確率モデルに基づく日本語ゼロ代名詞の照応解消

関 和広[†]　藤井 敦[†,††]　石川 徹也[†]
[†] 図書館情報大学
[††] 科学技術振興事業団 CREST
{seki,fujii,ishikawa}@ulis.ac.jp

1 はじめに

自然言語では自明な対象への言及や冗長な繰り返しを避けるため，代名詞などの照応表現が使われる．さらに，文脈や状況から読み手や聞き手が容易に推測できる対象は，代名詞すら利用されずに省略されることがある．このように省略された格要素をゼロ代名詞という．日本語では，主語がゼロ代名詞化されることが多い．

これら照応表現の指示対象を特定することを照応解消という．機械翻訳，自動要約，情報検索などの自然言語処理の応用分野は，照応解消によって高度化が期待できる [2, 4, 8, 13].

従来の照応解消に関する研究は，規則に基づく手法と統計的手法に大別できる．

規則に基づく手法 [3, 6, 14] では，助詞や文間距離などに基づく規則を利用して照応解消を行う．この手法では対象ごとに詳細な規則を設定することが可能である．しかし，人手で規則を作成するために恣意性が生じやすい．また，規則の数が増えるほど規則間の整合性を保つことが困難になる．

統計的手法 [5, 7, 9] では，照応関係が付与されたコーパスを用いて確率モデルや決定木などを学習し，照応解消を行う．この手法では，属性値を実データに基づいて推定するために恣意性が少ない．しかし，モデルが複雑になるほど推定する属性の数が増え，大きな学習データが必要になる．

本研究では，日本語に多く現れるガ・ヲ・ニ格のゼロ代名詞を対象に確率モデルを用いた照応解消手法を提案する．本手法では，照応関係が付与されたコーパスと付与されていないコーパスを利用して学習データ量の問題に対処する．

2 提案する照応解消手法

2.1 概要

本提案手法では，まずIPAL動詞辞書 [12] を参照し，入力文章中に欠落している必須格をゼロ代名詞と見なす．

次に，特定したゼロ代名詞 ϕ について，名詞（単語，複合語）を指示対象候補 C_i として文章中から抽出する．候補の抽出範囲は，ゼロ代名詞が出現した文から前段落の1文目までとする．

最後に，各 C_i についてゼロ代名詞 ϕ の指示対象となる確率 $P(C_i|\phi)$ を計算し，値が大きい順に出力する．

2.2 利用した属性

本手法では，ゼロ代名詞 ϕ と指示対象候補 C_i の属性として，先行研究を参考に以下の6つを利用した．

- 統語的な属性
  - 指示対象候補に共起する助詞 (p_i)
  - ゼロ代名詞と指示対象候補間の距離 (d_i)：
    両者が同一文にあれば 0，n 文前であれば n ($n > 0$) を値にとる．
  - 指示対象候補が連体修飾句に含まれるかどうか (m_i)：
    含まれれば 1，含まれなければ 0 とする．
  - ゼロ代名詞の格 (z)：
    係り受け情報とIPAL動詞辞書を利用して決定する．ガ，ヲ，ニのいずれかを値にとる．
- 意味的な属性
  - 指示対象候補の意味分類 (n_i)：
    分類語彙表 [11] の分類番号に基づいて決定する．
  - ゼロ代名詞の指示対象が持つ意味素性 (s)：
    IPAL動詞辞書の意味素性（19種類）を用いる．

2.3 確率モデル

ゼロ代名詞 ϕ が与えられたときに，名詞 C_i が ϕ の指示対象である確率を $P(C_i|\phi)$ とする．ここで，C_i と ϕ を 2.2 節で述べた属性で表現すると式 (1) が成り立つ．

$$P(C_i|\phi) = P(p_i, d_i, m_i, n_i|z, s) \quad (1)$$

C_i の属性のうち，d_i，m_i はそれ以外の属性との関連が低いので，式 (1) を式 (2) のように変形する．

$$P(C_i|\phi) = P(p_i, n_i|z, s) \cdot P(d_i) \cdot P(m_i) \quad (2)$$

$P(p_i, n_i|z, s)$ において，p_i と z はそれぞれ指示対象候補とゼロ代名詞の統語的属性であり，n_i と s はそれぞれ指示対象候補とゼロ代名詞の意味的属性である．そこで，p_i と z，n_i と s のみが依存関係にあると考え，式 (3) のように変形する．

