氏　名	颯々野　学

（論文内容の要旨）

本論文は、自然言語処理におけるマージン最大化に基づく分類器の実用的な利用法について論じたものであり、前半では、日本語の解析の効率よい手法をマージン最大化に基づく分類器での検証とともに論じ、後半では、マージン最大化に基づく分類器を用いた自然言語処理システムにおいて、効率的な正解事例の作成法を論じており、全９章から構成されている。

第１章は序論であり、機械学習手法であるマージン最大化に基づく分類器を用いる自然言語処理システムは、十分な正解事例を与えられれば高い精度が得られるが、実際の応用に適用する際には、実行時や学習時の計算コストが高いこと、及び、正解事例作成のコストが高いことを指摘し、本研究で解くべき課題を明確化している。

第２章では、日本語の係り受け解析について、時間計算量の上限を線形時間に抑えて、後戻りなく決定的に係り受け解析を行う手法を提案している。スタックを用いることで、係り受け関係の非交差の制約を守りつつ効率的に解析を行える。これをサポートベクタマシン（ＳＶＭ）と組み合わせて係り受け解析器を作成し、京大コーパスVersion 2に対して従来研究と比べて最も高い精度が得られることを示している。

第３章では、第２章のアルゴリズムを拡張し、日本語の文節認識と係り受け解析を１回の文のスキャンで同時に行うアルゴリズムを提案している。形態素単位の依存関係を文節内、文節間の２種類で表現し、従来の文節ベースの係り受け解析と互換性のある依存構造を表現している。このスキームを用いて、形態素単位で依存構造を決定することで、文節間の係り受け構造も決定できることを示している。

第４章では、日本語の三つの解析タスク、すなわち単語分割、文節認識、係り受け解析において、マージン最大化に基づく分類器であるＳＶＭと多数決パーセプトロンとの比較を行い、種々の観点で議論している。多数決パーセプトロンが精度面ではＳＶＭと同等であり、学習速度、判定速度の面ではＳＶＭを大きく上回ることを明らかにしている。

第５章では、ＳＶＭの能動学習を初めて日本語の単語分割に適用し、能動学習の効率改善手法を提案している。従来手法では、大きなラベルなしプールを用いると精度の立ち上がりが悪くなる欠点がある。これを避けるため、二つのラベルなし事例のプールを用い、大規模なクラスタリングを避けつつ、事例のサンプリングを行うプールの大きさを徐々に大きくする方法を提案している。97%の精度を得るのに必要な正解事例の数を通常の受動学習、従来の能動学習に比較して、それぞれ82.6%、40.7%削減することに成功している。

第６章では、日本語の係り受け解析において、２種類の部分的アノテーション付きコーパスの利用法を論じている。アノテートの手間がかかる品詞情報は避け、単語の区切り情報だけをアノテートした場合では、１文字、２文字の部分文字列の素性を追加して用いれば、高い精度の係り受け解析が行えることを示している。一方、隣接した文節が係る係らないかだけの情報を与え、長距離の文節間の依存情報を与えない場合では、精度は高くはないが、一定精度の係り受け解析器が構築可能であることを示している。

第７章では、文書分類において、ある文書に対して、少量の単語を追加・削除しても、その文書が属するカテゴリは変化しないとの仮定を置き、仮想的な事例を生成して、正解事例のセットに追加することで精度の向上を図る方法を提案している。この仮想事例の生成とＳＶＭとを組み合わせて、英語のニュース記事の文書分類の実験で検証を行い、正解事例が１５０文書の場合、マイクロ平均Ｆ値を１０以上改善できることを示している。

第８章では、仮想事例の代わりに、ラベルなしの実際の事例を用いる方法を提案している。ある事例から仮想事例を作り出す代わりに、大量のラベルなし事例の中から近いものを探して利用する方法である。日本語の単語分割タスクにおいて、ウェブから収集した９万４千文をラベルなし事例として用いて、一定の効果を確認している。

第９章は結論であり、本論文を総括している。

氏　名	颯々野　学

（論文審査の結果の要旨）

本論文は、自然言語処理におけるマージン最大化に基づく分類器の実用的な利用法について、特に、日本語の解析の効率よい手法と、効率的な正解事例の作成法に関する研究をまとめたもので、得られた主要な成果は以下のとおりである。

１．日本語の係り受け解析をスタックを用いて後戻りなく決定的に行なうアルゴリズムを提案し、その時間計算量の上限が理論的に線形時間で抑えられることを示し、それを実験でも確かめた。さらに、このアルゴリズムと改良された素性、サポートベクタマシン（ＳＶＭ）を組み合わせることで、京大コーパスVersion 2に対して最も高い精度が得られることを示した。このアルゴリズムを発展させ、文節認識と係り受け解析を同時に行えるアルゴリズムも提案した。また、文法情報や係り受け情報が部分的にしか与えられていない場合でも、訓練可能であることを示した。

２．ＳＶＭの能動学習を初めて日本語の単語分割に適用し、能動学習の効率改善手法を提案した。大規模なクラスタリングを避けつつ、二つのラベルなし事例のプールを用い、事例をサンプリングするプールの大きさを徐々に大きくする方法で、一定の精度を得るのに必要な正解事例の数を通常の受動学習、従来の能動学習に比較して、大幅に削減できることを示した。

３．文書分類において、ある文書に対して、少量の単語を追加・削除しても、その文書が属するカテゴリは変化しないとの仮定を置き、仮想的な事例を生成して、正解事例のセットに追加することで精度の向上を図る方法を提案した。この仮想事例の生成とＳＶＭとを組み合わせる方法を英語のニュース記事の分類に適用し、正解事例が少ない場合に、大きく精度の改善が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２０年８月２７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。