ゼロ代名詞の検出と補完を統合した確率的照応解消モデル

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冗長な繰り返しを避けるために，代名詞などの照応表現が使われる．特に日本語では，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頻繁に省略される．このように省略された格要素をゼロ代名詞という．省略を補完する処理（ゼロ代名詞の照応解消）は文脈解析，音声対話，機械翻訳などに有効であり，これまでも研究が行なわれてきた [7, 8].

ゼロ代名詞の照応解消処理には，ゼロ代名詞出現箇所の「検出」と指示対象の「特定」が必要である．既存の照応解消手法は，ゼロ代名詞があらかじめ正しく検出されているという前提で指示対象の特定だけに注目してきた．しかし，照応解消処理を実用化するためには，ゼロ代名詞の検出についても検討する必要がある．

そこで本研究では，ある格がゼロ代名詞化している確率を考慮し，ゼロ代名詞の検出と指示対象の推定を統合する手法を提案する．

2 本研究の焦点

ゼロ代名詞の検出を計算機によって自動的に行なった場合，真のゼロ代名詞以外の箇所まで余計に検出してしまったり，逆に真のゼロ代名詞を検出できないといった誤りが生じる．自動検出によるゼロ代名詞と真のゼロ代名詞の関係を図1に示す．

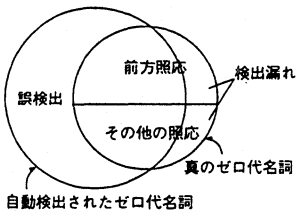

図 1: 自動検出のゼロ代名詞と真のゼロ代名詞の関係

図1では，左右の円がそれぞれ自動検出されたゼロ代名詞と真のゼロ代名詞の集合を表す．真のゼロ代名詞には前方照応のほか，外界照応や後方照応のゼロ代名詞も含まれているため，適切に照応解消を行なうためには，これらを区別しなければならない．

本研究では前方照応のゼロ代名詞に焦点を当てるため，自動検出したゼロ代名詞のうち「誤検出」，および前方照応以外の「その他の照応」の排除を試みる．また，図1における「検出漏れ」のゼロ代名詞が少なくなるように再現率（recall）を重視して処理を行う．

3 方法

3.1 ゼロ代名詞候補の検出

入力文の係り受け構造と動詞の格フレーム[1]を比較し，充足されていない格要素をゼロ代名詞の候補として検出する．検出の対象は，日本語ゼロ代名詞のほとんどを占める「ガヲニ」格に限定する．ただし，ガガ構文（例：「私は給料が上がった」）に見られるように，ガ格は同一の動詞に二つ同時に係ることがある．そこで，ガガ構文を取る動詞を考慮し，二つのガ格を動詞に近い方からそれぞれ「ガ$_1$」「ガ$_2$」として区別する．

動詞の格フレームは，IPAL 動詞辞書 [5] を参照して決定する．ただし，IPAL 動詞辞書は登録動詞数が 861 語と少ないため，辞書未登録の動詞に関しては IPAL 動詞辞書記載の類義語情報を利用して対処する．すなわち，意味的に類義の動詞は同一の格フレームを取りやすいことに注目し，ある動詞の類義語はその動詞と同一の格フレームを持つと考える．これにより，さらに 2,038 の動詞に関しても格フレームを与えることが可能となる．

一方，動詞には多義性があり，語義が異なれば格フレームが異なることも多い．よって，ゼロ代名詞の検

[1] ここでは，動詞の可能な必須格を記述した情報を格フレームと定義する．

出を正確に行うためには動詞の多義性を適切に解消する必要がある．しかし，動詞の多義性解消は難しい問題であり，語義の解消誤りはゼロ代名詞の誤検出につながる可能性がある．

そこで本手法では，ゼロ代名詞検出の再現率を高めることを優先して，格 c が動詞 v の任意の語義で必須格であれば格 c を格フレームに加える．例えば，動詞 v が語義によって「ガ$_2$，ガ$_1$，ヲ」「ガ$_1$，ニ」という2種類の格フレームを持つならば，動詞 v の格フレームを「ガ$_2$，ガ$_1$，ヲ，ニ」（順不同）と見なす．

IPAL動詞辞書（および類義語情報）を利用しても格フレームを決定できない動詞に関しては，「ガ$_1$」格のみを必須格とする．これにより，本手法では全ての動詞を対象にゼロ代名詞を検出することができる．