$$P(C_i|\phi) = P(p_i|z) \cdot P(d_i) \cdot P(m_i) \cdot P(n_i|s) \quad (3)$$

式 (3) の右辺において，$P(p_i|z) \cdot P(d_i) \cdot P(m_i)$ を統語モデル，$P(n_i|s)$ を意味モデルと呼ぶ．

2.4 確率値の推定

$Fr(x)$ を x の頻度と定義すると，式 (3) 右辺のそれぞれの項は式 (4)〜(7) で表される．

$$P(p_i|z) = \frac{Fr(p_i, z)}{\sum_j Fr(p_j, z)} \quad (4)$$

$$P(n_i|s) = \frac{Fr(n_i, s)}{\sum_j Fr(n_j, s)} \quad (5)$$

$$P(d_i) = \frac{Fr(d_i)}{\sum_j Fr(d_j)} \quad (6)$$

$$P(m_i) = \frac{Fr(m_i)}{\sum_j Fr(m_j)} \quad (7)$$

式 (5) において，分類番号 n_i は 660 通り，意味素性 s は 19 通りからなるので，その全組み合わせに対して $P(n_i|s)$ を正しく推定するためには大規模な学習データが必要である．しかし，学習データの作成はコストが高い．そこで，分類番号の各桁が上位から下位へ階層的な意味分類を表していることを利用し，分類番号の上位の桁を考慮してディスカウンティング（スムージング）を行い Fr を近似する．これを式 (8) に示す．なお，${n_k}^j$ は n_k の上位 j 桁を表す．

$$Fr^*(n_k, s) = Fr({n_k}^1, s) + Fr({n_k}^2, s) + \cdots + Fr({n_k}^5, s) \quad (8)$$

2.5 共起情報の利用

$P(n_i|s)$ を正確に推定するためには，照応関係を付与した学習データが必要である．また，IPAL 動詞辞書の未登録語に関しては $P(n_i|s)$ を計算できない．辞書に登録されていても IPAL 動詞辞書の意味素性は高々 19 であり，指示対象の意味分類を表すには十分ではない．そこで，照応関係が付与されていないコーパスから $P(n_i|s)$ を推定する必要がある．

指示対象の意味素性 s は，述語の語義とゼロ代名詞の格によって決まる．ここで，述語の語義が単一であり，ひとつの語義には重複する格がないと仮定すると式 (9) の近似が成り立つ．

$$P(n_i|s) \approx P(n_i|v, z) = \frac{Fr(n_i, v, z)}{\sum_j Fr(n_j, v, z)} \quad (9)$$

$P(n_i|v, z)$ はゼロ代名詞と指示対象の意味的な整合性を表す．しかし，ゼロ代名詞は本来表出すべき格要素の省略であるから，$P(n_i|v, z)$ は述語と格要素の関係に置き換えることができる．そこで，コーパス内の文の係り受け関係を解析し，述語と格要素の共起関係を抽出することで式 (9) を推定する．

3 評価実験

2 章で提案した確率モデルの有効性を評価するために以下の観点から照応解消実験を行った．

(1) 統語モデル，意味モデルを個別に，あるいは組み合わせて用いる．

(2) 式 (3) の各確率値の推定に利用する学習データ量を変化させ，学習データ量と正解率の関係を観測する．

(3) 共起情報を用い，式 (9) により $P(n_i|s)$ を推定する．

(4) 一定の閾値を設け，指示対象としての確信度が高い候補だけを出力する．

なお，式 (3) の各確率値の推定，および評価実験には，京都大学テキストコーパス [1] 収録の社説記事 30 件（平均 28.9 文/件），一般報道記事 30 件（同 14.1 文/件）を用いた．共起情報の獲得には'94〜'99 年の毎日新聞記事 [15] を用いた．

上記 (2) 以外の評価は交差確認法で行った．すなわち，1 記事をテストデータ，他の 29 記事を学習データとする試行を 30 回繰り返し，結果を平均した．

(1) 利用するモデルと正解率

統語モデル，意味モデルを個別に，あるいは組み合わせて用いた場合の照応解消実験結果を表 1 に示す．

表 1: 照応解消の実験結果

モデル	順位	社説記事 正解数	一般報道記事 正解数
統語	1 位	171 (34.4%)	**184 (52.6%)**
	1-2 位	245 (49.3%)	**219 (62.6%)**
	1-3 位	298 (60.0%)	243 (69.4%)
意味	1 位	123 (24.7%)	106 (30.3%)
	1-2 位	191 (38.4%)	158 (45.1%)
	1-3 位	238 (47.9%)	193 (55.1%)
組合せ	1 位	**183 (36.8%)**	168 (48.0%)
	1-2 位	251 (50.5%)	**219 (62.6%)**
	1-3 位	**304 (61.2%)**	**247 (70.6%)**
規則	1 位	175 (35.2%)	127 (36.3%)
	1-2 位	**254 (51.1%)**	179 (51.1%)
	1-3 位	292 (58.8%)	217 (62.0%)

