

論文内容の要旨

博士論文題目 日本語通時コーパスのための形態論情報アノテーションの研究

氏名 小木曾 智信

(論文内容の要旨)

近年、コーパスを用いた日本語研究が盛んになり、国立国語研究所においては、日本語史研究のための通時コーパスを構築する準備が進められている。通時コーパスには、現代語のコーパスと同様の形態論情報を付与することが必要とされているが、従来は歴史的な日本語資料に十分な精度で形態素解析を施すことができず、形態論情報のアノテーションは困難であった。

このような中、本研究は、日本語通時コーパスのための形態論情報アノテーションを実現するために自然言語処理技術を応用して、次の貢献を行った。

1. 古文の形態素解析を実現するための言語資源として、新たに古典語の辞書と学習用のコーパスを整備し、統計的機械学習にもとづく形態素解析技術を用いて、中古和文と近代文語文について実用的な精度（見出し語認定の F 値で 0.96 以上）が得られる形態素解析システムを実現した。
2. 上記の言語資源と通時コーパス自体の整備のために、辞書の見出し語とコーパスの出現形とを関連付け一貫性を保ちながら形態論情報の修正作業を行うことのできるデータベースシステム（国立国語研究所「形態論情報データベース」）を構築し、通時コーパス整備の基盤を整えた。
3. 通時コーパスに収録される多様なテキストに対して高い精度で形態論情報のアノテーションを行う方法を検討し、近世口語文・和漢混淆文・旧仮名遣いの口語文について、実際に形態論情報のアノテーションを行った。
4. 上記の形態素解析技術や形態論情報付きの通時コーパスを日本語研究者や人文科学系の研究者に使いやすい形で提供するために、新たなツールの作成・既存のツールの適用を行った。

| | |
|----|--------|
| 氏名 | 小木曾 智信 |
|----|--------|

(論文審査結果の要旨)

平成26年1月27日に開催した公聴会の結果を参考に平成26年2月21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。

小木曾 智信は、本博士論文において、日本語通時コーパスのための形態論情報アノテーションを実現するために、形態素解析用の辞書開発とそれを用いた形態素解析システムを実装し、種々の実験・考察を行った。具体的な研究成果は次のようにまとめることができる。

1. 古文の形態素解析を実現するための言語資源として、古典語の形態素辞書および学習用コーパスを整備し、統計的機械学習にもとづく形態素解析技術を用いて、中古和文と近代文語文に対して実用的な精度が得られる形態素解析システムを実現した。
2. 古文用の形態素辞書とそれに基づく通時コーパス整備のために、辞書の見出し語とコーパスの出現形とを一貫性を保ちながら形態論情報の修正作業を行うことのできるデータベースシステムを構築し、これにより、通時コーパス整備の基盤を整理した。
3. 通時コーパスの様々なテキストに対して高い精度で形態論情報の自動アノテーションを行う方法を検討し、近世口語文・和漢混淆文・旧仮名遣いの口語文について、形態論情報のアノテーションを実施した。
4. 古典語のための形態素解析技術や形態論情報付きの通時コーパスを研究者に使いやすい形で提供するために、新たなツールの作成を行った。

日本語の古文の形態素解析に関して、辞書および通時コーパスの整備と自動アノテーションのための様々なツールを作成し、研究者に利用可能な形で公開するまでにいたった本研究は、自然言語処理および日本語の通時研究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。