確率モデルを用いた日本語ゼロ代名詞の照応解析

関 和広[†]　　藤井 敦[††]，[†††]　　石川 徹也[††]

日本語では，読み手や聞き手が容易に推測できる語は頻繁に省略される．これらの省略を適切に補完することは，自然言語解析，とりわけ文脈解析において重要である．本論文は，日本語における代表的な省略現象であるゼロ代名詞に焦点を当て，確率モデルを用いた照応解析手法を提案する．本手法では，学習を効率的に行なうため，確率モデルを統語モデルと意味モデルに分解する．統語モデルは，ゼロ代名詞の照応関係が付与されたコーパスから学習する．意味モデルは，照応関係が付与されていない大規模なコーパスを用いて学習を行ない，データスパースネス問題に対処する．さらに本手法では，照応解析処理の精度を高めるために確信度を定量化し，正解としての確信が高いゼロ代名詞のみ選択的に結果を出力することも可能である．新聞記事を対象にした照応解析実験を通して本手法の有効性を示す．

キーワード:　照応解析, ゼロ代名詞, 確率モデル, コーパス, 文脈解析

Japanese Zero Pronoun Resolution using a Probabilistic Model

Kazuhiro Seki[†] and Atsushi Fujii[††], [†††] and Tetsuya Ishikawa[††]

In Japanese, entities which can easily be predicted are often omitted. Identifying appropriate antecedents associated with those ellipses, which is termed "anaphora resolution", is crucial in natural language processing, specifically, a discourse analysis. This paper proposes a probabilistic model to resolve zero pronouns, which are one of the major ellipses in Japanese. Our proposing model can be decomposed into two models associated with syntactic and semantic properties, so as to optimize a parameter estimation. A syntactic model is trained based on corpora annotated with anaphoric relations. However, a semantic model is trained based on a large-scale unannotated corpora to counter the data sparseness problem. We also propose a notion of certainty to improve the accuracy of zero pronoun resolution. We show the effectiveness of our method by way of experiments.

KeyWords:　*anaphora resolution, zero pronouns, probabilistic models, corpora, discourse analysis*

[†] 産業技術総合研究所, National Institute of Advanced Industrial Science and Technology
[††] 図書館情報大学, University of Library and Information Science
[†††] 科学技術振興事業団 CREST, CREST, Japan Science & Technology Corporation

1　はじめに

自然言語解析では，形態素解析，構文解析，意味解析，文脈解析などの一連の処理を通して，入力テキストを目的に応じた構造に変換する．これらの処理のうち，形態素・構文解析は一定の成果を収めている．また，意味解析に関しても言語資源が整ってきており，多義性解消などの研究が活発に行なわれている (Kilgarriff 1998)．しかし，文脈解析は依然として未解決の問題が多い．

文脈解析の課題の一つに，代名詞などの**照応詞**に対する指示対象を特定する処理がある．自然言語では，自明の対象への言及や冗長な繰り返しを避けるために照応表現が用いられる．日本語では，聞き手や読み手が容易に推測できる対象（主語など）は，代名詞すら使用されず頻繁に省略される．このような省略のうち，格要素の省略を**ゼロ代名詞**と呼ぶ．そして，(ゼロ) 代名詞が照応する実体や対象を特定する処理を**照応解析**と呼ぶ．

照応解析は，文間の結束性や談話構造を解析する上で重要であり，また自然言語処理の応用分野は照応解析によって処理の高度化が期待できる．例えば，日英機械翻訳の場合，日本語では主語が頻繁に省略されるのに対し，英語では主語の訳出が必須であるため，照応解析によってゼロ代名詞を適切に補完しなければならない (中岩, 池原 1993)．

照応詞の指示対象は文脈内に存在する場合とそうでない場合があり，それぞれを**文脈照応**(endophora)，**外界照応**(exophora) と呼ぶ．外界照応の解析には，話者の推定や周囲の状況の把握，常識による推論などが必要となる．文脈照応は，照応詞と指示対象の文章内における位置関係によって，さらに二つに分けられる．指示対象が照応詞に先行する場合を**前方照応**(anaphora)，照応詞が指示対象に先行する場合を**後方照応**(cataphora) と呼ぶ．以上の分類を図 1 にまとめる (Halliday and Hasan 1976)．ただし anaphora は endophora と同義的に用いられることもある．