このように，本手法ではIPAL動詞辞書に記載された必須格かガ格のみをゼロ代名詞として検出するため，係り受けが不可能な格をゼロ代名詞として検出することは少ない．

3.2 ゼロ代名詞の検出と補完を統合した確率モデル

動詞 v の充足されていない格 c がゼロ代名詞である確率を $P_{zero}(c|v)$，動詞 v の格 c に対応するゼロ代名詞 ϕ_c が指示対象候補 a を照応する確率を $P_{ana}(a|\phi_c)$ で表す．動詞 v の格 c がゼロ代名詞であり，かつそのゼロ代名詞が a を照応する確率を式 (1) で表す．

$$P_{ana}(a|\phi_c) \cdot P_{zero}(c|v) \tag{1}$$

ここで，$P_{zero}(c|v)$ は格 c がゼロ代名詞である度合を示すので，この値が大きければ格 c のゼロ代名詞 ϕ_c が真のゼロ代名詞である可能性は高くなる．すなわち，図1の「誤検出」である可能性は低くなる．また，$P_{ana}(a|\phi_c)$ は指示対象候補中に適当な候補がない場合，理論的には低い値をとる．よって，指示対象候補を前方の文脈から抽出した場合，文脈内に候補が存在しない外界照応などのゼロ代名詞に関しては，低い確率値をとると考えられる．

よって，式 (1) の値を利用してシステムの出力を制限すれば，図1の「誤検出」と「その他の照応」のゼロ代名詞を削減する効果が期待できる．そこで，式 (1) の値に基づいて式 (2) に定義する確信度を計算し，確信度が閾値以上の場合だけ結果を出力する．ここで $P_i(\phi_c)$ は，式 (1) の値によって指示対象候補を降順にソートしたときの i 番目の候補に関する式 (1) の値を表す．t は0～1の定数であり，現在は経験的に0.5としている．

$$C(\phi_c) \stackrel{def}{=} t \cdot P_1(\phi_c) + (1-t)(P_1(\phi_c) - P_2(\phi_c)) \tag{2}$$

確信度 $C(\phi_c)$ は，$P_1(\phi_c)$ と $P_2(\phi_c)$ の差が大きいほど，または $P_1(\phi_c)$ が大きいほど高くなる．

3.3 $P_{zero}(c|v)$ の推定

ある格 c がゼロ代名詞化しているかどうかは，動詞 v に関する係り受け関係と動詞 v の必須格によって決まる．ここで，与えられた係り受け関係が正しければ，充足されていない格 c は，動詞 v において必須格であればゼロ代名詞，必須格でなければ非ゼロ代名詞となる．すなわち，格 c がゼロ代名詞化しているかどうかは，格 c が必須格であるかどうかに依存する．しかし，必須格と任意格の判断基準は必ずしも明確ではない．

そこで本研究では，格 c が必須格となり得る度合を動詞 v と格 c の共起頻度に基づいて連続値で表す．すなわち，動詞 v が格 c を伴って現れることが多いほど格 c が必須格である確率が高く，よって格 c がゼロ代名詞化している確率 $P_{zero}(c|v)$ が高いと考える．

動詞 v に関して「ヲ」「ニ」格が必須格である確率は，動詞 v に対する両者の相対頻度で表す．一方，ガ格はほとんどの動詞において必須であることから，充足されていなければ常に省略されているものと見なす．ここでは，「ガ$_1$」をいずれの動詞にも必須の格と仮定する．「ガ$_2$」格に関しては，動詞 v とガ$_1$ 格の組に対する相対頻度で動詞 v の必須格である確率を表す．以上を式 (3) にまとめる．ここで $F(x)$ は x の頻度を表す．

$$P_{zero}(c|v) \stackrel{def}{=} \begin{cases} \dfrac{F(c,v)}{F(v)} & if\ c \in \{\text{ヲ}, \text{ニ}\} \\ 1 & if\ c = \text{ガ}_1 \\ \dfrac{F(\text{ガ}_2, \text{ガ}_1, v)}{F(\text{ガ}_1, v)} & if\ c = \text{ガ}_2 \end{cases} \tag{3}$$

3.4 $P_{ana}(a|\phi_c)$ の推定

ゼロ代名詞 ϕ_c が候補 a を照応する確率 $P_{ana}(a|\phi_c)$ を推定するために，筆者らが提案した確率モデル [2, 6] を利用する．意味クラス（分類語彙表 [4] の分類番号）n，後接する助詞 p，ϕ_c との文間距離 d，連体修飾句に関する制約 r で a を表現し，動詞 v と格 c で ϕ_c を表現し，式 (4) のように変形する．

$$\begin{aligned} P_{ana}(a|\phi_c) &= P(n,p,d,r|v,c) \\ &\approx P(n|v,c) \cdot P(p|c) \cdot P(d) \cdot P(r) \end{aligned} \tag{4}$$

