照応関係を考慮した新聞記事の固有表現抽出

佐竹 正臣 † 白井 清昭 † 奥村 学 ‡
†北陸先端科学技術大学院大学 情報科学研究科
‡東京工業大学 精密工学研究所

1 はじめに

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．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 [6][3][10]．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．

- 同一の対象に対して同じタグが付与されない
  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，つまり固有名詞として抽出されない場合がある．しかし，これらは共に固有表現として抽出し，同じ固有表現タグを付与するべきである．
- 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「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．両者は同一の対象を表しているので，これらには同じ固有表現タグを付与するべきである．

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．具体的には，照応解析によって同一の対象を表す固有表現を特定し，それらに同一の固有表現タグを付与するように初期の固有表現抽出結果を修正する．また，固有表現を対象とした新しい照応解析アルゴリズムも提案する．

2 提案手法

本研究では以下の手順で固有表現抽出を行なう．

1. 固有表現抽出を行ない，初期の固有表現タグを付与する
   ここでは，既存の固有表現抽出アルゴリズムをそのまま用いる．具体的には，関根らが提案した固有表現抽出システム rika[6] を使用する．ただし，rika で用いられている形態素解析器は JUMAN[11] であるが，本研究ではこれを ALTJAWS[9] に置き換える．3 節で述べる照応解析モジュールでは，単語の意味素の情報を必要とする．また，意味素として日本語語彙大系 [5] の意味属性を用いる．ALTJAWS は形態素解析結果と同時に日本語語彙大系の意味属性番号を出力することができるので，形態素解析器として ALTJAWS を用いる．
2. 照応解析を行なう
   照応解析して同一の対象を特定する．このアルゴリズムの詳細は 3節で述べる．
3. 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
   照応解析の結果から同一と判定された固有表現に対して，同一の固有表現タグが付与されるように初期の固有表現抽出結果を修正する．同一の対象を表す固有名詞に異なる固有表現タグが付与されていた場合，rika が出力するタグ付けの信頼度の最も高いタグが正しいとみなして，タグを統一する．また，タグが付与されていない固有名詞については，それと同じ対象を指す他の固有名詞に付与されたタグを新たに付与する．

3 照応解析モジュール

本節では固有名詞を対象とした照応解析アルゴリズムについて述べる．

3.1 参照表現抽出

まず，記事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．

参照表現は主に以下の 3 つに分類される．

1. 省略表現
   例：「松下電器産業」を単に「松下」と表す
2. 固有表現を指す普通名詞
   例：「東京大学」を単に「大学」と表す

3. 「同」を用いた表現
   例：「同社」「同県」

本研究では，初期の固有表現抽出で抽出されなかった固有名詞や，固有名詞を参照する普通名詞に対しても固有表現タグを付与するべきであると考える．そこで固有名詞以外に上記の 2，3 のような普通名詞も参照表現として抽出し，照応解析を行なう．参照表現の抽出は，初期の固有表現抽出の際に得られる ALTJAWS の形態素解析結果を元に行なう．まず，ALTJAWS によって固有名詞を表す品詞が与えられた形態素を参照表現として抽出する．また，普通名詞を表す品詞が与えられた形態素についても，ALTJAWS が出力する意味属性が，組織名や人名などのように固有名詞に近い場合には，参照表現として抽出する．さらに,「同」を用いた表現については「同」「両」「自」といった文字を手がかりにして抽出する．

ALTJAWS による形態素区切りは，固有名詞を表す単位としては細かすぎる場合がある．そこで，以下の場合には形態素を統合し，一つの参照表現として取り扱う．

1. 似ている意味属性を持つ名詞の連続
   例) 東京都＋杉並区，村山＋富市
2. 名詞 (句) ＋接辞
   例) 関西＋国際空港，社会＋党
3. 括弧で括られた名詞 (句)
   例) 「村山政権を支え社民リベラル政治をすすめる会」

3.2 照応解析に用いる素性

本研究では,「同」を用いた表現のときと，それ以外のときとで処理を分ける.「同」を用いた表現以外のときは，木谷 [2] の *LCS(longest common subsequence)* を用いて，同一の対象を指しているかどうかを判定する．*LCS* とは，2 つの文字列のうち，同じ順序で現れる同じ文字の数のことである[1]．2 つの参照表現の *LCS* が大きいとき，それらは同一の対象を表すとみなす．

