論文内容の要旨

博士論文題目　Multilingual Word Segmentation and Part-of-speech Tagging: A Machine Learning Approach Incorporating Diverse Features
(多言語の単語分割と品詞タグ付け：多様な素性を利用した機械学習によるアプローチ)

氏名　　中川 哲治

(論文内容の要旨)

自然言語解析システムにおいて，単語分割と品詞タグ付けは最も基本的な言語解析処理である．特に単語が分かち書きされない中国語や日本語では，多くの応用で単語分割処理が必要とされる．単語分割や品詞タグ付けにおいて，解析システムの辞書中に存在しない単語は未知語と呼ばれる．未知語に関する情報は解析システム中に存在しないため，未知語は解析失敗の大きな原因となっている．実用的な言語処理システムを実現するには，未知語の問題に対処し，高い精度で単語分割や品詞タグ付けを行う必要がある．また，統計的な言語解析処理に使用されるコーパス中には，しばしば誤った情報が付与されているため，コーパス中の誤りを自動的に検出してそれらを修正していく必要がある．

上記の課題に対して，本論文は，高い精度を持ち様々な言語に適用可能な単語分割と品詞タグ付けを機械学習を用いて実現する手法を提案している．機械学習に基づく言語解析で高い精度を得るには，できるだけ多くの有用な素性を利用することが重要であるため，様々な情報を効果的に用いることができる解析手法を検討している．具体的には，次のような手法の提案が行われている．

品詞タグ付けや未知語の品詞推定手法として従来から知られているマルコフモデルは，多様な素性を効率的に利用することが困難である．そこで，サポートベクターマシーン (SVM) を用いて多数の素性を効果的に利用することにより，高い精度で品詞タグ付けと未知語の品詞推定を行う手法を提案した．

多くの従来研究では，未知語の品詞はその前後の単語等の局所的な情報のみを用いて推定されるが，局所的な情報のみでは品詞推定が困難な場合がある．そこで，未知語の品詞間の相互作用を考慮することにより，局所的な情報だけではなく大域的な情報も用いて未知語の品詞推定を行う手法を提案した．

従来から知られている単語単位の単語分割手法は未知語の処理が困難であり，文字単位の単語分割手法は既知語に対する解析精度が低い傾向がある．そこでこれらの二つの手法を組み合わせ，単語単位の情報と文字単位の情報を同時に利用することにより，既知語に対しても未知語に対しても高い精度で単語分割を行うことができる手法を提案した．

SVM を用いて品詞タグ付けや未知語の品詞推定を行う場合，非常に多くの計算量が必要とされる．そこで，SVM のような表現力の高い学習モデルを計算量の少ない別の学習モデルと組み合わせることにより，少ない計算量で高い精度を達成する修正学習法を提案した．

タグ付きコーパス中で，機械学習による学習が困難な事例は，誤りである可能性が高い．そこで，SVMを用いてコーパス中の例外的な事例を発見し，その事例に基づいてコーパス中で不整合が生じている部分を抽出し，コーパス中の誤りを高い精度で検出する手法を提案した．