式 (4) 右辺の最初の要素 $P(n|v,c)$ を「意味モデル」，残りの要素 $P(p|c)\cdot P(d)\cdot P(r)$ を「統語モデル」と呼ぶ．意味モデルの推定には動詞と格要素の共起情報を用い（3.5 節参照），統語モデルの推定には照応関係を付与したコーパスを学習データとして用いる．

3.5 共起情報の収集

式 (3)，および式 (4) の意味モデルを計算するため，照応関係が付与されていない未解析のコーパスから動詞とその格要素の組を収集する．手順を以下に示す．

(1) 入力文を形態素解析する（JUMAN [3] を利用）．

(2) 次の規則によって係り受け関係を同定する．
名詞は後方最近傍の動詞に係ると仮定する．ただし，読点をまたいだ係り受けは係り受け関係に曖昧性があるので除外する．また，受身・可能・使役文は格の交替が起きるので除外する．

(3) 係り受け関係にある「名詞・格・動詞」を抽出する．助詞「を」「に」をそれぞれヲ格，ニ格として扱う．助詞「は」「が」はどちらもガ$_1$ 格として扱う．ただし，「は」「が」が同一の動詞に係っている場合は，動詞に近い方をガ$_1$ 格，遠い方をガ$_2$ 格とする．
（例 1）私 は$_{(ガ_2)}$ 給料 が$_{(ガ_1)}$ 上がった．
（例 2）私 は$_{(ガ_1)}$ 昼 に$_{(ニ)}$ 起きた．
例 1 の「～上がった」はガガ構文であり，ガ格を二つ取り得る．対して，例 2 の「～起きた」はガ格を一つしか取らない．しかし，例 3 のように，副詞句がガ格と誤認識される場合も考えられる．
（例 3）今日 は$_{(ガ_2)}$ 私 は$_{(ガ_1)}$ 昼 に$_{(ニ)}$ 起きた．
いずれの場合もガ$_1$ 格を必須格と仮定する．これは，一般に動詞の意味上重要な要素ほど動詞の近くに現われることが多いからである．一方，ガ$_2$ 格が必須格である確率は，任意格や例 3 のような副詞句を排除するため，ガ$_1$ 格が現われたときにさらにガ$_2$ 格が現われる頻度から計算する（式 (3) 参照）．

(4) 分類語彙表を用いて，抽出した名詞を意味クラス（分類番号）に汎化する．
分類語彙表に未登録で分類番号を与えられない場合は，名詞の表記そのものを意味クラスとする．

4 実験

4.1 実験方法

次の 2 種類のモデルを用いてゼロ代名詞の検出，および指示対象の特定に関する比較実験を行なった．

(1) $P_{ana}(a|\phi_c)$

(2) $P_{ana}(a|\phi_c)\cdot P_{zero}(c|v)$

(2) がゼロ代名詞検出の精度向上を目的とした本研究の提案モデルである．

4.2 実験に用いたデータ

3.5 節で説明した共起情報の収集には，毎日新聞 CD-ROM 10 年分（1991～2000 年版）を用いた．

ゼロ代名詞検出および指示対象特定の実験には，京都大学テキストコーパス ver.2.0 [1] 収録の報道記事 30 件を用いた．これらの記事には，あらかじめ人手でゼロ代名詞の出現箇所と照応関係を与えた．

4.3 評価尺度

ゼロ代名詞の検出に関しては，以下の評価尺度を用いた．

$$\text{再現率} = \frac{\text{検出に成功した前方照応ゼロ代名詞数}}{\text{前方照応のゼロ代名詞数}}$$

$$\text{正解率} = \frac{\text{検出に成功した前方照応ゼロ代名詞数}}{\text{システムが検出したゼロ代名詞数}}$$

ゼロ代名詞の指示対象特定に関しては，システムが検出に成功した前方照応のゼロ代名詞を対象に，以下の尺度を用いて評価した．

$$\text{被覆率} = \frac{\text{結果を出力した前方照応ゼロ代名詞数}}{\text{検出に成功した前方照応ゼロ代名詞数}}$$

$$\text{正解率} = \frac{\text{正解した前方照応ゼロ代名詞数}}{\text{結果を出力した前方照応ゼロ代名詞数}}$$

4.4 実験結果

(a) ゼロ代名詞検出に関する評価

まず，式 (2) の確信度を用いた出力制限を全く行わない場合のゼロ代名詞検出の実験結果を表 1 に示す．再現率に関しては 90.0%と比較的良好な結果が得られたものの，正解率は 25.9%にとどまった．

