

論文の内容

博士論文題目 Building Morphology-Based Statistical Models for Japanese Language
(統計的形態素解析を用いた言語モデル)

氏 名 竹内 孔一
(論文内容の要旨)

自然言語処理において、モデル化は重要な第一段階である。英語のように単語間に空白を挿入する言語では、言語モデルとして単語を基本としたモデルを構築し、有益な手法が開発されている。しかしながら、日本語テキストのように通常わかし書きを行わない言語ではこれらの手法を直接用いることはできない。そこで、我々は日本語に対する言語モデルの構築において、形態素解析システムを適応したモデルを利用する。

従来、形態素解析システムは人手により整備されてきた。しかし、最近、電子化された大規模コーパスを利用して、学習によりパラメータを推定できる統計的な形態素解析モデルが用いられるようになった。統計的手法を導入する利点は、解析対象とする分野のコーパスを用意するだけで目標とする分野に適応できることである。

隠れマルコフモデル (Hidden Markov Model) は音声処理で主に利用されてきたが、これを用いて日本語の形態素解析システムを構築することができる。隠れマルコフモデルは大規模な学習データを利用して Forward-Backward アルゴリズムという教師無し学習を行うことができるため、大変有用である。

そこで、本論文では隠れマルコフモデルを用いた統計的形態素解析システムを利用して日本語テキストにおける2つの問題に関して研究を行った。1つは形態素解析の精度の向上に関する研究で、他はOCR処理後テキストの誤り訂正に関する研究である。形態素解析の研究では、隠れマルコフモデルを用いて、少量の品詞タグ付きコーパスと大規模テキストデータからの学習によりパラメータを推定した結果、人手で整備されている形態素解析システムの精度を上回ることを明らかにする。また、OCR誤り訂正では、大規模テキストデータを仮定した場合、統計的形態素解析システムに統計的な単語モデル(単語 trigram モデルと単語共起モデル)を融合したモデルを構築し、ランダム誤りテキストに対して高い誤り訂正能力が得られることを示す。さらに、大規模テキストデータがない場合について、OCR処理後のテキストから学習する訂正システムを構築し、システムがOCRの誤りを改善することを示す。

これらの結果から、日本語のようにわかし書きされない言語において、統計的形態素解析システムを適応したモデルが大変有効であることを示す。

氏 名	竹内 孔一
-----	-------

(論文審査結果の要旨)

平成9年12月26日に開催した公聴会の結果を参考に平成10年2月16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。

竹内孔一君は、本博士論文において、統計に基づいた日本語の形態素解析の研究、および、光学文字読み取り装置(OCR)における読み取り誤りの自動修正への適用に関する研究を行った。本研究の成果は次のようにまとめることができる。

- 大規模な未解析コーパス(生のテキストデータ)を対象に、隠れマルコフモデル(HMM)を利用することにより、日本語形態素解析の統計的パラメータの推定を行う際に、単語の分かち書きがないという日本語の問題に対応する方法を提案した。HMMによるパラメータ推定では初期確率が問題になるが、種々の分野や大きさの解析済みコーパスを用意し、初期確率を得るためのコーパスの性質と、最終的なパラメータの性能の関係を調査した。さらに、文法的な制約を導入することによる性能の変化についても実験による調査を行った。これらの調査により、数百文程度の少量の解析済みコーパスと分野に依存しない文法的制約を用いることにより、従来行われていた人手によるパラメータ設定を凌ぐ性能をもつ日本語形態素解析システムがコーパスからの学習により自動構築できることを示した。形態素解析は日本語処理の基本的な技術であり、解析精度向上の自動化の手法を提案した本研究の意義は深い。
- OCRによって電子化されたテキストには数%程度の読み取り誤りが含まれることが多く、言語解析や検索の障害となっている。OCR読み取りテキストの文字誤りを自動修正するため、統計学習によって得られた形態素解析システム、および、分野における単語の出現傾向を統合的に利用する方法を提案した。大規模な新聞記事データを用い、単語の tri-gram と形態素解析のための確率パラメータを学習することにより、読み取り誤りのうち3割以上を自動修正することが可能であることを示した。また、学習に用いるデータの分野の差異によってその性能が大きく影響されることを実験により確認し、同分野における大規模な電子化データの存在の重要性を明らかにした。

以上のように、本研究は計算機による大規模な電子化データの取り扱いに関する重要な基礎技術の発展に貢献する重要な成果を上げており、情報科学、言語処理の分野における学術的貢献として高く評価できる。