審 査 の 結 果 の 要 旨

氏　　名　　風間　淳一

風間君の論文は，自然言語処理分野における機械学習アルゴリズムとして，最大エントロピー法(ME 法)という強力かつ頑健であることから自然言語処理において広く用いられている確率モデル推定法について研究し，自然言語処理に対する ME 法の性能を向上させる２つの手法を提案している．１つ目の手法は，不等式 ME 法という ME 法の新しい拡張であり，学習データに含まれる不確実性を推定時に考慮することによりオーバーフィッティングを制御する手法である．2 つ目の手法は，HMM 状態素性手法であり，教師なし学習により学習された隠れマルコフモデル（HMM）の隠れ状態を ME 法における素性の一つとして用いることにより，データスパースネスを低減し性能の改善を行う手法である．

第一の手法の不等式 ME 法は，学習データに含まれる不確かさを考慮しながら推定を行うように通常の ME 法を拡張したものである．通常の ME 法の推定では，各素性に関して，素性のモデルによる期待値が素性の学習データでの期待値に等しいという等式制約を満たすモデルの中から，最もエントロピーの大きなモデルを選択する．素性の学習データでの期待値は限られた量の学習データから計算されるために必ず不確かさを含んでおり，そのために等式制約を完全に満たすことはオーバーフィッティングとなる．提案手法の不等式 ME 手法では，この等式制約を緩和して素性の学習データでの期待値と素性のモデルによる期待値の差がある区間・幅にあるという条件として，これを不確かさに従って大きくすることによりオーバーフィッティングを適切に制御することが可能にするものである．この不等式制約の下でエントロピー最大化を行うと，通常の ME モデルのパラメトリック形式・最適化関数を修正した新しい ME モデルが得られる．修正項の形から，不等式 ME 法は，正規化学習と解釈されるが，不等式 ME 法には制約を最大限違反している素性以外のパラメータは推定の結果ゼロになるという顕著な性質があり，素性選択が推定に組み込まれていると解釈することができる．素性選択は頑健な学習のために重要な手法であり，これを内包する不等式 ME 法は既存の手法と比べてさらに頑健であることが期待される．本研究ではこの不等式幅決定方法を提案し，正規化法との統合についても調べた．この不等式 ME 法の有効性と一般性を確認するため，文書分類と固有表現認識という二つの自然言語処理タスクを対象にして実験が行われ，その結果，両タスクにおいて，不等式 ME 法が比較手法よりも一般化能力の点で高性能であることが分かった．加えて，素性選択能力によりそのような高性能が他の手法にくらべて大幅に少ない素性数で実現されていることも明らかになった．

第二の手法の HMM 状態素性は，第一の手法が結果的には与えられた素性集合から有効な素性を選択する手法だったのに対し，大量のテキストから教師なし学習で得られた確率モデルの隠れ状態を利用して，信頼性が高くデータスパースネスの問題が少ない素性を生成するものである．本論文では，タグ付けと呼ばれる自然言語処理タスクに対し，確率モデルとして隠れマルコフモデル（HMM）を対象に，タグ付きコーパスや辞書が存在しない新しい分野やタスクに対して，タグ付け器を作成するためのコストを削減することを目的とした．その場合，開発の初期段階において少量のタグ付きコーパスしか利用できず，データスパースネスが大きい時に，より高い精度を達成できる手法が求められる．HMM に対する教師なし学習法として確立している Baum-Welch アルゴリズムを単独で適用するだけでは，そのような辞書的な情報を用いない環境下で

は，常に精度が低下することが観測された．提案手法は，独で HMM を用いるのではなく，タグ付けモデルとしては高性能な ME モデルを用い，教師なし学習された HMM を素性として用いることにより，教師なし学習が難しい状況でも精度の改善を実現するものである．素性として利用する方法は，ビタビアルゴリズムで求められた最尤の状態列を素性関数の中で参照するもので，特徴として，状態がタグであるという仮定はもはや必要なく，状態数は状況に応じて自由に設定できることがある．HMM 側から見れば，これは状態とタグの対応を ME 法により推定していると解釈できる．Baum-Welch 学習の目的はモデルの尤度を上げることであり，元々の状態とタグの同一性を保つことは保証されない．ME モデル側から見れば，HMM の状態が信頼性の高い素性として働いていると解釈できる．これは，通常 HMM の状態数が語彙数よりも大幅に小さく設定され，状態が単語に対するスムージングになっているからである．実験では，英語の品詞タグ付けと日本語の単語分割という二種類のタグ付けタスクを用いて本手法の有効性と一般性を確認した．結果として，本手法が両タスクでタグ付きコーパスが少ない時の精度を大幅に改善することが示された．また，英語の品詞タグ付けではタグ付きコーパスを最大限使用した場合に既存の最高精度を上回るような精度を達成した．

以上のように，本研究は，ME 法に基づいた自然言語処理の性能を改善するための 2 つの手法を提案し，その有効性を詳細な実験を通して示したもので，コンピュータ科学の分野、特に自然言語処理と機械学習の分野において顕著な研究成果をあげている．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．

「審査の結果の要旨」の概要

１．	課程・論文博士の別	課程博士
２．	申請者氏名（ふりがな）	風間淳一（かざま　じゅんいち）
３．	学位の種類	博士（情報理工学）
４．	学位記番号	博情　第３号
５．	学位記授与年月日	平成１６年３月２５日
６．	論文題目	Improving Maximum Entropy Natural Language Processing by　Uncertainty-aware Extensions and Unsupervised Learning（最大エントロピー法に基づく自然言語処理の不確実性拡張と教師なし学習による改良）
７．	審査委員会委員	（主査）東京大学　教　授　今井　浩 宮野　悟 中川　裕志 浅井　潔 奈良先端大学院大学　教　授　松本　裕治

８．提出ファイルの仕様等

	ファイル名	使用ｱﾌﾟﾘｹｰｼｮﾝ	OS
使用文書ファイル名	風間淳一要旨.doc	WORD	Win XP
テキストファイル名	風間淳一要旨.txt	WORD	Win XP
画像ファイル （ある場合のみ）			

最 終 試 験 の 結 果 の 要 旨

論文提出者氏名　　風間　淳一

審査委員会は、平成１６年２月２日に論文提出者に対し、学位請求論文の内容及び専攻分野に関する学識について口頭による試験を行った結果、本人は博士(情報理工学)の学位を受けるに十分な学識と研究を指導する能力を有するものと認め、合格と判定した。

論 文 の 内 容 の 要 旨

論文題目

氏　　名