【30】

氏名	白川 真澄（しらかわ ますみ）
博士の専攻分野の名称	博士（情報科学）
学位記番号	第 25867 号
学位授与年月日	平成25年3月25日
学位授与の要件	学位規則第4条第1項該当 情報科学研究科マルチメディア工学専攻
学位論文名	Wikipediaを用いた汎用的な知識体系の構築に関する研究
論文審査委員	(主査) 教授 西尾 章治郎 (副査) 教授 藤原 融　教授 細田 耕　教授 薦田 憲久 教授 下條 真司　准教授 前川 卓也　准教授 原 隆浩

論文内容の要旨

汎用的な知識体系の構築は，自然文の意味をコンピュータに理解させるという大きな課題の実現において重要な役割を担っている．特に最近では，大規模な協調Web百科事典であるWikipediaから様々な知識を抽出する研究が注目を集めており，Wikipediaのエンティティを基盤とした知識体系が整備されつつある．Wikipediaでは，記事やカテゴリなどがそれぞれ識別子を持っており，知識体系を結合するためのハブとして機能する．そのため，Wikipediaをもとにした知識体系の構築は，研究者が協調して知識を蓄積できるという大きな利点を有している．しかし現時点では，エンティティの属性，エンティティ間の関係，エンティティの上位概念以外の知識についてはあまり整備されていない．そのため，Wikipediaを基盤とした知識体系の構築において，知識の種類を充実させることは重要な課題である．

そこで本研究では，既存の知識体系と連携可能な新たな知識をWikipediaから抽出することを目的とする．具体的には，既存の知識体系で整備されていない知識として，語句のトピック情報，自然文に対する関連語句，上位概念間の関係抽出を対象として知識の抽出を行う．これらの知識はそれぞれ，テキストのトピックへの分類，トピックが類似しているテキストの発見，未知の語句の意味推測などに利用できる．本論文は5章から構成され，その内容は次の通りである．まず，第1章において，序論として研究の背景について述べる．

第2章では，Wikipediaのカテゴリ構造を解析し，エンティティ（記事）がどのようなトピックに属するかという情報を抽出する．Wikipediaのカテゴリ構造は複数の親やループを持つネットワーク構造であるため，あるエンティティがどのカテゴリに属しているかという情報を親カテゴリをたどって抽出することが難しい．そこで，各カテゴリへの所属を，所属するか否かではなく，どの程度所属するかという確率として表現することで，上記の問題を解決する．確率値を算出するため，Wikipediaのカテゴリ構造を有向グラフとみなし，ランダムウォークを適用する．また，定常状態におけるランダムウォークによる確率を効率的に算出するため，べき乗法と呼ばれる数値計算手法を取り入れる．

第3章では，Wikipediaから抽出可能な様々な情報を組み合わせ，自然文から関連語句を推測する．既存手法では，自然文からの関連語句推測において解決すべき複数のタスク（キーフレーズ抽出，単一語句からの関連語句推測，関連語句の集約など）に対し，単純な加算によるスコアリングによってタスクを組み合わせているが，入力テキストに含まれるノイズに弱いという問題がある．そこで，ベイズ理論に基づく確率的なスコアを導入し，また，確率的な入力に対して適用可能な拡張ナイーブベイズを提案する．これにより，入力テキストに対してロバスト性の高い関連語句推測を実現する．

第4章では，Wikipediaの知識をもとに，大規模なテキストデータから上位概念間の関係を抽出する．関係抽出に関する研究では一般的にエンティティ間の事実関係を網羅的に抽出することを目的としているが，提案手法では，未知の事物に対する推測を行うため，汎化した上位概念レベルで関係を抽出する．上位概念間の関係抽出は，Wikipediaのエンティティに対して上位下位関係を定義した既存研究の成果を利用し，テキストから関係を抽出する際に語句を上位概念に置き換えることで実現する．語句から上位概念への変換においては，語句の曖昧性の問題が発生するが，Wikipediaのエンティティを介して行うことで高い精度を達成する．

最後に第5章では，本論文の成果を要約したのち，今後の研究課題について述べ，本論文のまとめとする．

論文審査の結果の要旨

汎用的な知識体系の構築は、自然文の意味をコンピュータに理解させるという大きな課題の実現において重要な役割を担っている。特に最近では、大規模な協調Web百科事典であるWikipediaから様々な知識を抽出する研究が注目を集めており、Wikipediaのエンティティを基盤とした知識体系が整備されつつある。しかし現時点では、エンティティの属性、エンティティ間の関係、エンティティの上位概念といった基本的な知識を定義しているのみであり、Wikipediaを基盤とした知識体系の構築において、知識の種類を充実させることは重要な課題である。この課題に対し、本論文では、既存の知識体系にはない新しい知識として、語句のトピック情報、自然文に対する関連語句、上位概念間の関係をそれぞれWikipediaから抽出するための手法を提案している。本論文の主要な研究成果を要約すると次の通りである。

(1) Wikipediaのカテゴリ構造をグラフとみなして解析することにより、エンティティがどのようなトピックに属するかという情報を確率値として抽出する手法を提案している。この手法ではWikipediaの任意のカテゴリをトピックとして選択するだけで、教師データの作成を必要とせずに語句のトピック情報を抽出できる。
(2) Wikipediaの記事や記事間リンクの情報から、特徴的な語句や語句の曖昧性解消などの確率値を計算し、ナイーブベイズを拡張した手法によりそれらの確率値を統合することで、自然文に対する関連語句を推測する手法を提案している。この手法はベイズ理論に基づいており、既存のヒューリスティックな手法と比べて高い

精度でテキストから関連語句を推測できる。

(3) 大規模なテキストデータから語句間の関係を抽出し、Wikipediaの記事を介して語句を上位概念に置き換えることにより、上位概念間の関係を抽出する手法を提案している。この手法では人が上位概念間の関係を学習する方法を模倣することで、精度の高い上位概念間の関係抽出を実現している。

以上のように、本論文はWikipediaを基盤とした汎用的な知識体系の構築における先駆的な研究として、情報科学に寄与するところが大きい。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。