(続紙 1)

京都大学	博士 (情報学)	氏名	萩行 正嗣
論文題目	Studies on Annotated Diverse Corpus Construction and Zero Reference Resolution in Japanese (日本語の多様な文書からなるタグ付きコーパスの構築及びゼロ照応解析に関する研究)		
<p>(論文内容の要旨)</p> <p>本論文は、日本語のゼロ照応解析について、多様なWeb文書へのタグ付与を通じた問題の明確化と、それに基づく新たな解析モデルの研究を行い、その成果をまとめたものである。本論文では従来のゼロ照応解析研究の主な対象であった新聞記事とWeb文書の違いとして文書の著者および読者の出現に注目している。これらの出現を文書中の著者・読者表現と外界ゼロ照応に分類し、Web文書に対するこれらのタグ付けの際の問題点の整理および著者・読者の談話への出現を考慮したゼロ照応解析システムの提案を行なっている。論文は、Webコーパスの構築、著者・読者表現の自動推定、ゼロ照応解析システムの提案に各1章を割り当て、序論と結論をあわせて、全5章で構成されている。</p> <p>第1章は序論であり、まず日本語ゼロ照応解析の概略について説明し、次に新聞記事とWeb文書の違いについて論じている。Web文書は多様な著者から多様な読者に向けて発信されており、話題も直接著者や読者に関連することが多い。著者・読者の談話への出現を、それらを明示的に指し示す表現(著者・読者表現)と外界ゼロ照応に分類し、ゼロ照応解析の観点からそれぞれの性質について論じている。また、自然言語処理におけるコーパス構築およびゼロ照応解析の歴史および従来研究で取られてきたアプローチについて概略をまとめている。</p> <p>第2章は多様なWeb文書からなるタグ付きコーパスの構築について述べている。タグ付きコーパスとは文書に人手で様々な言語情報を付与したものである。本研究で作成したコーパスは形態素情報、構文情報などに加えてゼロ照応関係を付与したものである。本章では、まず、Webを利用して多様な文書を集め、効率よくタグ付け作業を行なう手法を提案している。そして、Webの文書をタグ付けする上での2つの問題を明らかにし、タグ付け基準を整理している。1つ目の問題は著者・読者表現のタグ付けであり、これは従来のコーパスではタグ付けされてこなかったものである。本研究では著者・読者表現のWebでの登場パターンについて分析し、タグ付け基準を設定している。2つ目の問題は著者・読者が談話に登場することで述語項構造のタグ付けに曖昧性が生じる表現の存在であり、これらの表現について分類を行ないタグ付け基準を設定している。最後に、構築されたコーパス(DDLC: Diverse Document Leads Corpus)の統計およびその性質について論じている。DDLCで使用されているモデリティや敬語表現の傾向は新聞記事コーパスとは大きく異なり、多様なタイプの文書がDDLCに含まれている。また、多様な表現が著者・読者表現として用いられており、著者・読者に関するゼロ照応関係の割合が高い。作業員間のタグ付けの一致度について調査を行なったところ、作業員間でタグ付けが一致しなかったものの多くが作業員の見落としによるものであったため、実際のコーパス作成時には必ず一名以上の作業員による確認を行うこととしている。</p> <p>第3章では著者・読者表現の自動推定について述べている。著者・読者表現はゼロ照応解析において大きな手掛りとなるが、その情報をゼロ照応解析において利用するためには前処理として著者・読者表現を自動推定する必要がある。本論文では、著者・読者表現推定をランキング学習による機械学習によって扱う。著者・読者表現候補の周辺文脈を汎化した語彙統語パターンを定義し、機械学習の素性として用いた。第2章で構築したDDLCを用いて著者・読者表現推定の実験を行ない、高い精度で著者・読者表現を推定できることを示した。一方、再現率は5割程度にとどまっており、こ</p>			

これは、全体として著者・読者表現が出現しない文書の方が多く、モデルが「著者・読者表現なし」と推定しやすく学習されたことが原因であった。

第4章では外界ゼロ照応および著者・読者表現を考慮したゼロ照応解析モデルについて述べている。従来のゼロ照応解析モデル(ベースラインモデル)では、ゼロ代名詞の照応先として、文章中の談話要素のみを考慮していたが、提案手法では、外界ゼロ照応に対応する仮想的な談話要素を導入することにより外界ゼロ照応を扱っている。また、機械学習における素性表現を工夫することで、著者・読者表現に対応する談話要素は通常の談話要素と仮想的な著者・読者の両方の性質を持つものとして表現している。これにより、著者・読者と敬語表現などとの関係の学習を実現している。実験では、DDLCを用いてベースラインと提案手法の比較を行なった。提案手法では、著者・読者表現として第3章の手法による自動推定結果を用いる方法と、コーパスに人手で付与された著者・読者表現を用いる方法を評価した。実験結果では、自動推定の結果を用いた場合でも、文章内ゼロ照応およびゼロ照応全体において提案手法がベースラインの精度を上回った。一方、正しい著者・読者表現を用いた場合には、精度が大きく向上しており、著者・読者表現推定の精度を高めることで、ゼロ照応解析の精度がさらに向上する可能性があることを示した。

第5章は結論であり、本論文を総括し、今後の課題について述べている。

(続紙 2)

(論文審査の結果の要旨)

本論文は、日本語のゼロ照応解析について、多様なWeb文書へのタグ付与を通じた問題の明確化と、それに基づく新たな解析モデルの研究を行い、その成果をまとめたものである。得られた主要な成果は以下の通りである。

1. 従来のゼロ照応関係タグ付きコーパスは主に新聞記事コーパスを対象に作成されていた。本研究では、Webから収集した多様な文書を対象にしてゼロ照応関係タグ付きコーパスを構築した。Web文書は多様な著者から多様な読者に向けて記述されており、著者・読者に直接関連する話題が多く含まれるという特徴を持つ。そのため、文書の解釈に著者・読者が大きな影響を持ち、これがゼロ照応関係においては特に顕著である。本研究では、ゼロ照応関係のタグ付けと著者・読者の関係について分析し、タグ付け基準の設計を行なった。また、作成されたコーパスを分析することで、新聞記事とWeb文書の文体や話題の傾向の違いについて考察した。作成されたコーパスは今後研究者コミュニティ向けに公開される予定であり、今後のゼロ照応解析研究の発展に資すると考えられる。

2. 著者・読者を指す表現はゼロ照応解析において重要な手掛りとなる。一方、Web文書では様々な著者・読者が存在し、人称代名詞だけでなく固有表現など様々な表現が著者・読者表現として利用される。そのため、これらを単語の表層表現だけから同定することは困難である。本論文では、著者・読者表現推定をランキング問題として捉え、候補表現の周辺の語彙・統語パターンを手掛りとして著者・読者表現推定を行なった。作成したWebコーパスを用いた実験の結果、高い精度で著者・読者表現が推定できることを示した。

3. 従来のゼロ照応解析では、照応先が文章中に出現しない外界ゼロ照応にはほとんど注意がはらわれてこなかった。外界ゼロ照応には照応先が著者・読者となるものも多く、著者・読者表現と密接な関係にある。本論文では、外界ゼロ照応および著者・読者表現を考慮したゼロ照応解析モデルを提案している。本モデルでは著者・読者を表す仮想的な談話要素を導入し、機械学習による述語項構造解析としてゼロ照応解析を行う。作成したWebコーパスを用いた実験の結果、提案手法は既存モデルよりも高い精度でゼロ照応解析が行えることを示した。

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26年2月19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。