図 1: 照応の分類

照応解析に関する先行研究の多くは前方照応を対象にしている．これらは人手規則に基づく手法と統計的手法に大別できる．

人手規則に基づく手法は，照応詞と指示対象候補の性・数の一致や文法的役割などに着目

した規則を人手で作成し，照応解析に利用する (Brennan, Friedman and Pollard 1987; Hobbs 1978; Kameyama 1986; Mitkov, Belguith and Stys 1998; Okumura and Tamura 1996; Strube and Hahn 1996; Walker, Iida and Cote 1994; 中岩, 池原 1993; 村田, 長尾 1997)．これらの手法では，人間の内省に基づいて規則を作成するため，コーパスに現れないような例外的な言語事象への対処が容易である．その反面，恣意性が生じやすく，また，規則数が増えるにつれて規則間の整合性を保つことが困難になる．

これに対して，1990 年代には，コーパスに基づく統計的な照応解析手法が数多く提案された (Aone and Bennett 1995; Ge, Hale and Charniak 1998; Soon, Ng and Lim 1999; 江原, 金 1996; 山本, 隅田 1999)．これらの手法は，照応関係（照応詞と指示対象の対応関係）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，照応解析に利用する．統計的手法ではパラメータ値や規則の優先度などを実データに基づいて決定するため，人手規則に基づく手法に比べて恣意性が少ない．しかし，モデルが複雑になるほど推定すべきパラメータ数が増え，データスパースネスが生じやすい．

本研究は日本語のゼロ代名詞を対象に，確率モデルを用いた統計的な照応解析手法を提案する．本手法は，統語的・意味的な属性を分割して確率パラメータの推定を効率的に行なう点，照応関係が付与されていないコーパスを学習に併用してデータスパースネス問題に対処する点に特長がある．なお，本研究は日本語に多く現れる前方照応（図 1 参照）を対象とする．

以下，2 章において本研究で提案するゼロ代名詞の照応解析法について述べ，3 章で評価実験の結果について考察し，4 章で関連研究との比較を行なう．

2　本研究で提案するゼロ代名詞の照応解析手法

2.1　システムの概要

本研究で提案するゼロ代名詞照応解析システムの構成を図 2 に示す．以下，この図に沿って処理の流れを説明する．

まず，システムは入力テキストを形態素・構文解析し，述語を中心とした係り受け情報を抽出する．本システムでは，形態素解析に JUMAN (黒橋, 長尾 1998)，構文解析に KNP (黒橋 1998) を用いる．

次に，入力テキスト中の全てのゼロ代名詞を特定する．ここでは，省略されている必須格要素をゼロ代名詞として検出する．具体的には，入力テキスト中の係り受け情報と動詞の格フレームを照合し，入力テキスト中で充足されていない必須格をゼロ代名詞と見なす．

続いて，ゼロ代名詞の指示対象を特定する．検出された各ゼロ代名詞について，テキストにおける前方の文脈から（複合）名詞を指示対象の候補として抽出する．原理的には，ゼロ代名詞の出現箇所以前の全文脈が探索範囲となりうる．しかし，一般的にゼロ代名詞は文章の論理的な構造を無視して極端に離れた対象を照応することは少ない．そこで本手法では，段落が文

