

氏 名	との いけ まさ つぐ 外 池 昌 嗣
学位(専攻分野)	博 士 (情 報 学)
学 位 記 番 号	情 博 第 244 号
学位授与の日付	平 成 19 年 3 月 23 日
学位授与の要件	学 位 規 則 第 4 条 第 1 項 該 当
研 究 科 ・ 専 攻	情 報 学 研 究 科 知 能 情 報 学 専 攻
学 位 論 文 題 目	Natural Language Processing Exploiting Topics in the Web Text Archive (ウェブテキストアーカイブにおける文書のトピックを利用した自然言語処理)
論文調査委員	(主 査) 教 授 黒 橋 禎 夫 教 授 石 田 亨 教 授 河 原 達 也

論 文 内 容 の 要 旨

本論文は、ウェブ文書のトピックを利用した自然言語処理について論じたものであり、専門用語対訳集の自動生成のタスクにおいては、用語のトピックを判定する方法、及び、トピックとの整合性に基づき訳語候補の正誤を判定する方法を論じており、質問応答における解選択のタスクにおいては、ウェブを用いて質問文のトピックから最も強く連想される解を選択する手法を論じており、全5章から構成されている。

第1章は序論であり、ウェブは、大規模かつ、様々な言語の文書を含む言語資源・知識源であり、これまでの自然言語処理研究で広く利用されてきたことを指摘し、これらの研究を総括している。そして、ウェブ文書のトピックを利用した事例について分析した上で、専門用語対訳集自動生成及び、質問応答における解選択のタスクにおいてウェブを言語資源、知識源として用いる場合のトピックの扱いについて議論している。専門用語対訳集の自動生成のタスクにおいては、タスク全体を、対象トピックの文書の収集、対象トピックの用語の抽出、用語の訳語推定の3つに分解し、それぞれに対して、トピックを考慮した操作が必要であることを指摘し、本研究で解くべき課題を明確化している。質問応答における解選択においては、ウェブを利用して、質問文のトピックから最も強く連想される答えを選ぶ方法を提案している。

第2章では、与えられた用語が対象トピックに属するか否かを、ウェブを利用して判定する手法を提案している。従来、用語のトピック判定を行うには、対象トピックのコーパスを準備する必要があるため、このため、対象トピックのコーパスが準備できない場合に適用することができなかった。また、用語が文書に出現する頻度を利用していたため、安定した性能でトピック判定をすることが困難であった。提案手法では、ウェブを利用することにより、多様なトピックに適用可能とし、また、用語のトピック判定の課題を、用語を含む文書のトピック判定の問題に置き換えることにより、安定した性能を達成した。そして、対象トピックの新しい用語を発見する実験により、その有効性を示している。

第3章では、訳語のわからない専門用語の訳語を推定する手法を提案している。予備調査の結果、約88%の専門用語が構成的であることから、訳語推定対象用語の構成要素の訳語を対訳辞書で調べ、これらを結合して訳語候補を生成する方法を採用している。従来手法では、ウェブ全体を利用して訳語候補の検証をしているが、この方法では、トピックによって訳語が違う場合、そのトピックで使われている訳語を正しく選択することができない。これに対して、提案手法は、ウェブからあらかじめ収集した対象トピックの文書を用いて訳語候補の検証を行うことにより、そのトピックで使われる訳語候補を正しく選ぶことを可能にしている。本論文では、両者のウェブ利用法の比較に焦点をあて、実験の結果、ウェブ全体を用いるとカバレッジが高いが、精度は低い反面、対象トピックのコーパスを利用すれば、カバレッジは低下するが、高い精度を達成できるという得失を明らかにしている。

第4章では、質問応答のタスクにおいて、質問文のトピックと解候補の連想の強さを利用して、解選択を行う手法を提案している。従来のウェブを用いた解選択手法においては、質問文の解析、及び、質問文との照合に計算コストがかかるため、ウェブ全体の一部の文書しか利用できなかった。これに対して、提案手法は、質問文のトピックと解候補の連想の強さを測

定するため、計算コストがかからず、ウェブ全体の知識を用いることができる。実際の手法としては、あらかじめ決めておいた語の重要度に基づきトピックを同定し、答えの選択を行う手法と、トピックと解候補の連想の強さを利用してトピック同定と答えの選択を同時に行う方法を提案している。そして、4択クイズを対象とした実験により、質問文のトピックと解候補の連想の強さを利用して、どの程度質問に答えることができるのかを明らかにしている。

第5章は結論であり、本論文を総括している。

論文審査の結果の要旨

本論文は、ウェブ文書のトピックを利用した専門用語対訳集の自動生成と、質問応答における質問文のトピックと選択肢の間の連想を利用した解選択に関する研究をまとめたもので、得られた主要な成果は以下のとおりである。

1. トピック判定対象の用語と、対象トピックが与えられたとき、その用語が対象トピックに属するか否かを、対象トピックの既知用語を利用してウェブから収集したコーパスと、対象用語を含むウェブ文書の間の類似性を評価することによって判定する手法を提案した。そして、既存の専門用語辞書に含まれる用語のトピック判定の実験により、高い性能でトピック判定ができることを示し、さらに、対象トピックの新しい用語を発見する実験により、実際に対象トピックの未知用語を発見し、専門用語集を自動生成することが可能であることを示した。

2. 訳を知りたい用語の構成要素の訳語を既存の対訳辞書で調べ、構成要素の訳語を組み合わせることで訳語候補を生成する手法において、訳語候補の検証時に、ウェブ全体ではなく、ウェブから収集した専門分野コーパスを用いる方法を提案した。そして、ウェブ全体を用いるとカバレッジが高いが、精度は低い反面、対象トピックのコーパスを利用すれば、カバレッジは低下するが、高い精度を達成できるという得失を実験によって示した。

3. 質問応答のタスクにおいて、質問文のトピックを同定する方法と、質問文のトピックと解候補の連想の強さをサーチエンジンのヒット数から計算できる2つの尺度を利用して測定することにより、解選択を行う手法を提案した。そして、この連想を用いた解選択手法を利用することにより、実際にどの程度の質問に答えることができるのかを明らかにした。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。

また、平成19年2月6日実施した論文内容とそれに関連した試問の結果合格と認めた。