颯々野　　学（黒橋教授）

「Practical Use of Large Margin Classifiers in Natural Language Processing」（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）

平成20年9月24日授与

コンピュータとネットワークの飛躍的な広がりにより、企業内、家庭内における電子化文書の量は大幅に増えた。これらを処理する自然言語処理アプリケーション（機械翻訳、テキストマイニング、検索エンジンなど）の研究開発も活発に行われている。従来、アプリケーションで利用される言語処理の基盤技術（形態素解析や構文解析など）はルールベースで作成されていた。ルールベースのシステムは、規模がそれほど大きくなければ人間に理解しやすい、修正しやすいなどの長所がある。しかし、一定の規模を超えると、保守の困難さが大きくなり、精度の向上も難しくなる短所がある。

90年代以降、システムの出力の正解となる結果を人手で用意しておき、それを用いてシステムを訓練する方式の有効性が確かめられた。初期には比較的シンプルな確率モデルが中心であったが、近年では機械学習を用いる手法が中心となっている。中でも最も有望なのは、マージン最大化に基づく分類器（large margin classifiers）を用いるものである。ただ、これを用いても、計算コストが高いことや正解データを作るコストが大きいことなどの課題は残っている。企業内での研究開発や実用的なアプリケーションへの適用を考えると、これら課題の解決は非常に重要である。本研究では上記課題の解決を目指す。

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（構文解析）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（図１）。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。

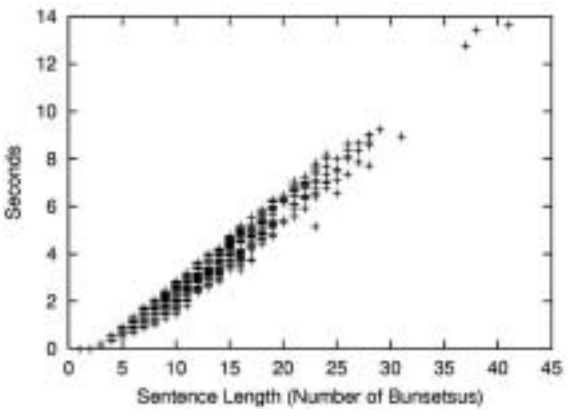

図１．文の長さと解析時間の関係

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。

今後は、これらの知見をもとに、精度が高く、よりいっそう実用的な自然言語処理システムの研究開発に取り組んでいきたい。