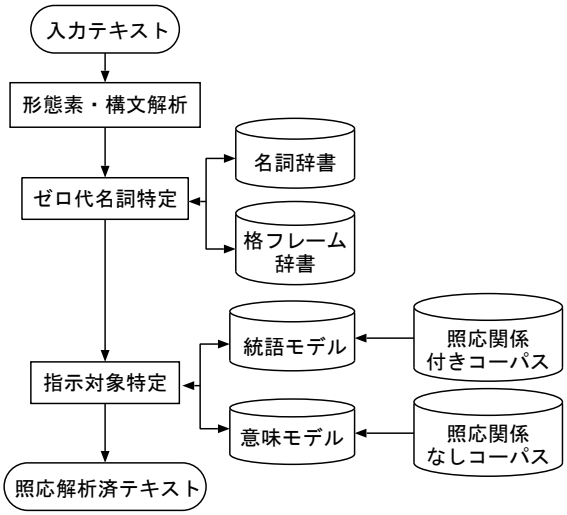

図 2: ゼロ代名詞照応解析システムの構成

章の論理構造に関連することに着目し，ゼロ代名詞の出現箇所から前段落の先頭文までを指示対象候補の探索範囲とする．

最後に，各ゼロ代名詞に対する複数の指示対象候補を尤度に基づいて順位付けし，出力する．本手法では，候補 a_i がゼロ代名詞 ϕ の指示対象である確率 $P(a_i|\phi)$ を計算し，その値によって順位付けを行なう．しかし，$P(a_i|\phi)$ を表層的な情報だけを用いて推定することは困難なので，a_i や ϕ を抽象的な属性で表現する必要がある．

以下，2.2 節でゼロ代名詞の特定方法について述べ，2.3 節でゼロ代名詞 ϕ と指示対象候補 a_i を表現するための属性について説明する．2.4 節以降で提案する確率モデルの詳細とその推定方法について説明する．

2.2　ゼロ代名詞の特定法

本手法では，動詞に関する係受け情報と格フレームを比較することで，充足していない必須格をゼロ代名詞として検出する．

ここでは，格フレーム辞書として IPAL の基本動詞辞書 (情報処理振興事業協会技術センター 1987) とサ変動詞辞書を利用する．IPAL 基本動詞辞書は，和語動詞 861 語を意味的・統語的特性から下位範疇化した 3,379 のサブエントリからなり，動詞あたり平均 3.9 個のサブエントリがある．また，サ変動詞辞書はサ変動詞 50 語に関する 94 のサブエントリからなり，動詞あたり平均 1.9 個のサブエントリがある．以降，両者を合わせて「IPAL 動詞辞書」と呼ぶ．図 3 に，動詞「臨む」の格フレームを例示する．ここで「HUM」や「ORG」などは，それぞれ「人間」や「組織」などの名詞意味素性を表す．

フィールド名	サブエントリ 1	サブエントリ 2	サブエントリ 3
見出し語	のぞむ	のぞむ	のぞむ
表記	臨む	臨む	臨む
意味記述	集会などに出席する	ある重要な場面に出会う	ある所がどこかに面している
格形式 1	ガ	ガ	ガ
意味素性 1	HUM/ORG	HUM/ORG	LOC
格形式 2	ニ	ニ	ニ
意味素性 2	ACT	ABS	LOC
類義語	臨席する，出る	直面する，際する	面する，向く
....	....	....	....

図 3: IPAL 動詞辞書に記述された動詞「臨む」のサブエントリ（抜粋）

IPAL 動詞辞書の検索は次のように行う．まず，システムはテキスト中に現れる全ての動詞について辞書を「表記」で検索し，適合する全ての格フレームを取得する.「表記」で適合する格フレームがない場合は,「見出し語」（読み）で動詞辞書を検索する.「見出し語」でも格フレームが得られない場合は，さらに「類義語」で検索を行なう．これは，類義の動詞は同一の格フレームを持ちやすいという知見に基づく．また，IPAL 動詞辞書に記載された「類義語」は，同一の文脈でサブエントリの動詞と置き換え可能な表現かどうかなどの基準で選定されているため，多くの類義語はサブエントリの動詞と格フレームが一致する．例えば，図 3 のサブエントリ 1 の場合,「臨席する」「出る」はいずれも「HUM/ORG ガ ACT ニ」という格フレームを取り得る．このように「類義語」情報を利用することにより，IPAL 動詞辞書未採録の 2,038 語の動詞（3,150 のサブエントリ）についても，事実上，格フレームの利用が可能となる．

なお,「表記」「見出し語」「類義語」のいずれの検索でも格フレームが得られない場合は，ゼロ代名詞特定の再現率を上げるため，意味素性「不明」のガ格のみを必須格と仮定する．よって，本システムでは対象とする動詞に制限はない.