次に，ゼロ代名詞検出の正解率を向上させるため，式 (2) の確信度を用いて出力数を制御した．すなわち，確信度が閾値以下のゼロ代名詞を削除した．閾値を 0.0018 としたときに削除された検出ゼロ代名詞数を表 2 に示す．ここで表右端の「削除率」は，確信度に基づいて削除された検出ゼロ代名詞の割合を「前方照応」「その他の照応」「誤検出」の種別ごとに示しており，仮にゼロ代名詞を無作為に削除した場合は種別によらず同率となる．よって，「前方照応」の削除率が他と比較して小さいほど，本手法が「前方照応」の識別に有効であることを意味する．

表 1: ゼロ代名詞検出の実験結果

全ゼロ代名詞数	検出数	検出成功数	全前方照応数	検出成功数（前方照応）	再現率	正解率
627	1,561	576	449	404	90.0%	25.9 %

表 2: 削除された検出ゼロ代名詞数（閾値=0.0018）

モデル	種別	削除前の数 (a)	削除後の数 (b)	削除率 ((a-b)/a)
(1)	前方照応	404	189	53.2%
	その他の照応	172	56	67.4%
	誤検出	985	154	84.4%
(2)	前方照応	404	183	54.7%
	その他の照応	172	54	68.6%
	誤検出	985	88	91.1%

表 2 のモデル (1) では，「前方照応」のゼロ代名詞は「その他の照応」「誤検出」のゼロ代名詞と比較して削除率が 14〜30 ポイントほど低かった．モデル (2) では，「前方照応」と「誤検出」の削除率の差がさらに広がり，前方照応の識別における本手法の有効性が示された．

次に，閾値を変えながらゼロ代名詞検出の再現率と適合率の関係を調べた結果を図 2 に示す．

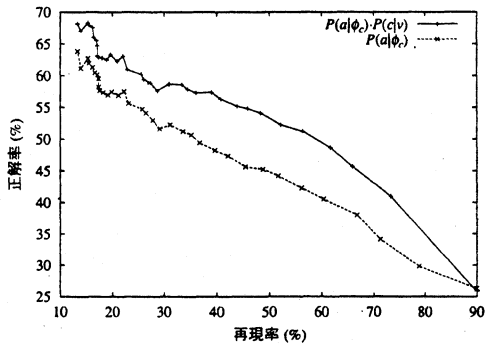

図 2: ゼロ代名詞検出の再現率と正解率の関係

図 2 より，前方照応のゼロ代名詞検出に関して，提案モデル (2) では再現率によらずにモデル (1) 以上の正解率を得ることができた．

(b) ゼロ代名詞の指示対象特定に関する評価

式 (2) の確信度に対する閾値を変えながら，指示対象特定の正解率と被覆率の関係を調べた（図 3）．

図 3 より，格 c がゼロ代名詞化している確率を考慮したことによる指示対象特定処理への悪影響（副作用）は見られない．すなわち，指示対象特定処理の精度を低下させずにゼロ代名詞の検出精度を向上できた．

5 おわりに

ゼロ代名詞の検出は，ゼロ代名詞の照応解消処理を実用化する上で非常に重要な問題である．本研究は日本語ゼロ代名詞の検出と指示対象特定処理を統合して照応解消を行う手法を提案した．新聞記事を対象に評価実験を行なった結果，本手法が前方照応のゼロ代名詞と（誤検出を含む）その他の照応との識別に有効であることが分かった．さらに，指示対象特定処理においてもゼロ代名詞検出処理の統合による副作用は見られず，指示対象特定の精度を低下させることなくゼロ代名詞の検出精度を向上できることが確認された．

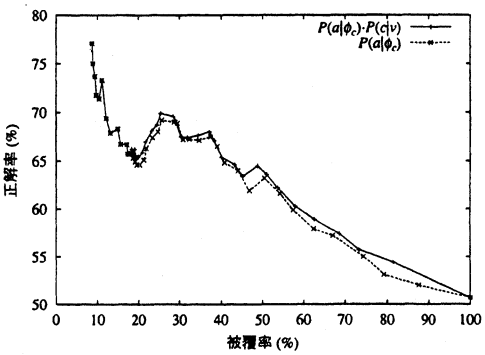

図 3: 指示対象特定の正解率と被覆率の関係

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] Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. A probabilistic model for Japanese zero pronoun resolution integrating syntactic and semantic features. In *Proceedings of the 6th Natural Language Processing Pacific-Rim Symposium*, pp. 403–410, 2001.

[3] 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.

[4] 国立国語研究所（編）. 分類語彙表. 秀英出版, 1964.

[5] 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.

[6] 関和広, 藤井敦, 石川徹也. 確率モデルに基づく日本語ゼロ代名詞の照応解消. 言語処理学会第 7 回年次大会発表論文集, pp. 510–513, 2001.

[7] 中岩浩巳, 池原悟. 日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析. 情報処理学会論文誌, Vol. 34, No. 8, pp. 1705–1715, 1993.

[8] 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.