審 査 の 結 果 の 要 旨

氏　名　　趙　漢哲

本論文は、自然言語処理における固有表現抽出について、教師付き学習によるアプローチを試み、その際に課題となる疎データ問題について解決を試みている。固有表現抽出は、web や学術文書などの構造化されていない文書から、人名や地名などの固有名詞、時間表現、日付表現、数量表現などの固有表現を抽出する技術である。文書を形態素に分解し、それぞれの品詞を予測する形態素解析において、固有表現は辞書に登録されていない限り未知語として認識されるため、解析における誤りの原因となる。従って、適切に固有表現を認識することは重要である。しかしながら、人間が実際に大量の文書を読んで固有表現を抽出することは、コスト面や必要な時間から困難である。そこで、計算機を用い自動的に大量の文書から固有表現を抽出できれば、自然言語処理における形態素解析やより高次の解析の精度向上に繋がり極めて有用である。計算機を用いた固有表現抽出は、ルールベースの方法や統計的手法などさまざまな方法が提案されているが、現在、主流の方法は、人間がタグ付けしたものを正解データとして含む教師付き学習の手法にほとんどは基づいている。しかしながら、この教師付き学習手法を実際に運用すると、学習データに含まれない事例が多数出現し、固有表現抽出の予測能力は低下してしまう、いわゆる疎データ問題が生じる。疎データ問題に対しては、単語そのものではなく、品詞や固有表現の種類（チャンクラベル）のような、より一般化された情報を学習に用いる方法や、大量の教師無し文書をクラスタリングし利用するなどの方法が提案されている。本論文では、前者のアイデアを発展させ、複数種類の粒度の異なる固有表現ラベルを同時に用いることで過学習を避けつつ高い予測能力を達成することに成功している。また、固有表現と共起することの多い「手がかり表現」に着目し、予測能力を向上させる方法の開発にも成功している。

本論文は五つの章からなる。第一章では、固有表現抽出について、具体例を挙げながらその基本的な概念、克服すべき困難について説明している。第二章では、固有表現抽出における三つのアプローチとして、辞書に基づく方法、ルールベースの方法に加え、近年研究が集中的に進んでいる統計的機械学習に基づく方法について説明している。第三章では、固有表現抽出の精度を向上させる新しい教師付き学習の方法を提案している。提案手法の基本的なアイデアは、単語の境界情報を表すラベル集合としてさまざまな粒

度のものを統合的に利用する点にある。粒度の細かいラベル集合を用いると、十分なデータが利用できる場合は高い予測能力を得ることができると予想されるが、データ量が不十分であれば過学習を起こし性能低下に繋がる。逆に、粒度の荒いラベル集合は、データ量が少ない場合は過学習を避け適当な予測能力を得ることができるが、情報のロスが大きくモデルは十分な柔軟性を持たず、結果予測能力を低下させる原因ともなる。本論文では、利用する複数種類のラベル情報間の一対一関係を手がかりに、複数種類のラベル情報を統合的に利用する斬新なアイデアを得、そのアイデアを条件付き確率場モデルに基づく統計的学習理論として定式化している。また、エントロピーを用いることにより辞書中の固有表現を解析し、提案モデルの精度向上に繋がる基礎データを得ている。提案手法の性能評価として、CoNLL2003固有表現抽出タスク、遺伝子名認識タスクを用い、従来法と比較し、従来法と並ぶ高い適合率を保持したまま再現率を改善することにより、高い予測能力を得ることに成功している。第四章では、固有表現と共起する手がかり表現を用いた方法について説明している。注目している単語と文法的な依存性において高頻度に共起している単語を利用し、一次の条件付き確率場を用いて固有表現抽出を行う方法を提案している。単純枚挙に従うと、膨大な数の共起情報が抽出されることになり、実用的ではない。そこで、本論文では、文法的な関連の長さは5以内、かつ、データマイニングにおける信頼度により絞り込みを行い、遺伝子名認識タスクを具体例としてその有効性を示している。第五章では、研究の概括と今後の展望を述べている。

本論文は、自然言語処理における形態素解析などの解析の基盤の一つとなる固有表現抽出において、さまざまなラベル情報の統合による新たな方法論を与え、かつ、共起する手がかり表現の利用による精度向上を示し、固有表現抽出における疎データ問題に対して現実的かつ効果的な解決策を与えるものであり、今後の研究への指針を示すという点で大きな貢献をなすものであり評価に値する。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。