一方，格フレームに記述された意味素性との照合のため，入力テキスト中の名詞に対しても意味素性を付与する必要がある．ここで，名詞辞書として IPAL 基本名詞辞書 (情報処理振興事業協会技術センター 1996) などを利用することが考えられる．しかし，当該辞書は登録語数が 1,081 語と少ないため，本システムでは分類語彙表 (国立国語研究所 1964) を用いる．分類語彙表には 87,743 語が登録されており，そのうち名詞が 55,443 語含まれる．各登録語には分類番号（意味クラス）が 5 桁の数値で与えられており，名詞の場合，544 種類の意味クラスがある．分類語彙表の構造を図 4 に示す.

なお，一つの名詞が複数の意味クラスに対応する場合，その名詞はいずれの意味クラスにも対応するものとして扱う．また，名詞シソーラスに登録されていない名詞には，一律に「未知語クラス」を与える.

IPAL 動詞辞書の意味素性と分類語彙表の意味クラスとの対応付けは，村田と長尾 (1997) の作成した対応表（表 1）を用いた．この表において，例えば「120」は上位 3 桁が「120」である意味クラス全てに対応する．なお,「未知語クラス」の名詞は格フレームの選択に有効な情報

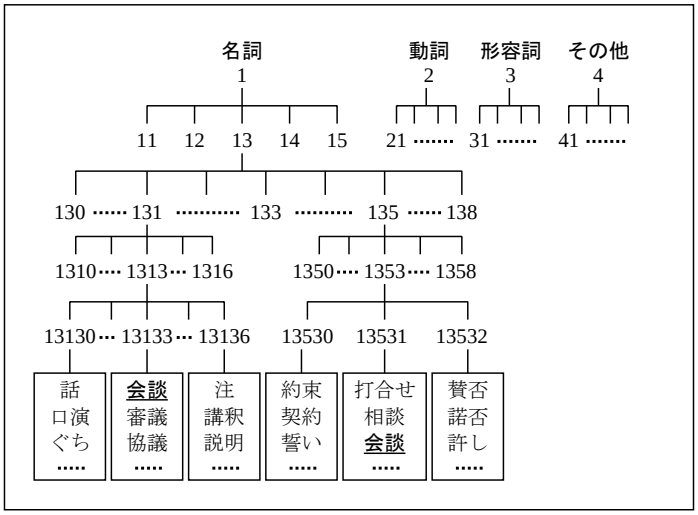

図 4: 分類語彙表の構造

を与えないため，いずれの意味素性にも対応しないものとする．

表 1: 意味素性と意味クラスの対応

IPAL 基本動詞辞書の意味素性	分類語彙表の意味クラス（上位 3 桁）
ANI （動物）	156
HUM （人間）	120, 121, 122, 123, 124
ORG （組織・機関）	125, 126, 127, 128
PLA （植物）	155
PAR （生物の部分）	157
NAT （自然物）	152
PRO （生産物・道具）	14
LOC （空間・方角）	117, 125, 126
PHE （現象名詞）	150, 151
ACT （動作・作用）	133, 134, 135, 136, 137, 138
MEN （精神）	130
CHA （性質）	112, 113, 114, 115, 158
REL （関係）	111
LIN （言語作品）	131, 132
TIM （時間）	116
QUA （数量）	119
CON （具体物）	11, 125, 126, 13, 158
ABS （抽象物）	12, 14, 152, 155, 156, 157
DIV （制限緩やか）	1

次の例を用いて，ゼロ代名詞の特定処理を説明する．

（例 1） 会談に 臨む ことは党内に強い異論のある並立制受け入れとなるため，激しい反発を呼ぶのは必至である．

下線部「臨む」には図 3 に示す 3 通りの格フレームが対応する．そこで，最適な格フレームを

選択するために，ニ格の「会談」をそれぞれの格フレームのニ格と比較する．ここで,「会談」は図 4 より意味クラス「13133」「13531」に対応するので，表 1 より意味素性「ACT」に対応する．よって，図 3 のサブエントリ 1 が選択される．その結果，例 1 中でガ格が省略されていることが分かるので，ガ格をゼロ代名詞として検出する．なお，一つの名詞が複数の意味クラスに対応する場合，その名詞はそれぞれの意味クラスが対応する意味素性全てに対応するものとして扱う．また，充足格を利用しても複数の格フレームが候補として残る場合は，残った候補のうち IPAL 基本動詞辞書で先に記載されている格フレームを選択する．本来ならば，動詞の多義性解消などによって最適な格フレームを選択すべきである．しかし，多義性解消はそれ自身で非常に難しい研究課題であるので，本研究では扱わなかった．

