ISS-P-51

中心化理論の支援のための辞書構築

佐藤 駿† 三浦 孝夫†
† 法政大学理工学部創生科学科

1. 前書き[1]

照応解析とは代名詞や指示詞，ゼロ代名詞を補完する処理のことである．照応解析の利点は，ゼロ代名詞や代名詞，指示詞によって文章に出現しない要素を補完して，文章の曖昧性を除去できる点である．それによって，自動要約，クラスタリング，文章分類の支援をすることができる．

照応解析に関する代表的な理論として中心化理論がある．しかし，中心化理論では二つの大きな問題がある．一つ目は，直前の発話で参照された実体しかゼロ代名詞の候補として扱えないという問題である．二つ目は顕現性に影響する要因が文法役割以外にもあるがそれをとらえていないという問題である．

本研究ではゼロ代名詞が示す文法以外の指示対象の特徴,ゼロ代名詞と同時に出現する動詞，格助詞と助詞「ハ」，それと同時に出現する名詞とその頻度で辞書を構築する．

2 提案手法

文末の動詞と同時に出現する，助詞「ハ」と格助詞に対する名詞，頻度，名詞の有無(出現する:1，出現しない:0)を計測する．

次にこの辞書を用いて，照応解析を支援する方法を述べる．先ほどの再現に失敗した例を用いて述べる．

U1:戦後κ周年を機会に来るべきκ年を展望し、αと飛躍を図るため、αに取り組む。
U2:第κはα。
U3:φ(村山富市首相:ガ)規制緩和、特殊法人の見直し、地方分権の推進、情報公開などの行政改革を実行する。

表 9 提案手法の事例

まずゼロ代名詞がある文の文末の動詞，ゼロ代名詞に対応する格助詞を見る．記事中と辞書中で出現する名詞と一致するものを抽出する．例えば，この例ではゼロ代名詞の格助詞はガ格，文末の動詞は「する」である．記事中の単語は{戦後κ周年，機会，κ年，α，飛躍，第κ，規制緩和，特殊法人，見直し，地方分権，推進，情報公開，行政改革}である.その文において助詞「ハ」と格助詞に対しての名詞の有無を(0，1)で表したものと抽出された名詞と共に助詞「ハ」と格助詞に対しての名詞の有無を(0，1)であらわしたもので cos 類似度を計算する．同時に頻度も計算する．そして，cos 類似度≧0.65，1≦頻度≦6 を満たすものを抽出して，辞書情報として用いる．ここで，辞書情報自体の再現率=(候補の中に正解が含まれる数)/(全事例)，絞り込めた割合=(指示対象の候補の数の平均)/(記事の単語のタイプ数の平均)を計算する．ここで辞書情報考慮した Cf ランキング「主題>視点>ガ格>ニ各>ヲ格>その他の格>辞書情報」を定義する．これを用いて，照応解析を行う.

Cf(U1)と Cf(U2)に共通する要素はない．Cf(U2)={第κ(ハ)，{対処，表現，必要性，懸案，機会，対応，中，課題，社会，困難，κ年，世代，わが国，教育，心，村山富市首相，政治，展望}(辞書情報)}，Cb(U2)=不定，Cp(U2)={第κ(ハ)}，φ=第κ 以外の時は，遷移=CONTINUE を得られるため，φ= {対処，表現，必要性，懸案，機会，対応，中，課題，社会，困難，κ年，世代，わが国，教育，心，村山富市首相，政治，展望} が辞書を用いた場合の指示対象の候補であり，”村山富市首相”を含んでいる．

4. 実験

4.1 実験の手法

辞書を用いない照応解析と辞書を用いる照応解析の二つの手法で照応解析をする．CD-毎日新聞’95 年度のテキストデータ，Mecab-0.98，Naist Text Corpus1.5 を用意する．

4.2 実験結果

辞書の再現率と絞り込めた割合は表 10 のようになる．社説記事に比べて，社説記事以外では再現率が 4%減少する.また絞り込めた割合について，社説記事に比べて，社説記事以外のほうが約 2.4 倍単語を絞り込むことができる．また辞書のデータ形式はテキスト形式で，サイズは 9,289,575 バイトとなる．

	単語タイプ数	絞り込めた割合	再現率
社説記事(100件)	152→18	0.1204	0.32
社説記事以外(100件)	88→9	0.1059	0.28

表 10 辞書の再現率，絞り込めた割合

辞書を用いない照応解析に比べて，辞書を用いる照応解析の方が照応解析の再現率が約 23%向上する．しかし指示対象の候補数の平均が辞書を用いない照応解析に比べて，約 5 倍になる．

	指示対象の候補数の平均	再現率
辞書を用いない照応解析	3	0.177
辞書を用いる照応解析	15	0.403

表 11 照応解析の再現率 結果

5. 考察と結論

記事中の単語からゼロ代名詞の特徴(格助詞，cos 類似度，頻度)を用いてゼロ代名詞の候補を抽出した．それらを集めた辞書により，照応解析を支援する手法を提案した．辞書を用いない場合に比べて，再現率は約 23%向上したが，指示対象の候補数の平均が約 5 倍となった．辞書を用いた場合，ゼロ代名詞のある文の直前に指示対象がない場合，直前の文のゼロ代名詞を辞書によって照応している場合において再現率が向上する．ただし指示対象の候補数の平均は辞書を用いない場合の方が小さくなる．よって照応解析の再現率という点において，辞書を用いた照応解析の有用性を示すことができた．

参考文献

[1] 白松俊，宮田高志，奥乃博，橋田浩一，”ゲームの理論による中心化理論の解体と実現語データに基づく検証 ”，言語処理学会，