論文審査の要旨及び担当者

報告番号	甲　乙　第　　号	氏　名	山田　育矢

論文審査担当者	主　査　環境情報学部教授　兼政策・メディア研究科委員　武藤佳恭
	副　査　環境情報学部教授　兼政策・メディア研究科委員　清木康
	副　査　総合政策学部教授　兼政策・メディア研究科委員　徳田英幸
	副　査　国立情報学研究所　教授　武田　英明
学力確認担当者：	

（論文審査の要旨）

山田育矢君の学位請求論文は「Entity Linking with a Knowledge Base」と題し、テキスト中のエンティティと呼ばれる固有名詞を中心とした用語を、Wikipedia等の知識ベース上の該当するエントリに結びつける手法である「エンティティリンキング」について研究したものである。

エンティティリンキングは、エンティティに関する曖昧性を排除して、コンピュータに正しいエンティティの意味を認識させる新しい方法として注目されており、2007年頃から、多数の論文が出版されている。この技術を用いると、テキスト中に出現したキーワードを該当する知識ベースのエントリに結びつけることが可能となる。例えば、「Apple」がアップルコンピュータ社を示しているのか、林檎を示しているのかをテキストの文脈を用いて判定することができるようになる。このため、コンピュータによる正確なテキストの理解に役立つと考えられている。

また、エンティティリンキングを用いることで、テキストに対して、知識ベース上にある人手で編集されたエンティティに関する高品質な情報（エンティティに関する説明文や構造データなど）を付加することができる。テキストに対して、背景となる知識を自然な形で導入することができるようになるため、上述したエンティティの曖昧性の問題の解決だけでなく、情報検索、質問応答などの様々な自然言語処理の分野の基礎的な技術として有用であると考えられている。

しかし、エンティティリンキングは比較的新しい研究分野であることもあり、最先端のシステムにおいても、その精度は十分に高いといえる状況ではない。また、エンティティリンキングを有効に活用するアプリケーションについても、十分に検討されているとはいえない。そこで、本論文では、高精度なエンティティリンキングの実現及びエンティティリンキングの実用性の評価の研究を行った。

本論文では、エンティティリンキングに関する四つの研究テーマを取り扱っている。うち

論 文 審 査 の 要 旨

二つは、高精度なエンティティリンキングを実現するための新しい手法の研究であり、(1) 分散表現と呼ばれる自然言語処理の技術を用いて、単語及びエンティティを同一のベクトル空間にマップし、これを用いてテキストの文脈をモデル化することで、エンティティリンキングを高精度に実現する手法、及び、(2) 固有表現抽出器に依存せずに、エンド・ツー・エンドで処理を行うことで、ツイッターのようなノイズが多く口語的な文書に対しても高精度にエンティティリンキングを行う手法を提案している。

さらに、エンティティリンキングを用いた二つのアプリケーションに関する研究を行った。(1) 本論文で提案したツイッター向けのエンティティリンキング手法を用いて、自然言語処理の古典的なタスクである固有表現抽出の精度を向上させる方法、及び、(2) エンティティリンキングを用いて、ウェブページ上に出現したユーザが検索したくなるようなエンティティ名を自動的にリンクに変換することで、エンティティに関する詳細な情報を、用語を選択するだけで簡単に調べられるようにする開発者向けのアプリケーションである「Linkify」を提案した。

本論文は、７章から構成されており、第１章では、本研究の動機と目的、研究対象、主な貢献のサマリーなど、本研究の概要を説明している。

第２章では、分散表現技術を用いて、単語とエンティティを同一のベクトル空間にマッピングする技術を提案し、それをエンティティリンキングの精度向上に用いる手法を提案している。本手法では、Skip-gramと呼ばれる大規模なテキスト中から単語のベクトル表現を獲得する手法を拡張し、単語とエンティティを同一ベクトル空間にマップするベクトル表現をWikipedia記事中のリンクおよびリンク周辺にある単語から高速に学習することを可能とした。そして、文書とエンティティの意味的な近さを、学習したベクトル空間上で簡単に計量する手法をあわせて提案した。具体的には、文書中の全ての名詞ベクトルの平均と、文書中の曖昧でないエンティティのベクトルの平均の二つのベクトルを用いて、文書をベクトル空間上にマップし、文書のベクトルとエンティティのベクトルのコサイン類似度を用いて、文書とエンティティの意味的な近さの計算を行う。また、エンティティリンキングを行う対象文書には、意味的に近しいエンティティが出現しやすいという仮説に基づき、この意味的な近さをエンティティリンキングの文脈として用いて、従来手法で用いられてきた特徴量と機械学習（Gradient Boosted Regression Trees）を用いてあわせることで、エンティティ

論 文 審 査 の 要 旨

リンキングのモデルを構築した。このモデルの精度を検証するため、AIDA/CoNLL、および、TAC 2010の二つのエンティティリンキングの研究において最もよく使われるデータセットを用いて評価を行った。この結果、最新のシステムのスコアを大きく更新し、既存の発表文献上での最高スコアを更新した。