2.3　ゼロ代名詞と指示対象候補を表現する属性

日本語ゼロ代名詞の照応解析に関する先行研究では，指示対象候補に後接する助詞や指示対象候補とゼロ代名詞間のテキスト内での距離などの属性によってゼロ代名詞や指示対象が表現されている．特に，助詞は焦点（話題の中心）の推移をモデル化するセンタリング理論 (Grosz, Joshi and Weinstein 1995) に基づく照応解析手法において中心的な役割を果たす (Kameyama 1986; Walker et al. 1994)．他にも，指示対象候補とゼロ代名詞間の意味的な整合性や語の頻度などが属性として利用されている．本研究では，コーパスに基づく予備調査を通して照応解析に有効な属性について検討し，以下に示す 6 つの属性を用いてゼロ代名詞 ϕ と指示対象候補 a_i を表現する．

- ゼロ代名詞 ϕ に関する属性
  - **格 (c)**：　本研究では，日本語に多く現れる「ガ」「ヲ」「ニ」格のゼロ代名詞を扱う．よって，格 c として取り得る値は「ガ」「ヲ」「ニ」のいずれかであり，それぞれどの格が省略されたのかを表す．この値は，ゼロ代名詞の特定処理時に格フレーム辞書に基づいて決定される（2.2 節参照）．
  - **意味素性 (s)**：　ゼロ代名詞，すなわち省略された格要素に対応する意味素性を表す．本手法では，IPAL 動詞辞書で定義される 19 種類の意味素性を用いる（表 1 参照）．上記の「格」属性と同様に，ゼロ代名詞の特定処理において格フレーム辞書を参照することで決定される．
- 指示対象候補 a_i に関する属性
  - **助詞 (p_i)**：　候補 a_i に後接する助詞を表し，可能な値は形態素解析で助詞と品詞付けられた語の全てである．助詞は従来の手法でも照応解析の有効な手がかりとして用いられている．
  - **文間距離 (d_i)**：　ゼロ代名詞と候補 a_i 間の文数を表す．両者が同一文中にあれば 0，指示対象候補がゼロ代名詞より n 文前にあれば n $(n > 0)$ とする．一般に，ゼロ代名詞からの距離が遠い候補ほど，指示対象になりにくい．なお，距離を計る単位としてゼ

ロ代名詞と候補 a_i 間の語数や文節数なども考えられる．しかし，日本語は語順が比較的自由であり，また副詞句や形容詞句など様々な挿入句が可能であるため，文間距離を用いて大まかな距離を示し，より詳細には助詞 p_i で区別する方法を採用する．

連体節に関する制約 (r_i)：　候補 a_i が連体修飾節に含まれるかどうかを表す．含まれれば真，含まれなければ偽の 2 値を取る．連体節に含まれる名詞句は照応されにくいという知見 (江原, 金 1996) を利用するために導入する．

意味クラス (n_i)：　候補 a_i の意味素性を表す．本手法では，分類語彙表 (国立国語研究所 1964) で定義される 544 種類の分類番号を利用する．

2.4　確率モデル

ゼロ代名詞 ϕ が候補 a_i を照応する確率を $P(a_i|\phi)$ と定義する．ここで，a_i と ϕ を 2.3 節で述べた属性で表現すると式 (1) が成り立つ．

$$P(a_i|\phi) = P(p_i, d_i, r_i, n_i|c, s) \tag{1}$$

式 (1) は推定すべきパラメータ数が膨大であり，大量の学習データを必要とする．しかし，照応関係を付与したコーパスの作成は高価であり，学習に十分な大きさのコーパスを作成するのは現実的ではない．そこで，モデルの妥当性を保持しつつ確率値の推定を容易にするため，以下の近似を行なう．

