氏名	内元清貴（うちもときよたか）
学位の種類	博士（情報学）
学位記番号	論情博第50号
学位授与の日付	平成16年3月23日
学位授与の要件	学位規則第4条第2項該当
学位論文題目	Maximum Entropy Models for Japanese Text Analysis and Generation （日本語テキスト解析・生成のための最大エントロピーモデル）
論文調査委員	（主査）教授 松山隆司　教授 河原達也　助教授 佐藤理史

論文内容の要旨

本論文は，最大エントロピーモデル（MEモデル）を用いて日本語テキストの解析および生成処理を実現する手法について述べたもので，8章から構成されている。

第1章は序論で，日本語の解析および生成処理について概観し，本論文で扱う5つの処理過程の概要を示している。

第2章は，日本語処理のためのMEモデルを推定するための枠組みを提案している。この枠組みは，多種多様な情報源からの情報を統合する枠組み（MEモデリング）と，問題をMEモデリングに適した形に変形する手法（問題の変形）からなる。問題の変形により，コーパス等の言語データは学習データに変換され，MEモデリングにより，処理を実際に行う最大エントロピーモデル（MEモデル）が推定される。この枠組みを用いることにより，解析および生成に関する多様な問題を解く処理系が，比較的容易に構築できることを示している。

第3章から第7章は，この枠組みに沿って，日本語の解析および生成における5つの処理過程を高精度で実行する処理系が，言語データからの学習によって構成できることを実験的に示している。

まず第3章では，MEモデルに基づいて形態素解析を実現する方法を提案している。形態素解析においては，辞書に記述されていない未知語が出現することが避けられない。この問題を解決するための手法として，どのような文字列が形態素になるかだけではなく，どのような文字列が形態素にならないかということも合わせて学習する方法を提案し，未知語を含んだ場合でも比較的高い精度で形態素解析が実現できることを実験的に示している。

第4章では，依存構造解析のためのMEモデルを提案している。係り受け関係を，後方文脈を考慮してモデル化することにより，高い精度を持った係り受け解析が実現できることを実験的に示している。

第5章では，意味解析の一つとして，固有名詞抽出をとりあげ，それを実現するMEモデルを提案している。抽出すべき固有名詞は，しばしば複数の形態素から構成されたり，一つの形態素の一部分であったりする。こうした問題に対処するため，形態素に付加する固有名詞ラベルを工夫するとともに，MEモデルで固有名詞ラベルの推定を行った後に，書き換え規則を併用して推定精度を高める方法を提案し，良好な結果が得られることを実証している。

第6章と第7章では，文生成を目指した処理過程に対するMEモデルの適用法を提案している。

第6章では，キーワード（内容語）のリストから依存構造を生成するMEモデルを提案している。まず，キーワードから想定できる様々な依存構造を生成し，それらの適切さをMEモデルを用いて評価することにより，適切な依存構造を決定する。

第7章では，こうして得られた依存構造から，同様のgenerate-and-test法により，依存構造中に含まれる文節の順序を決定するMEモデルについて提案し，自然な語順の日本語文が生成できることを実験的に示している。

第8章は，結論であり，本研究のまとめと今後の展望を述べている。

論文審査の結果の要旨

本論文は，最大エントロピーモデル（MEモデル）を用いて日本語テキストの解析および生成処理を実現する手法について述べたもので，得られた成果は以下の通りである。

(1) 日本語テキストを処理するためのMEモデルを推定する枠組みを提案した。この枠組みは，多種多様な情報源からの情報を統合する枠組み（MEモデリング）と，問題をMEモデリングに適した形に変形する手法（問題の変形）からなる。問題の変形により，コーパス等の言語データは学習データに変換され，MEモデリングによる処理を実際に行う最大エントロピーモデル（MEモデル）が推定される。この枠組みを用いることにより，解析および生成に関する多様な問題を解く処理系が，比較的容易に構築できることを示した。

(2) 上記の枠組みが日本語テキストの解析および生成処理において有効に機能することを，5つの具体的な処理過程（形態素解析，依存構造解析，固有名詞抽出，依存構造生成，語順推定）に対して実証した。

(3) コーパス等の言語データをMEモデリングに適した形に変形する適切な方法を，上記の5つの処理過程に対して示した。形態素解析では未知語を適切に扱う方法，依存構造解析では後方文脈の考慮，固有名詞抽出では形態素単位と抽出単位の不整合の解決，生成ではgenerate-and-test法を基にしたMEモデルによる生成対象の適切さの評価法，精度向上に有効な問題変形法を具体的に示した。その結果，それぞれの処理過程において既存の方法に匹敵する高精度の処理系が学習によって構成できることを実験的に示した。

以上本論文は，確率モデルの一つである最大エントロピーモデルを用いて日本語テキストの解析および生成が実現できることを示すとともに，解析および生成の中心的な5つの処理過程を高精度で実現する処理系を，入手可能な言語資源からの学習によって構成する方法を具体的に示しており，学術上，実際上寄与するところが少なくない。よって，本論文は博士（情報学）の学位論文として価値あるものと認める。

また，平成16年1月8日に実施した論文内容とそれに関連した試問の結果合格と認めた。