一方,「同」を用いた表現のときは,「同社」などの表現は長さが短く，表記から得られる情報が少ないため，*LCS* によって先行詞を特定することは難しい．そこで,「同」を用いた表現のときの照応解析の手がかりとなる素性として，以下の 3 つを考える．これらの素性は，固有名詞の先行詞のなりやすさの指標として用いる．

- **センタリング理論に基づく文法属性**
  センタリング理論 [1][8][4] では，名詞の格などの文法的な属性に着目し，式 (1) のような順序で名詞が先行詞になりやすいとしている．

  主題 > 視点 > ガ格 > ニ格 > ヲ格 > その他 (1)

  「主題」は固有名詞が主題化されているとき，ガ格，ニ格，ヲ格はそれぞれの表層格の格要素になっているときを表す．本研究では,「視点」を除き，式 (1) の順序で固有名詞が先行詞になりやすいとする．
- **距離**
  距離とは，固有名詞と先行詞との間に存在する単語数であると定義する．ここでは，距離が小さいほど先行詞になりやすいとする．
- **言及クラス**
  本研究では，照応解析に用いる素性として新たに言及クラスを定義する．言及クラスとは，ある対象が次にどう表現されているかを表すクラスであり，同一の対象ごとに与えられる．言及クラスには以下の 3 つのクラスがある．

  相違言及クラス ある対象と同一の対象が異なる表現で言及されたもの．例えば「九条署」が後に「同署」となった場合，これらの言及クラスは相違言及となる．

  未言及クラス ある対象が文章内で始めて出現し，それ以前にはまだ言及されていないもの．

  同一言及クラス ある対象と同一の対象が同じ表現で言及されたもの．例えば「九条署」が同じ文書内でもう一度「九条署」として出現した場合，これらの言及クラスは同一言及となる．

  新聞記事を調べたところ，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．したがって,「同」という表現が出現したとき，それ以前に様々な表記で出現した対象を指しやすいと考える．そこで本研究では，相違言及，未言及，同一言及の順に固有名詞が先行詞になりやすいとする．また，先行詞の候補となる複数の固有名詞が同じ言及クラスに属する場合，言及回数の多い固有名詞ほど先行詞になりやすいとする．言及回数とは，同一の対象を指す固有名詞が記事中に出現した回数である．

「同」を用いた表現の照応解析に関する先行研究としては，木谷 [2]，若尾 [7] らの研究が挙げられる．これらは「同社」,「両社」,「自社」など，特定の表現に限定して照応解析を行なっている．これに対し，本研究では,「同社」以外の幅広い表現 (例えば「同県」「同氏」など) についても考慮している．

[1]例えば,「松下電器」と「松下」の *LCS* は 2 である．

3.3 照応解析アルゴリズム

ここでは，参照表現間の照応解析を行うアルゴリズムについて説明する．まず，3.1 項で述べた手法で抽出された参照表現の集合を $r_1, \cdots, r_n$ とする．ただし，r_i は文書中の出現順序にしたがって並べられているとする．また，先行詞候補のリスト AL を用意する．AL は，先行詞の候補となりうる既出の参照表現を，先行詞のなりやすさの順序で並べたリストである．そして，AL は，参照表現を組織名，人名などのクラスに分類し，各クラス毎に個別に作成する．このクラス分けは固有表現タグ付けと似ているが，初期の固有表現抽出結果は用いずに，ALTJAWS が出力する意味属性だけを用いて簡易に行う．