コーパス分析に基づく我々の予備調査によると，候補 a_i に関する属性のうち，文間距離 d_i，連体節に関する制約 r_i は，それ以外の属性との関連が比較的低い．そこで，d_i と r_i の独立性を仮定すると，式 (2) が得られる．

$$P(a_i|\phi) \approx P(p_i, n_i|c, s) \cdot P(d_i) \cdot P(r_i) \tag{2}$$

$P(p_i, n_i|c, s)$ において，助詞 p_i と格 c はそれぞれ指示対象候補とゼロ代名詞の統語属性であり，意味クラス n_i と意味素性 s はそれぞれ指示対象候補とゼロ代名詞の意味属性である．そこで，統語属性と意味属性間の独立性を仮定すると，式 (3) が得られる．

$$P(a_i|\phi) \approx P(p_i|c) \cdot P(n_i|s) \cdot P(d_i) \cdot P(r_i) \tag{3}$$

式 (3) の右辺において，統語属性だけからなる要素 $P(p_i|c) \cdot P(d_i) \cdot P(r_i)$ を**統語モデル**，意味属性だけからなる要素 $P(n_i|s)$ を**意味モデル**と呼ぶことにする．

式 (3) のそれぞれのパラメータは，照応関係が付与されたコーパスから得られる頻度情報を用いて，式 (4) によって計算できる．ここで，$F(x)$ はコーパスにおける事象 x の出現頻度を

示す.

$$\begin{aligned} P(p_i|c) &= \frac{F(p_i,c)}{\sum_j F(p_j,c)} \\ P(n_i|s) &= \frac{F(n_i,s)}{\sum_j F(n_j,s)} \\ P(d_i) &= \frac{F(d_i)}{\sum_j F(d_j)} \\ P(r_i) &= \frac{F(r_i)}{\sum_j F(m_j)} \end{aligned} \tag{4}$$

ただし，意味モデル $P(n_i|s)$ の推定については，2.5 節において照応関係が付与されていないコーパスの利用を検討する.

2.5　照応関係付きコーパスを必要としない意味モデルの推定法

データスパースネス問題を避けるため，2.4 節で式 (1) の確率モデルを統語モデルと意味モデルに分解した．しかし，式 (4) において，意味クラス n_i と意味素性 s の全ての組み合わせに対して意味モデル $P(n_i|s)$ を正しく推定するには，なお大量の学習データが必要であり，データスパースネスを生じやすい．そこで本研究では，照応関係が付与されていないコーパスを利用して意味モデルを推定する手法を提案する.

格要素（名詞）の意味素性 s は，動詞の語義とゼロ代名詞の格によって一意に決まることが多い．これは名詞がゼロ代名詞化されている場合も同様である．しかし，動詞の語義を付与したコーパスは高価であるため，ここではさらに次の近似を行なう．すなわち，動詞に多義性がなく，重複する格が係らないと仮定する．すると，意味素性 s は動詞 v とゼロ代名詞の格 c の組み合わせで表現することができ，式 (5) が成り立つ.

$$P(n_i|s) \approx P(n_i|v,c) \tag{5}$$

$P(n_i|v,c)$ は，動詞 v の格 c がゼロ代名詞化しているとき，その指示対象の意味クラスが n_i である確率を表す．しかし，ゼロ代名詞は省略された格要素であるため，動詞 v の格 c に本来埋まるべき名詞とゼロ代名詞の指示対象は同じ意味素性に対応すると考えてよい．そこで，$P(n_i|v,c)$ は $\langle n_i,\, v,\, c\rangle$ という係り受け（共起関係）を用いて推定することができる．すなわち，本手法は照応関係を付与していないコーパスを学習に利用することができる．なお，照応関係を付与していないコーパスの利用可能性は村田と長尾 (1998) も指摘している.

