論文内容の要旨

博士論文題目　Machine Learning and Data Mining Approaches to
Practical Natural Language Processing
(機械学習とデータマイニングを用いた現実的な自然言語処理)

氏名　工藤 拓

(論文内容の要旨)

大量の電子テキストから, 有用な知識を抽出, 管理するためには, 高性能, 頑健, かつ高速な言語処理が必要不可欠である. 本論文では, 機械学習とマイニングアルゴリズムを用いた, 「現実的な」言語処理手法を提案する.

近年, Support Vector Machines (SVMs) を始めとする高精度な機械学習手法が提案される一方で, 要求される処理コストは増大しつつある. 高精度と高速処理は一般にトレードオフの関係にあり, 両立することは難しい. 例えば, 近年注目を浴びている Kernel 法は, 実際の応用における計算コストが問題となっている. 本論文では, まず, Kernel 法を用いた, 高性能でかつ頑健な言語解析システムを提案する. さらに, 応用範囲を広げるべく, それらの高速化手法を提案する.

機械学習を用いた言語処理における, もう 1 つの問題として, 素性の設計が挙げられる. 自動もしくは半自動の素性設計が理想的であるが, 従来は, 試行錯誤によって素性を設計することが多かった. 汎用的で, 再利用可能な言語処理を行うためには, 半自動の素性選択は必要不可欠である. 本論文では, 1) Kernel 法, 2) データマイニング, という 2 つの方法論に基づき半自動の素性選択手法について議論する.

さらに, 半自動素性選択の一例として, 半構造化テキストの分類アルゴリズムを提案する. テキストは, 種々の言語処理ツールを適用することで, 品詞や係り受け情報が付与された「半構造化テキスト」に変換できる. テキストマイニングといった高次の言語処理では, 半構造化テキストから, タスク依存の有益な情報 (素性) を部分的に選択して利用する. しかし, どのような情報が有益なのかすら分からないことが多く, 選択自身が問題となっている. 本論文では, 発見的に情報を選択するよりは, 半構造化テキストをそのままの形で処理するほうが一般的であるという考えに基づき, 半構造化テキストの構造情報を直接反映できる学習/分類アルゴリズムを提案する. 本手法は, 膨大な素性集合 (情報) から, 有益な素性 (部分情報) を自動的に選択する.

本論文の構成は, 以下の通りである. まず, 2 章にて, SVMs の概要を述べ, 3,4 章にて, SVMs を用いた, 高性能で頑健な言語処理 (Text Chunking, Dependency Parsing) を提案する. 5 章にて, それらの高速処理を提案する. 6 章にて, 半構造化テキストを対象とする学習/分類アルゴリズムを提案する.

氏 名	工藤 拓

(論文審査結果の要旨)

平成15年12月26日に開催した公聴会の結果を参考に平成16年2月17日に本博士論文の審査を行った．以下のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

工藤 拓は，本博士論文において，種々の機械学習手法とデータマイニング技術を用いた高性能の自然言語解析システムを提案し，様々な観点から解析システムの性能評価および拡張を行っている．本論文の貢献は，以下のようにまとめることができる．

1. Support Vector Machinesを利用した日本語係り受け解析に対して２つのモデルを提案し，極めて高い精度の日本語の係り受け解析システムを実現した．また，作成したソフトウェアをフリーソフトとして公開し，研究分野の発展に寄与した．

2. 汎用のチャンキングシステムを提案し，英語の基本句解析において高い性能を実現できることを示した．

3. Support Vector Machinesは，高い性能の学習を実現するが，実行速度が遅いという欠点があった．SVMによる学習結果として得られるsupport vectorの集合に共通に現れる組合せ素性をマイニングによって得，それを利用して線形カーネルのSVMを構築し直すことにより，精度を落とすことなく，大幅な実行速度の改善を実現することを可能にした．また，高速化について，転置インデックスに基づく別手法をも提案し，種々の言語解析システムを実用レベルにまで効率化することに成功した．

4. 言語処理のタスクによっては，事例を単に単語の集まりと見るだけでは高い性能を得られないことがある．言語がもつ複雑な構造を直接的に学習のための素性として扱うために，木構造マイニングとBoosting技術を組合せ，複雑な素性を人手によって追加するのではなく，自動化することに成功した．

これら機械学習に基づく種々の自然言語処理手法を提案し，実用レベルの言語処理システムを達成した本研究は、独創性が高く，しかも実用的であり，自然言語処理の分野において高い貢献があると評価する。

よって、本論文は、博士（工学）の学位論文として価値あるものと認める。