 寛康

（論文審査結果の要旨）

平成14年1月23日に開催した公聴会の結果を参考に平成14年2月13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。

山田 寛康は、本博士論文において、文書に対する表層的な言語処理を、機械学習手法を利用して精度よく行う手法を提案し、次のような2つの重要なタスクに対して実用的な言語処理システムを実現し、解析システムの性能評価を行っている。

1. 文書に現れる人名、組織名、地名等の固有表現を抽出する手法をSupport Vectore Machineを利用して実装し、その性能について評価を行った。その結果、これまでに行われた研究と同様の学習データとテストデータを利用し、従来のどの結果をも上回る精度で固有表現抽出を行えることを示した。

2. 英語に対する部分構文解析を学習によって行う手法を提案した。学習にはSupport Vector Machineを利用し、学習時間を押さえるための様々な工夫を行った。英語の解析研究で共通に利用されるPenn Treebankを用いて学習と解析実験を行い、同様の手法に基づく過去のどの方法よりも高い解析精度を達成することができることを示した。

いずれの成果も、過去の同様の研究結果を上回る解析性能を達成しており、また、分野に依存する特殊な知識の導入を最低限にしている点からも、適用範囲の広い方法であると考えられる。現実的な文書に対する固有表現抽出と構文解析法を学習に基づいて提案した本研究は、分野に依存せすに適用可能な頑健な手法である。その独創性高く、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。

よって、本論文は博士(工学)の学位論文として価値あるものと認める。