意味モデルの学習は次のように行う．まず，照応関係などの付加情報が与えられていない未解析のコーパスを形態素・構文解析し，その結果得られる係り受け関係に基づいて動詞と格要素の共起を自動的に抽出する．続いて，名詞シソーラス（分類語彙表）を利用して格要素を意味クラスに汎化し，意味クラス・動詞・格の共起関係 $\langle n_i,\, v,\, c\rangle$ を収集する．最後に，共起関係の頻度に基づいて意味モデルを生成する.

なお，意味モデルのパラメータ値 $P(n_i|\,v,c)$ の推定にはデータスパースネス問題を避けるため，線形ディスカウンティング法 (linear discounting) (Ney, Essen and Kneser 1994) を用いる．ただし，動詞 v の格 c に関して何ら共起情報が得られない場合，および候補 a_i が名詞シソーラスに未登録で「未知語クラス」が与えられている場合は，全ての意味クラス（544 種類）に関して等確率を与える．式 (6) に意味モデルの計算式を示す．ここで α は比例定数（ディスカウント係数）である．また，N_0 は動詞 v，格 c が与えられたときに，$F(v,c)>0$ かつ $F(n_i,v,c)=0$ であるような意味クラス n_i の総数を表す．

$$P(n_i|v,c)=\begin{cases}\alpha\cdot\dfrac{F(n_i,v,c)}{F(v,c)} & if\ F(n_i,v,c)>0\\ \dfrac{(1-\alpha)}{N_0} & else\ if\ F(v,c)>0\ \text{かつ}\ F(n_i,v,c)=0\\ \dfrac{1}{544} & else\ if\ F(v,c)=0\ \text{または}\ n_i=\text{未知語クラス}\end{cases}\tag{6}$$

候補 a_i に複数の意味クラスが与えられている場合は，候補 a_i と各意味クラスが等確率に対応すると仮定する．すなわち，全ての意味クラスについて意味モデルのパラメータ値を計算し，その平均を候補 a_i の意味モデルのパラメータ値とする．

以上の操作によってモデルを構築することができたものの，動詞の多義性を無視することは言語学的には必ずしも妥当ではない．しかし，3 章の評価実験において，本手法は照応関係を付与したコーパスを用いた場合よりも，ゼロ代名詞の照応解析において有効であることを示す．

2.6　照応解析に関する確信度

照応解析の結果を機械翻訳など他の処理に応用する場合，照応関係の特定を誤ると後続の処理にも悪影響を及ぼす．このような場合，テキスト中の全てのゼロ代名詞を処理することよりも，誤りを犯さないように確実な結果だけを出力することが重要である．言い換えれば，照応解析の被覆率（coverage）よりも精度（accuracy）が重視される場合がある．

照応解析処理の精度を向上させるためには，解析結果が正解である確信が高いゼロ代名詞だけを出力すればよい．そこで，確信度を定量化し，その値が一定の閾値よりも大きい場合だけ結果を出力する（原理的に被覆率は低下する）．具体的には，式 (3) で定義される確率スコア $P(a_i|\phi)$ を利用し，以下の特性 (a) と (b) に基づいて確信度を計算する．なお，$P_j(\phi)$ は j 番目に大きい確率スコアとする．

(a)　$P_1(\phi)$ $(=\max_i P(a_i|\phi))$ が大きいほど確信度が高い
(b)　$P_1(\phi)$ と $P_2(\phi)$ の差が大きいほど確信度が高い

式 (7) に確信度 $C(\phi)$ の計算式を示す．

$$C(\phi)=t\cdot P_1(\phi)+(1-t)(P_1(\phi)-P_2(\phi))\tag{7}$$

ここで，右辺の第 1，2 項はそれぞれ上記 (a) (b) に対応し，t は両者の影響を制御する定数である．

3　評価実験

3.1　実験方法

2 章で提案した照応解析手法の有効性を実験によって評価した．実験には，京都大学テキストコーパス ver. 2.0 (Kurohashi and Nagao 1998) を利用した．当コーパスは，毎日新聞 1995 年版の報道記事と社説記事を各 1 万文ずつ JUMAN と KNP（本システムで用いた形態素・構文解析器）で解析し，その結果を人手で修正したものである．図 5 に京都大学テキストコーパスの一部を示す．図 5 において，第 1 行目は文 ID であり,「*」で始まる行が文節の先頭，行末の「3D」は文節「3」に係ることを示している．それ以外の行は文節に含まれる形態素情報である．

```
# S-ID:950111045-020 KNP:97/09/05
* 0 3D
次に つぎに * 副詞 * * *
* 1 3D
教育 きょういく * 名詞 サ変名詞 * *
委員会 いいんかい * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 3D
注文 ちゅうもん * 名詞 サ変名詞 * *
が が * 助詞 格助詞 * *
* 3 -1D
ある ある ある 動詞 * 子音動詞ラ行 基本形
。 。 * 特殊 句点 * *
```

