審 査 の 結 果 の 要 旨

氏　　名　　岡野原　大輔

自然言語処理の分野において、コーパスに基づいた自然言語処理システムは、機械翻訳、情報抽出、構文解析、情報検索など多くの問題で成功を収めてきた。近年特に、非常に巨大なコーパスが手に入るようになるにつれ、システムは高精度であるだけでなく、高効率、かつスケーラブルであることが求められている。本論文は、これらの要求を満たす大規模な自然言語処理システムを実現するための、効率的かつスケーラブルな手法の開発を提案している。本論文が提案するこの自然言語処理システムは、オンライン学習、文字列アルゴリズム、データ構造、疎パラメータ学習などの技術を統合することにより、高効率と高精度の両面を達成する自然言語処理システムを実現したものである。

大規模な自然言語処理の問題点は(1) サンプル数が多い、(2) 特徴種類数が多い、 (3) 候補解が多い、の3つのケースに大きく分けられる。(1) に対してはオンライン学習など様々な手法が提案されてきていることから、本論文では他の(2)、(3) の問題を中心に扱って解決を試みている。

問題(2)の例として、文書における部分文字列特徴があげられる。文書分類や文書クラスタリングにおいては、文書中に出現する任意の部分文字列の出現情報は、文書のラベルを決定するのに有効な特徴となりうるが、これらの種類数は文書長の2乗に比例し、そのまま扱うには非常に大きなコストが必要となる。本論文では異なる統計量（文書頻度等）を持つ部分文字列の種類数が高々文書長しかないことを示し、拡張接尾辞配列を用いて有効な部分文字列を漏れ無く効率的に探索する方法を提案している。本手法を文書分類と文書クラスタリングのタスクに適用することにより、この手法が既存手法を超える精度を達成することが示されており、また、本手法が100万文書の大規模文書群から有効な部分文字列を20 分で求めることが可能になるなど、スケーラビリティが高いことが示されている。

問題(2)の別の事例として、組み合わせ特徴の考察も行われている。自然言語処理では複数の基本特徴の組み合わせが有効である場合が多い。しかし、有効な組み合わせ特徴は少ないにも関わらず、組み合わせ特徴の候補数は非常に大きいため、有効な組み合わせ特徴の抽出は重要な課題であった。本論文では有効な特徴から順に最適化問題にGrafting アルゴリズムを採用し、最適解を保証しながら学習を効率的に行う方法を提案している。さらに、組合せ特徴の統計量を効率的に計算するために、単純なフィルタリングとオンラインでの統計量の計算を組み合わせたアルゴリズムを提案されている。係り受け解析に対する実験結果より、提案手法が膨大な組み合わせ特徴を効率的に処理し、既存手法と比較し同精度の結果を達成しながら非常にコンパクトなモデルが得られることが明らかにされている。

問題(3)の例としては、言語モデルの問題に取り組んでいる。言語モデルは、与えられた文が正しいかどうかの判定、または与えられた文脈から次の単語を予測するタスクであり、機械翻訳、音声認識、手書き文字認識など多くのアプリケーションで重要な役割を担っている。単語候補数は膨大であるため、従来の機械学習に基づく手法はそのまま適用できず、頻度情報に基づく単純な統計モデル（たとえばＮグラムモデル）が利用されてきた。本論文では、まず識別言語モデルを提案している。この識別言語モデルは、与えられた文に対し直接、正しい文か非文かを分類するモデルを構築するものである。このモデルの学習に必要な非文は

一般にコーパスから入手できないが、確率的言語モデルから生成された文を非文として利用することで学習を行うことが提案されている。実験結果より、このように学習して得られたモデルが、既存の言語モデルや構文解析では識別不可能な文の識別問題を75%の精度で分類できることが示されている。

これとは異なる言語モデルとして、本論文では階層型ロジスティック回帰モデルの研究についても取り組んでいる。このモデルは大量の候補がある問題を効率的に解くための学習モデルである。このモデルでは各候補（単語）が葉に対応し、内部節点のそれぞれにロジスティック回帰モデルが付随するような階層木を構築する。そして、ある候補に対する確率を、階層木の根から、その候補に対応する葉までの道上にある各節点での分類結果の積として定義する。このモデルでは任意の特徴が利用可能であることに加え、全候補を列挙する手間を省き確率が最大となる候補を効率的に求めることが提案手法により実現できることが示されている。提案手法と既存手法を比較し、本手法が高精度で単語を予測でき、かつ効率的に高確率の単語を推論できることがわかる。

以上をまとめるに、本論文は大規模自然言語処理システムに関して、スケーラブルで高精度を実現する方法論を、オンライン学習、文字列アルゴリズム、データ構造、疎パラメータ学習などでの先端技術を展開・統合することにより確立しており、コンピュータ科学分野に大きく貢献するものといえる。よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。

なお本論文の成果は共同研究によって得られたものであるが、申請者が主体的に研究を行って得られたものであることを確認している。