参照表現 r_i の先行詞を決定する手続きは以下の通りである．

1. AL を空にする．$i = 1$ とする．
2. $AL = \{a_1, \cdots, a_m\}$ であるとする．要素 a_j について，リストの先頭から順に，r_i の先行詞が a_j となるかどうかを調べる．a_j が r_i の先行詞となる条件は以下の通りである．
   - r_i が「同」を用いた表現以外のとき
     a_j と r_i の LCS が 2 以上のとき，r_i の先行詞は a_j であるとみなす．
   - r_i が「同」を用いた表現のとき
     a_i の ALTJAWS による品詞タグが「固有名詞」であるなら，r_i の先行詞は a_j であるとみなす．
3. AL の中から r_i の先行詞が見つからなければ，r_i は文書中の最初に出現した固有名詞であるとみなす．また，r_i は後続する固有名詞の先行詞となる可能性があるので，これを AL に追加する．
4. AL の要素の順序を並び替える．順序は 3.2 節で述べた素性をもとに決定する．4 節で述べる評価実験では，以下の 3 つの基準で先行詞の候補の並び替えを行った．

   基準 1 言及クラス+言及回数 > 文法属性 > 距離

   基準 2 文法属性 > 言及クラス+言及回数 > 距離

   基準 3 距離

   “基準 1” では，まず「言及クラス」「言及回数」によって AL の要素の順序を決める．これが同じときには「文法属性」によって AL の要素の順序を決め,「文法属性」も同じときには「距離」によって決定する．“基準 2” と “基準 3” も同様である．
5. $i = n$ なら終了．それ以外は 2. へ戻る．

表 1: 照応解析の結果

	基準 1	基準 2	基準 3
再現率	34.78	34.78	31.52
精度	54.42	54.23	52.16
F 値	42.44	42.38	39.29

4 評価実験

提案手法を評価する実験を行った．実験に使用したのは IREX の NE_DRYRUN データ 34 記事である．これらの記事に対して，提案手法により固有表現抽出を行なった．使用した固有表現タグは 8 種類である[2]．IREX のデータには，正解となる固有表現タグが付与されている．この正解とシステムの出力とを比較し，再現率，精度，F 値[3] を調べて評価を行なった．さらに，照応解析結果についても，再現率，精度，F 値による評価を行った．正解となる参照表現とその先行詞は人手で与えた．

4.1 照応解析の結果

まず，照応解析の結果を表 1 に示す．表 1 における “基準 1”, “基準 2”, “基準 3” は，3.3 項で述べた先行詞候補のリスト AL の順序を決定する 3 つの基準を表わす．この表からわかるように，照応解析の再現率，精度ともに十分高いとは言えない．この原因のひとつは，ALTJAWS の形態素解析の誤りである．特に，再現率が低い理由として，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挙げられる．これについては，3.1 項で説明したように，複数の形態素を連結して参照表現を抽出するように工夫はしているが，十分な成果が得られていないことがわかった．

本研究では,「同」を用いた表現の先行詞を決定するために,「言及クラス」「言及回数」といった素性を新たに導入した．しかし,「言及クラス+言及回数」を最も重視した基準 1 と,「文法属性」を最も重視した基準 2 とでは，F 値にほとんど差が見られない．これは，実験に用いた新聞記事の中に,「同」を含む表現があまり存在しなかったためである．評価した記事の中に参照表現は 460 個あったが，そのうち「同」を含む表現はわずかに 13 であった．したがって，本研究で提案する「言及クラス」や「言及回数」が，照応解析に用いる素性として有効であることを確かめることができなかっ

[2] 組織名 (ORGANIZATION)，人名 (PERSON)，地名 (LOCATION)，固有物名 (法律名，商品名など)(ARTIFACT)，日付表現 (DATE)，時間表現 (TIME)，金額 (MONEY)，割合表現 (PERCENT)．

[3] 係数 β は 1 とする．

た．今後,「同」を含む表現の事例を増やして，評価実験を行う予定である．

4.2 固有表現抽出の結果

rika による固有表現抽出結果と，本研究で提案した手法による固有表現抽出の結果を表 2 に示す．rika の結果は [6] で報告されている値よりも劣っている．特に，ARTIFACT の抽出精度が極端に悪い．これは special dictionary[6] を使用していないことが原因であると考えられる．

提案手法は，rika と比べて，F 値がわずかに向上した．表 1 に示したように，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と言える．照応解析の性能が向上すれば，固有表現抽出の精度や再現率もさらに改善されると期待できる．

