

(続紙 1)

京都大学	博士 (情報学)	氏名	武田 浩一
論文題目	Building Natural Language Processing Applications Using Descriptive Models (記述的モデルを利用した自然言語処理アプリケーション構築)		
(論文内容の要旨)			
本論文は、記述的モデルを利用した自然言語処理アプリケーションについて論じたものであり、漢字複合語分割、機械翻訳、概念表現の言換え、意見表現の集約という4種類のアプリケーションを構築するために記述的モデルを導入し、その計算言語学的諸性質とアプリケーションの核となる計算手法を考察しており、全7章から構成されている。 第1章は序論であり、1980年代からの産業的なアプリケーション需要を背景にした自然言語処理技術の発展と、それらの要素技術を利用したアプリケーションの構築に記述的モデルが利用された経緯について考察したうえで、大規模な自然言語処理資源とプログラムを安定的に利用し、段階的にその精度やアプリケーションの種類を拡大する手段として記述的モデルが有用であることを明確にしている。これにより、各アプリケーションの記述的モデルの計算言語学的諸性質と、アプリケーションに必要とされる出力を生成する能力を明らかにすることが、一連の自然言語処理アプリケーションの統合を目的とする本研究の主要な研究課題として提示されている。 第2章では、統計的な漢字複合語モデルを利用した複合語分割手法を提案している。複合語分割は、情報検索、機械翻訳、テキストマイニングなどに不可欠なアプリケーションであるが、従来手法では人手作業による辞書登録が分野適応および実用的精度実現の障害となっていた。本手法ではコーパスからの教師なし複合語モデル学習により、事前の辞書登録なしに95.0%の漢字複合語分割を実現できることを示した。さらに複合語モデルを利用して、接辞の制約や特定の状態遷移を制限した学習方法の改善および少数の正解事例の付与により97.3%という実用的分割精度を達成できることを示した。 第3章では、知識ベースを利用した中間言語方式の機械翻訳手法を提案している。本手法の記述的モデルは文の概念的表現をピボットとして、原言語の解析と目的言語の生成を双方向プロセスとして定義し、副作用を持たない単一化文法のもとでこの双方向性が実際に満足されることを示した。また自然言語クラスシステムという概念表現は、代名詞や疑問詞を表現する変数、並列句に対応する集合、半関数的動詞を表現する関数といったメタクラスを含んでおり、これにより表層表現と概念表現との対応をより直接的に表現でき、文解析および文生成時に文の統語構造と概念表現との変換を容易にできることを示した。また、計算機マニュアル分野の約25,000概念および一般分野の約50,000概念からなる機械翻訳システムを実装し、計算機マニュアルの翻訳実験において実用的な精度を実現できることを示した。			

第4章では、意味的にはほぼ等価ではあるが表現上は微妙に異なる概念表現を相互に言換える手法を提案している。このために、3種類の節点および枝によって概念情報や統語情報を表現する3色非巡回グラフという記述的モデルを定義し、訳出されるべき情報、反駁不可能な情報、反駁可能かつ訳出を必ずしも要求しない情報の区別を可能とし、文生成の制御がより容易になることを示した。また3色非巡回グラフ上の節点と枝とを逐次的により制限の弱い種類に縮退させることで、グラフのサイズの高々3倍のステップで停止する概念表現の言換え操作が定義できることを示した。

第5章では、原言語と目的言語の構文規則を文脈自由文法の対によって表現した同期文法のパターンを記述的モデルとする機械翻訳手法を提案している。文脈自由文法の表現能力を超えない範囲で、訳語の対応および統語的な制約を表現しやすいように、語幹および定数個の2値素性を記述できるように同期文法を拡張し、これによってインターネットのWebページを対象としたオープンドメインでかつ構文的にも多様な言語表現を翻訳できることを示した。原言語の解析は、文脈自由文法のEarleyアルゴリズムによる効率的な処理が適用でき、さらに同一区間上に同一の非終端記号に対応する複数の解析結果が存在する場合に、より字句的なパターンを含む解析結果を優先し、また左隅固定排他規則という優先的なパターンを導入することで、より正解を上位に残した枝刈り戦略が実現できることを示した。6分野からWebページをランダムに選択し翻訳した実験では、構文的不自然さは残るものの、平均的に、正しく理解可能な翻訳結果を生成できることを示した。

第6章では、インターネット上のソーシャルメディア等の文書集合に表明される意見表現を集約する手法を提案している。従来の文書クラスタリングや論理的な無矛盾性に基づく意見表現の集約では、肯定的および否定的意見表現の混在する文書集合からの意見間の依存性を把握することが困難であることを示すとともに、個々の意見表現の出現頻度に基づいて局所的に肯定・否定形の意見表現を含む文書部分集合の包含関係が成立する場合に、このような意見表現の依存関係を階層化して可視化できることを示した。与えられた文書集合における無矛盾で極大な文書部分集合を計算する複雑さがNP完全になることと、出現頻度に基づく意見表現の包含関係は同頻度の意見表現の順列を除けば一意な系列になり、かつ包含関係の計算が多項式時間で実行できることが示せるため、本提案手法による意見集合の集約および階層関係の可視化は実用的に有望な手法であることが結論づけられている。またデジタルカメラに関する消費者のレビュー意見を対象とした実験によって、実際に意見表現の包含関係が存在しており、本手法により階層化可能であることを検証した。

第7章は結論であり、本論文を総括している。

(論文審査の結果の要旨)

本論文は、自然言語処理アプリケーションにおける記述的モデルの役割に注目し、主要なアプリケーションを実際に構築したうえでその記述的モデルの計算言語学的な諸性質と実用的な精度に関する研究をまとめたものであり、得られた主要な成果は以下のとおりである。

1. 記述的モデルを利用した複数の自然言語処理アプリケーションについて考察を行い、共有される形態素解析、構文解析、意味解析等の言語処理基盤からの各アプリケーションで必要とされる情報の生成と、記述的モデル間の対応と写像に基づいてアプリケーション統合の可能性を示した。形態素解析における漢字複合語の分割、機械翻訳のための概念表現の生成、意見表現の抽出と集約といった実際のアプリケーションの構築とその記述的モデルの提案により、これらの記述的モデルを併用する現実的な解法を明らかにした。

2. 従来手法による漢字複合語分割における最大の課題である辞書整備の問題を、実用的といえる95%の精度を実現しつつ、コーパスからの教師なし機械学習手法によって解決した。さらに比較的少数の高頻度の漢字複合語の正しい分割パターン付与と接辞や語基の制限により分割精度が97.3%まで改善できることを示した。このような分割精度の改善ができたことは、記述的モデルを採用したことによる利点の1つである。

3. 2種類の記述的モデル(概念表現を中間言語とした知識ベース型翻訳モデルおよび同期文法を利用したパターンによる機械翻訳モデル)による機械翻訳システムの設計と実装により、分野を限定した多言語翻訳からインターネット翻訳までの広い範囲で必要とされる機械翻訳技術の諸性質を明らかにし、実験により翻訳精度を検証した。特にパターンによる機械翻訳方式では、約3万件の字句的な翻訳パターンを定義することにより、商用インターネット翻訳システムとしての精度を達成できることを示した。

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22年2月24日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。