表 1 の「順位」は，$P(C_i|\phi)$ の値に基づく順位であり，「1 位」「1-2 位」「1-3 位」の各行は，それぞれ上位 1，1〜2，1〜3 位に含まれた正解数である．

最下段の「規則」は，従来の手法で典型的に用いられる順位付け規則に基づいた照応解消結果である．具体的には，1) ゼロ代名詞と指示対象候補の意味的整合性，2) ゼロ代名詞と指示対象候補の文間距離，3) 指示対象候補に共起する助詞に関する規則を用いた．

社説記事と一般報道記事の正解率を比べると，後者の方が総じて正解率が高かった．これは対象とした文の長さだけでなく，分野の特殊性に起因する照応解消の難易度の違いを反映している．

また，意味モデルよりも統語モデルの方が良い結果が得られた．2 つのモデルを組み合わせた場合，社説記事では正解率が向上したのに対し，一般報道記事では 1 位の解の正解率が統語モデルよりも低下した．この結果から，一般報道記事では助詞や距離などの統語的な要因が照応関係により強く影響していることが分かる．

次に「組合せ」確率モデルによる手法と「規則」に基づく手法を比べると，社説記事はほぼ同程度の正解率であったのに対して，一般報道記事では確率モデルを用いた方が 10 ポイント程度良い結果が得られた．この結果から，確率モデルによる手法は，対象とする文章の性質を捉えやすいと考えられる．

(2) 学習データ量と正解率

式 (3) の確率値の計算に用いた学習データの記事数と 1 位の正解率の関係を図 1 に示す．

社説記事，一般報道記事とも，学習データの増加とともに正解率が向上した．しかし，20 記事を過ぎた辺りから正解率が下がる場合があった．

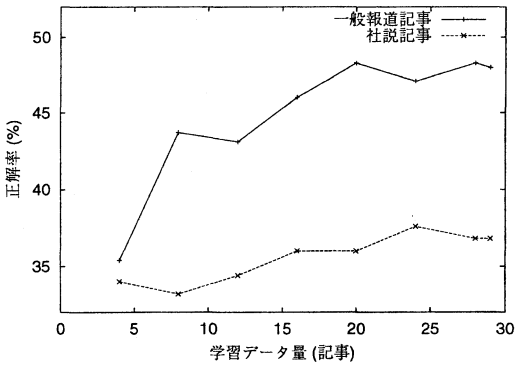

図 1: 学習データ量と正解率の関係

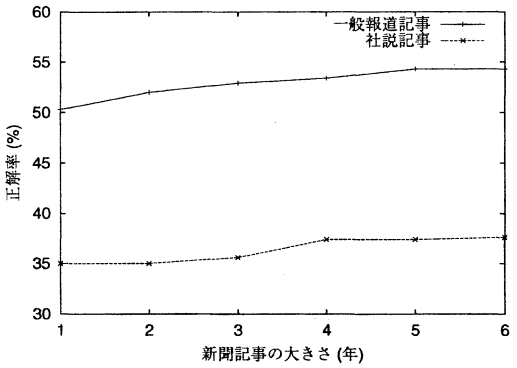

図 2: 新聞記事の大きさと正解率の関係

(3) 共起情報の利用

照応関係が付与されていないコーパス（新聞記事）を構文解析し，その結果得られた述語と格要素の係り受け関係を照応解消に利用した．係り受け関係は，新聞記事を JUMAN [10] で形態素解析し，「名詞（句）は後方最近傍の述語に係る」という比較的簡単な規則を仮定して抽出した．利用する新聞記事の量を 1〜6 年分まで変化させたときの照応解消の正解率を図 2 に示す．

社説記事，一般報道記事とも，学習に利用する新聞記事の量とともに正解率が向上した．特に 3〜4 年分以上の新聞記事を用いた場合，1 位の正解率は共起情報を使わない場合（表 1）の最高値以上の結果が得られた．これは，次の理由によるものと考えられる．

