

(続紙 1)

京都大学	博士 (情報学)	氏名	泉 朋子
論文題目	Normalization and Similarity Recognition of Complex Predicate Phrases Based on Linguistically-Motivated Evidence (題目和訳：言語学的特徴を用いた述部の正規化と同義性判定)		
(論文内容の要旨)			
<p>本論文は、自然言語文の述部に焦点を当て、同じ出来事を表す述部を計算機で認識させるための述部の正規化及び同義性判定手法について論じたものである。述部は文の核となる情報を表しており、述部の意味を正確にとらえることは自然言語処理アプリケーションにおいて必須のタスクである。たとえば、「使いたい」「接続できない」「わからない」などの述部表現は要望・苦情・疑問を表しており、意見マイニングや、質問応答/対話システムにおいて重要な情報を表している。同様に、「メモリを消費する」と「メモリを食う」などの異なる述部表現間同義性を自動で認識できれば、検索エンジンなど情報抽出技術の精度向上が期待できる。しかし、計算機で述部の意味を正確にとらえることは自然言語表現の多様性ゆえに簡単ではない。例えば、「解約したい」という表現も「解約を行って欲しいのですが」「キャンセルしたいんだけど」など様々な形で表現される。</p> <p>本論文の貢献は、これらの多様な述部表現を、言語学的な分析をもとに、同じ出来事を表す単純な表現に正規化し、さらに文脈を考慮して同義となる述部を自動で認識する高精度な同義判定技術を提案したことである。</p> <p>第1章は序論であり、述部に関する言語学的分析及び、実際の自然言語処理アプリケーションにおいて問題となる述部の多様性の原因について分析を行っている。述部の多様性には、形態素レベルでの多様性、構造レベルでの多様性、意味レベルでの多様性という3つの要因が関連していると論じている。</p> <p>第2章では、形態素レベルでの多様性に焦点をあて、日本語の文末表現の正規化を提案している。同じ出来事を表す最も単純な表現に正規化することを目的とし、出来事の意味に影響を与える文末表現を、形式意味論の時制、モダリティ、否定という3つの軸をもとに整理し、これらに属す表現のみを最低限残す正規化ルールを構築した。この正規化ルールにより、多様な文末表現の異なりを大幅に削減することができ、テキストマイニングアプリケーションを想定した偏り述部抽出タスク(特定の文書集合に特徴的に現れる述部表現の獲得)において提案手法の有用性を確認した。</p> <p>第3章では、「構造レベルでの多様性」に焦点をあて、「解約を行う」「納得が行かない」などの機能動詞構造の正規化を提案している。日本語は他の言語に比べ機能動詞の数が多いため、機能動詞が誤って述部として抽出される問題が頻発している。例えば、「納得が行かない」という機能動詞構造の場合、本来は「納得できない」と抽出されるべきであるが、単純な処理では「行かない」が誤って述部として抽出されてしまう。第3章では、これらの機能動詞構造の簡略化に焦点を当て、機能動詞の文法機能のみを保持しつつ、単純な述部に言い換えるための正規化パターンを構築した。さらに、機能動詞構造と本動詞構造の曖昧性解消に関して、大規模な新聞・ブログコーパスを用いて曖昧性解消辞書を構築した。この言い換えパターンと曖昧性解消</p>			

辞書を用いることで、約9割の精度で機能動詞構造を単純な述部に正規化することができ、より正確に述部の意味を抽出することが可能となった。

第4章では、意味レベルでの多様性に焦点を当て、「メモリを消費」と「メモリを食う」のような文脈によって同義になり得る述部も対象として、述部の同義性判定手法を提案している。同義の述部の特徴を様々な言語学的視点で分析し、辞書定義文、用言属性、分布類似度、機能表現などの言語情報を同義判定の素性として用いた。さらに、今まで言語処理では困難であった反義関係を表す述部と同義関係を表す述部の識別を、反義関係に特化した言語的特徴を用いることで大幅に改善し、既存手法に比べて高精度な述部同義性判定を実現した。また、提案手法を用いて同義の述部をブログデータから自動で抽出した結果、高精度に同義述部を獲得することが確認でき、多様な述部表現に対する言語資源の自動獲得の可能性を示した。

第5章は結論であり、本論文の貢献をまとめている。

(続紙 2)

(論文審査の結果の要旨)

本論文は、これまで計算機での意味解析が困難とされていた述部に焦点を当て、多様な述部表現の同義性を自動で認識するための正規化技術と同義性判定技術を提案している。本論文において得られた主要な成果は以下の通りである。

1. 従来のテキストマイニング技術や情報抽出技術では、述部の主辞 (head) を用いた分析が中心であったため、「～できる」「～したい」「～かもしれない」といった可能・要望・推量表現など述部の文末表現によって表現される意味を正しく認識することができなかった。本論文では述部表現を時制、モダリティ、否定という3つの軸をもとに正規化することにより、計算機で正しく述部の意味を抽出・集計することを可能とした。また、テキストマイニングアプリケーションを想定した偏り述部抽出タスクにおいて、この正規化技術によって従来手法では抽出できなかった述部表現が発見できることを確認した。この技術はすでに実サービスに導入されている。
2. テキストマイニングや翻訳技術など多数の言語処理タスクにおいて、「変更を行う」や「納得が行かない」などの複雑な述部構造である機能動詞構造は、述部の誤抽出や誤翻訳の大きな原因となっている。本論文では、150種類の機能動詞を対象とし、言語学的な分析をもとに、網羅的に機能動詞構造を単純な述部に正規化するパターンを構築した。さらに、組み合わせる名詞によって機能動詞と本動詞の曖昧性を持つ構造に関しては、大規模な新聞・ブログコーパスから曖昧性解消辞書を構築することで、高精度に機能動詞構造を単純化させることが可能となった。本技術は、マイニング技術のみならず機械翻訳の前処理としても有用なものと考えられる。
3. 従来の述部の同義性判定・同義語獲得技術では、言語処理リソースのカバレッジや自動獲得された同義語の精度に問題があった。本論文では、述部の同義性判定において有益な特徴を言語学的分析をもとに抽出し、高精度に述部の同義性を判定することを可能にした。特に、従来技術では困難とされていた同義述部と反義述部の識別に関して、反義述部の特徴を取り入れることにより、同義述部の判定を高精度化した。本論文における同義性判定手法が達成した精度87%は実サービスにも十分耐えうるものであり、提案手法を用いた同義述部獲得実験でも同義語辞書の自動構築を示唆する性能を示した。

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25年12月25日に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。