論文の内容の要旨

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）

氏　　名　　風間　淳一

最大エントロピー（ME）法は，強力かつ頑健であるという特長のため，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である．しかし，ME 法には改良の余地があり，それを実現することは自然言語処理分野に対して大きな貢献となる．本論文では，自然言語処理に対するME法の性能を向上させる２つの手法を提案する．提案手法は，統計的な機械学習において一般的に問題となるデータスパースネスやオーバーフィッティングを解決するための手法であり，これが顕著に表れる自然言語処理において大きな効果をもつことが期待される手法である．1 つ目の手法は，我々が「不等式 ME 法」と呼ぶ 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，我々が「HMM 状態素性」と呼ぶ手法で，これは，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「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 法を拡張し再定式化したものである．

ME 法では，ある特徴が事象に現れているかどうかを示す関数（素性関数あるいは素性）を多数用意して事象をモデル化する．通常の ME 法の推定では，各素性に関して，

(素性 i のモデルによる期待値)＝（素性 i の学習データでの期待値）(式 1)

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．ME 法では，一般的に，素性は独立である必要はなくオーバーラップする様々な粒度の素性を矛盾なく同時に用いることができる．また，エントロピー最大化は，制約を満たすモデルの中で最も一様分布に近い，つまり，最もオーバーフィッティングをしていないモデルを選択することを意味する．これらの特長から，ME モデルは他の確率モデルと比べて頑健であり，実際に様々な自然言語処理に応用され高性能であることが示されている．

しかし，ME 法によってもデータスパースネス問題が完全に解決したわけではない．原因のひとつは，等式制約（式１）である．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，必ず不確かさを含む．そのため，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を意味し，必ずしも最適ではない．そこで，本手法では,ある程度ならこの等式制約を違反することを許す以下の不等式制約を用いる．

-Bi <= （素性 i の学習データでの期待値）-（素性 i のモデルによる期待値）<= Ai

（ただし, Ai > 0, Bi > 0）

Ai，Bi は最大の違反の大きさを決める幅であり，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なる．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，「規格化された(regularized)」学習と解釈され，これは現在 ME 法においてオーバーフィッティングを低減する手法として最もよく用いられているガウス型事前分布を用いた事後確率最大化（Gaussian MAP 推定）(Chen and Rosenfeld, 1999）と同様のアプローチである． ただし，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．また，これは，不確かさに従って幅を大きくすることでより多くの素性を削減できることを意味し，直感的にも尤もらしい素性選択手法である．本研究ではこの幅を決めるためのいくつかの方法を述べる．また，上述の Gaussian MAP 推定との組み合わせなどの拡張についても述べる．不等式制約を用いる ME 法にはこれまでいくつかの提案（Khudanpur, 1995）があったとされるが，以上述べたように具体的に実現し,また，自然言語処理において詳細に評価したのは本研究が初めてである．

本研究では，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を行った．その結果，両タスクにおいて，不等式 ME 法が比較手法の Gaussian MAP 推定など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「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手法である．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）を用いる場合について述べる．ここでの目的は，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である．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 は，品詞タグ付けなど様々なタグ付けに適用され，成功を収めてきた確率モデルである．タグ付けは文中の各記号（単語）にタグを付与していく処理であり，HMM を用いたタグ付けでは，隠れ状態がタグを表していると考え，記号列に対し最尤の状態遷移列を見つけるビタビアルゴリズムによりタグ付けを行う． HMMに対してはBaum-Welchアルゴリズムという教師なし学習法が確立されおり，これを適用し，タグ付きコーパスなしに高精度の英語品詞タグ付け器の学習に成功した例もある．教師なし学習は，それにより高精度が得られるのであれば究極のデータスパースネスの解決法となるが，Baum-Welch 学習も常に上手く働くとは限らないという報告もあり，実際，我々の設定では，HMM によるタグ付け器を Baum-Welch アルゴリズムで学習しても，常に精度が低下することが観測された．これは，Baum-Welch 学習の有効性を確認した既存の研究とは異なり，我々の設定では上述の理由から辞書的な情報を一切用いないことにしたためと考えられる．

提案手法は，このように単独で 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，教師なし学習が一見上手く働かないと思われる状況でも，精度の改善に役立てようとする手法である．素性として利用する方法は単純で， 下の例のように，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．

「もし現在位置に対する最尤の状態が X であり，その時タグ Y を付けるなら 1 を返し，それ以外なら 0 を返す」

注目すべきは，本手法においては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である．

HMM 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．そのため，Baum-Welch 学習が進むにつれて同一性が崩れ精度の低下につながる．本手法の利点は，状態とタグの関係が ME 法により柔軟に学習でき，また，尤度が改善された HMM を生かすことができることである．

逆に，ME モデルにとっては，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と考えられるからである．これは， 最尤の状態が動的に決まるという違いはあるが，単語バイグラムモデルにおける単語のクラスタリング（Brown et al., 1992）に非常に近いものである．

また，本論文では，教師なし学習された HMM を用いたが， 概念的には隠れ状態を持つ確率モデルならば HMM 以外の確率モデルでも適用可能であり，本手法は教師なし学習をデータスパースネス低減に利用する一般的なアプローチを示している．

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.

本研究では, ME 法に基づいた自然言語処理の性能を改善するための二つの手法を提案し, その有効性を示した. これらを今回扱わなかった自然言語処理にも応用し有効性をさらに検証するとともに, 手法のさらなる改良を行いたいと考えている.