第３章では、ツイッター上の文書に対して高精度にエンティティリンキングを行う手法を提案した。エンティティリンキングでは、一般に固有表現抽出器と呼ばれる外部ツールを用いて、文書中のエンティティ名の開始位置・終了位置を特定してから、それに対応するエンティティを探すという二段階の処理を行うのが一般的だった。しかし、固有表現抽出器の精度自体が、ツイッターなどの口語的でノイズの多いテキストにおいては非常に低いため、この二つのステップで処理を行う構成が、精度の大きなボトルネックとなっていた。そこで、本手法では、固有表現抽出器を用いずに、全ての可能な単語の並び（N-gram）を抽出の候補とするエンド・ツー・エンドなエンティティリンキングの手法を採用した。また、エンティティの時系列的な注目度の変化に着目し、ツイートの投稿日時におけるWikipedia上でのエンティティのページビューを機械学習の特徴量として取り入れることで、精度の向上を行った。ページビューを含むWikipedia等から取得可能な様々な特徴量を用いて、機械学習器（ランダムフォレスト）を訓練し、エンティティリンキングのモデルを作成した。提案モデルの評価を行うため、International World Wide Web Conference（WWW）というウェブ関連の研究における著名な国際会議の併設で開催されたコンペティション「NEEL Challenge」に参加した。この結果、企業や学術機関等が提案した他の手法と比較して、1.5倍以上の精度を達成し、優勝した。

第４章では、テキスト中から固有名詞等のキーワードを抽出する古典的な自然言語処理のタスクである固有表現抽出をツイッターのテキストに適応した際の精度をエンティティリンキングを用いて改善する手法を提案した。上述したように、第３章で提案したエンティティリンキングシステムにおいては、従来の手法とは異なり、固有表現抽出器を用いずにエンド・ツー・エンドで処理を行っている。この特性を利用して、固有表現抽出を行ってからエンティティリンキングを行う従来のアプローチを逆にして、エンティティリンキングを行ってから、その結果を活用して、高精度な固有表現抽出を行う方法を提案した。具体的には、エンティティリンキングと既存の固有表現抽出システム（Stanford Named Entity Recognizer）の双方の結果を特徴量とするモデルをランダムフォレストを用いて作成し、高

論 文 審 査 の 要 旨

精度に固有表現抽出を行う手法を提案した。提案手法の評価を行うため、The Annual Meeting of the Association for Computational Linguistics（ACL）という自然言語処理における著名な国際会議に併設されたコンペティション Shared Task #1 of Workshop on Noisy User-generated Textに参加した。この結果、提案したシステムは、固有表現抽出の分野において著名な学術機関等に対して大きな差（F1スコアで5-10ポイント）をつけて優勝した。

第５章では、エンティティリンキングを用いて、文書中のユーザが興味を持ちそうなエンティティ名を検出し、自動的にWikipediaやGoogle等へのリンクに変換することで、用語を選択するだけで調べることができるようにするアプリケーション「Linkify」を提案した。この研究での最も重要な貢献は、ドキュメント中のエンティティ名にユーザが興味を持つかどうかを機械学習を用いて自動的に判定するモデルを開発した点である。Wikipediaから取得したエンティティの著名度や固有度などの指標、Linked Open Data（Freebase、DBpedia）から取得したエンティティの意味クラス（例：歌手、テレビ番組）、ドキュメントと対象エンティティ名の意味的類似度などの様々な特徴量を用いて機械学習を行うことで、類似手法と比較して高精度にこのタスクを解くことが可能になることが分かった。様々な機械学習手法を評価した結果、ランダムフォレストが、該当タスクを最も高精度に解くことができることが分かった。また、クラウドソーシング（Amazon Mechanical Turk）を用いて、ドキュメントが与えられた際に、文中のエンティティ名がユーザにとって面白いかどうかをアノテーションした新しいデータセットを作成し、オンラインにて公開した。開発したアプリケーションはモバイル開発者向けのウェブサービスとして、一般公開を行った。

第６章では、先行研究のサーベイを行っている。特に上述した４つの研究分野の先行研究を中心に、エンティティリンキングに関連する論文を概観し、既存手法と比較した提案した手法の新規性・優位性を示した。また、第７章では、上述した研究のサマリーをまとめた上で、本論文の結論および今後の課題について議論を行った。

本論文では、エンティティリンキングに関する４つのデータセットで世界最高精度を達成しており、様々な自然言語処理のタスクにおいて重要となるエンティティリンキングを高精度に実現することに大きく貢献した。また、従来、実用性が知られていなかった複数の分野における応用可能性を示した。これらの成果は、著者が研究活動を行うために必要な豊かな学識を有することを示したものと言える。よって本論文の著者は博士（学術）の学位を受ける資格があるものと認める。