萩行 正嗣（黒橋禎夫教授）

「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）

平成 26 年 3 月 24 日授与

近年、Web の広まりとともに情報検索や機械翻訳などの自然言語処理アプリケーションが広く利用されるようになってきたが，その精度向上には基礎的な自然言語処理解析の精度向上が不可欠である。その基礎的な自然言語処理解析の中の一つにゼロ照応解析がある。ゼロ照応とは用言の項が省略される現象である。例えば「(Φ1 ガ)パスタが好きで、(Φ2 ガ)(Φ3 ヲ)毎日食べています。」という文では、「誰が」パスタが好きで、「誰が」「何を」毎日食べているのか、という情報は明示的には書かれていない。しかし、実際には、Φ1 および Φ2 はこの文の著者であり、Φ3 は前方で言及された「パスタ」である。ゼロ照応解析は、このような省略された項を推定する処理である。特に日本語においては、ゼロ照応が頻出するため、ゼロ照応解析は日本語の意味解析において必須的な技術の一つであると言える。

従来のゼロ照応解析の研究は主に新聞記事を対象としてきたが、本論文では Web 文書をその対象としている。本論文では、これらの違いとして、文書の著者・読者に着目している。新聞記事では、著者は新聞記者に、読者は購読者に限られており、事件などを報道すること目的のため、著者や読者が話題中に登場することはほとんどない。一方、Web 文書は様々な著者によって様々な読者に向けて書かれており、その内容も blog や通販サイトなどを中心に、著者や読者が話題中に出現することが多い。著者や読者は、省略されやすい、敬語表現などがゼロ照応解析の手掛かりになるといった特徴を持つため、ゼロ照応解析において著者や読者を特別に扱うことは重要である。著者や読者は、上述の例における文書の著者のように明示的に言及する表現が存在しない、外界ゼロ照応として出現する。従来のゼロ照応解析研究の多くは外界ゼロ照応を扱ってこなかったが、著者・読者を扱うには、この外界ゼロ照応を扱うことは必要不可欠であると言える。

本論文では、まず、Web 文書からのタグ付きコーパスの構築について論じている。タグ付けにあたり、著者・読者に関するタグ付けの問題を分析している。一つ目の問題は「私」「我々」「あなた」などの文書の著者や読者について言及する表現の存在である。本論文では、そのような表現を著者・読者表現として定義し、タグ付け基準を整理した。二つ目の問題は、著者・読者が談話構造に出現することで、著者、読者および不特定の人のいずれか複数とも解釈できるようになる項である。本論文では、このような表現を分析し、タグ付けの基準を設定した。そして、1,000 文書からなら Web コーパスを構築し、ゼロ照応関係を含む、様々な意味関係の付与を行った。

そして、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の提案している。提案手法では、初めに語彙統語パターンを利用して著者・読者表現の自動推定を行う。その後、述語項構造解析の一部として、ゼロ照応解析を行う。この際、外界ゼロ照応に対応する仮想的な談話要素を考慮することで、外界ゼロ照応を扱っている。提案モデルでは、外界ゼロ照応や著者・読者表現を扱わない手法に比べて高い精度を達成している。

ゼロ照応解析は、表層には現れていない現象を扱うという点において、意味解析の一種といえる。今後は、さらに意味解析の技術が発展し、計算機による言語理解が一層進むことが期待される。

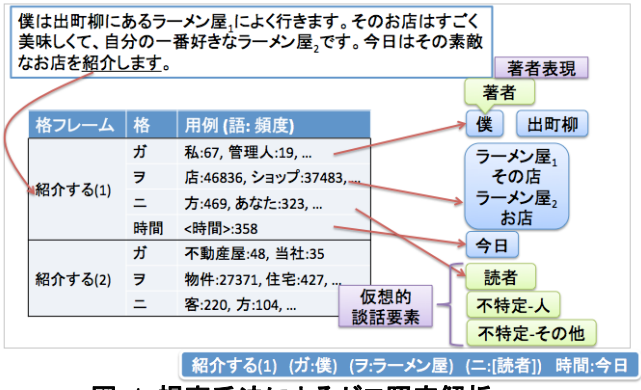

図 1 提案手法によるゼロ照応解析