

Title	新聞記事の固有表現を対象とした参照関係の解析
Author(s)	佐竹, 正臣
Citation	
Issue Date	2002-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1558
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士



新聞記事の固有表現を対象とした参照関係の解析

佐竹 正臣 (910052)

北陸先端科学技術大学院大学 情報科学研究科

2002 年 2 月 15 日

キーワード: 固有表現抽出, 参照表現, 照応解析, 言及クラス.

固有名詞に組織名や人名などの属性タグを付与する固有表現抽出は, テキスト処理における基礎的な技術として重要である. 特に新聞記事には, 時間表現や数値表現などの固有表現が多く含まれているため, 新聞記事を対象に固有表現抽出を行なう研究が数多く行なわれている. 固有表現抽出の先行研究の多くは, 固有名詞の周辺にある単語の情報を手がかりに, 固有表現タグを付与する規則を自動的に学習している. また, 固有表現タグの付与は, 同一文書にある他の固有表現に対するタグの付与とは独立に行なわれる場合が多い. そのため, 以下に挙げる 2 つの問題点がある.

- 同一の対象を表す固有表現が抽出されない

例えば, 同一文章中に「公正取引委員会」と「公取委」という 2 つの固有名詞があるとする. このとき, 前者には“組織名”というタグを付与するが, 後者には固有表現タグを付与しない, すなわち固有名詞として抽出されない場合がある.

- 同一の対象に対して同じタグが付与されない

例えば, 同一文章中に「山岸章」と「山岸」という固有名詞があり, 両者は同一の対象を表しているとする. しかし, 従来の固有表現抽出技術では, 独立に固有表現タグを付与するため, 前者に“人名”, 後者に“組織名”といったように, 異なる固有表現タグを付与する可能性がある.

このような問題に対処するために, 本研究では新聞記事を対象に記事内の固有名詞の照応解析を行ない, その結果を利用して固有表現抽出の精度を向上させることを目的とする. また, 固有名詞を対象とした新しい照応解析アルゴリズムを提案する.

まず, システムの概略について述べる. 最初に固有表現抽出を行ない, 初期の固有表現タグを付与する. 次に照応解析を行ない, 同一の対象を表す表現を特定する. 最後に同一の対象を表す固有表現には同一の固有表現タグを付与するように固有表現タグの整合性を取り, 初期の固有表現抽出結果を修正する.

次に、本研究で提案する照応解析手法について説明する。はじめに、文書から照応解析の対象となる名詞を抽出する。本研究では「固有名詞」または「固有名詞を参照している表現」を参照表現と呼び、照応解析の対象とする。参照表現は大きく分けて次の3種類がある。

省略表現 例えば、「松下電器産業」を単に「松下」と表す

固有表現を指す普通名詞 例えば、「東京大学」を単に「大学」と表す

「同」を用いた表現 例えば、「同社」「同県」など

次に照応解析とそれに用いる素性について述べる。基本的には、先行詞の候補と照応詞の表記の類似度を *LCS* で評価し、*LCS* の高い候補を先行詞とみなす。さらに、本研究では3つの素性、センタリング理論に基づく文法属性、距離、言及クラスを用いる。文法属性とは名詞の格などの文法的な属性であり、センタリング理論では名詞の先行詞のなりやすさの順序を決める素性として用いられる。距離は先行詞と照応詞の近さであり、距離が短いほど先行詞になりやすいとする。また、言及クラスとは、ある同一の対象が次にどう表現されたかを表すクラスである。新聞記事を分析したところ、ある同一の対象が異なる表現で出現した場合、それ以降もそれまでとは異なる別の表記で出現する傾向がみられた。このような表現は、「同」を用いた表現の先行詞になりやすいと考えられる。このことから、同一の対象が異なる表記で複数回言及されている場合を相違言及クラス、全く言及されていない場合を未言及クラス、同じ表記で複数回言及される場合を同一言及クラスと定義し、この順で先行詞になりやすいとした。これらの3つの素性を手がかりに照応解析アルゴリズムを提案し、システムを実装した。

評価実験の結果、照応解析の再現率、精度はそれぞれ 34.78%、54.42%であった。十分な成果が得られなかった原因のひとつは、ALTJAWS の形態素解析の誤りである。すなわち、ALTJAWS の形態素区切りが参照表現の区切りと一致せず、参照表現として取り出すべき名詞が取り出されていないことが多かった。一方、固有表現抽出結果は、提案手法によって、照応解析を用いない従来の固有表現抽出システムと比べて、F 値が 1% 向上した。照応解析の再現率や精度が十分高くないのにも関わらず、固有表現抽出の F 値が向上したことから、照応解析の結果を利用して固有表現抽出のタグ付け結果を修正することは有効であることがわかった。