# 論文内容の要旨

博士論文題目　Corpus-based Japanese Morphological Analysis
(コーパスに基づく日本語形態素解析)

氏名　浅原 正幸

(論文内容の要旨)

本研究の目的は、コーパスに基づく日本語の形態素解析（単語への分ち書きと品詞付与）を改善することである。ここでは、日本語形態素解析を３つの問題に分割して扱っている。１つ目は、既知の単語によるモデル、２つ目は、未知の単語を扱うモデル、そして最後は、タグ付きコーパスの管理の問題である。

最初にマルコフモデルに基づく既知語モデルの改善を行った。単純なマルコフモデルでは困難な現象について検討し、単語の品詞化、単語や品詞の position-wise grouping、撰択的 trigram といった拡張を提案して、精度の高い解析システムを実現した。次に、未知語処理では、情報抽出の手法に基づいて、オフラインの未知語抽出法を提案した。形態素解析システムの結果を文字列に分割して、文字毎の素性として記述し、文字列をまとめあげることによって未知語抽出する方法を提案した。タグ付きコーパス内の既知語をいくつかの基準によって未知語とみなすことにより学習データを生成し、それらを用いて学習したシステムの性能を評価した。高性能の形態素解析や未知語処理を行うには、学習データとなるタグ付きコーパスが整合性高いものでなれればならない。関係データベースを用いて、タグ付きコーパスと辞書を連携させて管理するシステムを提案し、両資源の同期を取りつつ誤り修正や資源の拡張を行うことのできる枠組を提案した。

これらのシステムの応用として、日本語の固有表現抽出や話し言葉データのフィラーを同定する手法を提案した。形態素情報としての素性を付与された文字列をまとめあげる手法を適用することにより、必ずしも形態素区切りと整合しない固有表現をも抽出可能な一般的な方法を実現することができた。実験により、従来提案されたどの手法よりも高性能であることを示した。話し言葉に頻出するフィラーや言い澱みは、話し言葉の解析の妨げとなる。文字列をまとめあげる手法の適用により、このような言語現象もある程度同定可能であることを示した。

| 氏 名 | 浅原 正幸 |
|---|---|

（論文審査結果の要旨）

平成15年11月4日に開催した公聴会の結果を参考に平成15年12月4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。

浅原 正幸は、本博士論文において、日本語形態素解析の性能改善を行う種々の手法を提案し、また、次に示すような様々な処理への応用および解析システムの性能評価を行っている。

1. コーパスからの学習に基づく日本語形態素解析の解析モデルの改善のため、単語の品詞化、単語や品詞の position-wise grouping、撰択的 trigram などの方法を提案した。

2. 日本語文書中に現れる未知語を自動同定する高性能な手法を提案した。

3. 日本語の固有表現や、話し言葉に現れるフィラーや言い澱みを同定する手法を提案した。

4. タグ付きコーパスと辞書を関係データベース上で連携させて管理するシステムを提案し、両者の整合性を取りつつ修正や拡張を行うことが可能なシステムを実現した。

日本語の形態素解析の性能を向上するために提案されたこれら一連の方法は、ほとんどのタスクにおいて従来手法を上回る高い性能を達成している。本研究は、独創性が高いだけでなく、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。