精度と再現率を見ると，照応解析結果を用いることにより，精度は低下したが再現率は向上している．再現率が向上したのは，rika で固有表現として抽出されなかった参照表現も，照応解析結果に基づくタグの修正によって，新たに固有表現タグを付与することができたためと考えらえる．

表 2: 固有名詞抽出の実験結果

	rika		提案手法	
	再現率	精度	再現率	精度
ORG	42.55	66.67	54.26	62.96
PERSON	74.63	78.12	79.10	69.74
LOCATION	61.58	68.02	70.00	62.44
ARTIFACT	0.00	0.00	0.00	0.00
DATE	75.89	70.83	81.25	72.80
TIME	63.64	73.68	63.64	70.00
MONEY	100.00	100.00	100.00	100.00
PERCENT	83.33	100.00	83.33	100.00
ALL SLOTS	59.70	72.58	66.57	68.07
F 値	65.51		67.32	

5 おわりに

本稿では固有表現抽出の精度を向上させるために照応解析の結果を利用する手法を提案し，その有効性を実験により確認した．また，固有名詞を対象とした照応解析の新たな素性として「言及クラス」を提案し，他の素性と併用する照応解析手法を提案した．

最後に，今後の課題を 2 つ挙げる．まず，提案手法は言い換え表現には対応していない．言い換え表現とは，例えば「第二電電」と「DDI」などのように，日本語表記が英語表記の省略形になった表現や,「朝鮮民主主義人民共和国」と「北朝鮮」などのように，省略とは違う形で言い換えられる表現である．言い換え表現は *LCS* によって関連性を見つけることが難しく，固有名詞の言い換え表現を収集した辞書が必要となる．

また，3.1 項で挙げた 3 つの参照表現のタイプのうち，固有表現を指す普通名詞と省略表現は *LCS* を用いて先行詞を特定していたが，*LCS* だけでは先行詞が特定できない場合もある．例えば「北陸先端大」の後に単に「大学」とあれば，*LCS* では同じ順序で文字が現われていないため，同一の対象とは見なされない．*LCS* によるスコアが低くても，ALTJAWS が出力する意味属性が等しければ先行詞とみなす手法も試みたが，精度は悪化した．したがって，*LCS* に代わる判定基準も検討する必要がある．

参考文献

[1] Barbara J.Groze, Arvind K.Joshi, and Scott Weinstein. Providing a Unified Account of Definite Noun Phrases in Discourse. *21st Annual Meeting of the Association for Computational Linguistic*, pp. 44–50, 1983.

[2] Tsuyoshi Kitani. Merging Information by Discourse Processing for Information Extraction. *tenth IEEE Conference on Aritficial Intelligence for Applications*, pp. 412–418, 1994.

[3] 内元清貴, 村田真樹, 小作浩美, 馬青. ME モデルと書き換え規則に基づく固有表現抽出 — IREX-NE 本試験における評価 —. IREX ワークショップ予稿集, pp. 133–140, 1999.

[4] 石崎雅人, 伝康晴. 談話と対話, 第 5 章 談話構造と照応. 東京大学出版会, 2001.

[5] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[6] Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. *COLING-ACL'98,Proceedings of the Sixth Workshop on Very LargeCorpora*, pp. 171–178, 1998.

[7] Takahiro Wakao. Reference Resolution Using Semantic Patterns in Japanese Newspaper Articles. *In Proceedings of COLING 94*, pp. 1133–1137, 1994.

[8] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese Discourse and the Process of Centering. *Computational Linguistics*, Vol. 20, No. 2, pp. 193–232, 1994.

[9] ＮＴＴコミュニケーション科学基礎研究所知能情報研究部 翻訳コミュニケーション研究グループ. ＮＴＴ日本語形態素解析ライブラリ ＡＬＴＪＡＷＳ Version 2.0, 1999.

[10] 山田寛康, 工藤拓, 松本祐治. Support Vector Machines を用いた日本語固有表現抽出. 情報処理学会研究会報告,No.NL-142-18, pp. 121–128, 2001.

[11] 松本裕治, 黒橋禎夫, 山地治, 妙木裕, 長尾真. 日本語形態素解析システム JUMAN Ver.3.1, 1997.