- IPAL 動詞辞書の未登録語に関しても，意味的な整合性 $P(n_i|s)$ を計算できた．
- 述語とゼロ代名詞の格の組み合わせを意味分類的に用いることで，意味素性数の影響を受けなかった．

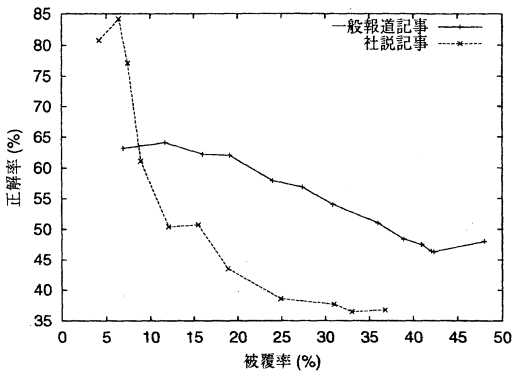

図 3: 被覆率と正解率の関係

(4) 確信度に基づく照応解消

照応解消の結果を機械翻訳など他の言語処理で応用する場合，全てのゼロ代名詞を処理するより，被覆率 (coverage) は低くても精度 (accuracy) が高い方が好ましいことがある．そこで，ある閾値を設けて，$P(C_i|\phi)$ が閾値以上の候補 C_i だけを出力した場合の被覆率と正解率の関係を図 3 に示す．

図 3 から，確信度の高い候補だけを出力することで，より高い正解率が得られることが分かる．

4 関連研究との比較

本手法は，確率モデルを用いてガ格の補完処理を行う江原ら [7] の手法に似ている．しかし，本手法は以下の点で江原らの手法と異なる．

- 江原らの手法が重文の分割によって生じるガ格の省略の文内照応（同一文内に指示対象がある照応）のみを扱っているのに対して，本手法はガ・ヲ・ニ格の前方照応（ゼロ代名詞の出現個所以前に指示対象がある照応）を対象としている．
- 本手法では，照応関係が付与されていないコーパスを学習に用いることで，学習データ量の問題や辞書未登録語の問題に対処した．
- 本手法では，ゼロ代名詞の指示対象として確信度の高い候補だけを出力することで，照応解消処理の精度向上が可能であることを示した．

5 おわりに

本研究では，確率モデルに基づく日本語ゼロ代名詞の照応解消法を提案し，評価実験によって一般報道記事，社説記事に対する有効性を示した．また，照応関係が付与されていないコーパスを用いて正解率を向上させることができた．さらに，確信度の導入によって，より高い精度で照応解消を行うことができた．

参考文献

[1] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pp. 719–724, 1998.

[2] Elizabeth DuRoss Liddy. Anaphora in natural language processing and information retrieval. *Information Processing & Management*, Vol. 26, No. 1, pp. 39–52, 1990.

[3] Ruslan Mitkov, Lamia Belguith, and Malgorzata Stys. Multilingual robust anaphora resolution. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, pp. 7–16, 1998.

[4] Thomas S. Morton. Coreference for NLP application. In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, pp. 173–180, 2000.

[5] Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. Corpus-based learning for noun phrase coreference resolution. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 285–291, 1999.

[6] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese discouse and the process of centering. *Computational Linguistics*, Vol. 20, No. 2, pp. 193–233, 1994.

[7] 江原暉将, 金淵培. 確率モデルによるゼロ主語の補完. 自然言語処理, Vol. 3, No. 4, pp. 67–86, 1996.

[8] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 言語理解とコミュニケーション研究会（編）, 自然言語処理と情報提示技術, pp. 1–24. 電子情報通信学会, 1998.

[9] 山本和英, 隅田英一郎. 決定木学習による日本語対話文の格要素省略補完. 自然言語処理, Vol. 6, No. 1, pp. 3–28, 1999.

[10] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.61. 京都大学大学院情報学研究科, 1998.

[11] 国立国語研究所（編）. 分類語彙表. 秀英出版, 1964.

[12] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL (Basic Verbs) 解説編, 1987.

[13] 中岩浩巳, 池原悟. 語用論的・意味論的制約を用いた日本語ゼロ代名詞の文内照応解析. 自然言語処理, Vol. 3, No. 4, pp. 49–64, 1996.

[14] 村田真樹, 長尾真. 用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定. 自然言語処理, Vol. 4, No. 1, pp. 87–109, 1997.

[15] 毎日新聞. 毎日新聞 CD-ROM 版 1994～1999 版. 日外アソシエーツ.