図 5: 京都大学テキストコーパスの一部

当該コーパスから社説記事 30 件，報道記事 30 件を無作為抽出し，ゼロ代名詞の照応関係を人手で付与し，正解セットを作成した．なお，社説と報道記事は文体等の違いにより照応関係にも顕著な差があると考え，両者を区別して実験に用いた．実験に用いたコーパスの特徴を表 2 に示す．

表 2 に示されるように，新聞記事には記事種によらず前方照応のゼロ代名詞が最も多く現われた．特に，報道記事では 7 割以上の照応が名詞を指す前方照応であった．一方，社説記事では照応の種類ごとのゼロ代名詞数が比較的分散しており，文章外や後方に対しても報道記事より多くの照応表現が用いられていることが分かる．

照応関係を付与したコーパスに対して，2.2 節で述べた方法によりゼロ代名詞特定処理を行なった．結果を表 3 に示す．

表 2: 照応の種類ごとのゼロ代名詞数

記事種	記事数	文数	ゼロ代名詞数（割合 [%]）				
			外界照応	後方照応	前方照応		計
					名詞	文や節	
社説	30	867	371 (33.5)	48 (4.3)	627 (56.6)	62 (5.6)	1,108
報道	30	423	157 (25.0)	7 (1.1)	449 (71.6)	14 (2.2)	627
計	60	1,290	528 (30.4)	55 (3.2)	1076 (62.0)	76 (4.4)	1,735

表 3: ゼロ代名詞特定処理の結果

記事種	全てのゼロ代名詞 (a)	特定数 (b)	特定成功数 (c)	再現率 (c/a)	適合率 (c/b)	評価対象数
社説	1,108	2,141	968	87.4%	45.2%	498
報道	627	1,130	553	88.2%	48.9%	355
計	1,735	3,271	1,521	87.7%	46.5%	853

人手で特定した全てのゼロ代名詞に関して，全体で 87.7%のゼロ代名詞がシステムによって特定された．しかし，特定に成功したゼロ代名詞数 (c) に比べ，システムが特定したゼロ代名詞数 (b) が倍以上あり，適合率は全体で 46.5%とやや低い．ゼロ代名詞特定処理の精密化は，今後の課題である．

本実験の焦点は，ゼロ代名詞の出現箇所特定ではなく照応解析にある．そこで以下の実験では，システムが特定に成功したゼロ代名詞のうち前方の名詞を照応するゼロ代名詞 853 箇所だけを評価の対象とする．すなわち，照応解析処理はゼロ代名詞特定処理と区別して評価する．

評価実験には次のような交差確認法を用いた．すなわち，社説記事・報道記事のそれぞれについて，29 記事を確率モデルの学習，残り 1 記事を評価用の入力テキストとして利用し，入力テキストを変えながら同様の試行を 30 回繰り返し，その結果を平均した．

意味モデルの推定に用いる動詞と格要素の共起情報の抽出には，1994〜1999 年の毎日新聞 6 年分に含まれる約 480 万文を用いた．動詞と格要素の共起情報は，新聞記事を JUMAN で形態素解析し，(複合) 名詞を後方最近傍の動詞に係ると仮定して抽出した．ただし，使役・可能・受身文は格の交替が起きるので共起関係の抽出には用いなかった．ここでは，動詞の活用形が未然形であるか語尾が「〜できる」という表層パターンに一致する動詞を使役・可能・受身のいずれかであると見なし，共起関係の抽出から除外した．また，(複合) 名詞に読点が続くと係り先が必ずしも最近傍でないことが多いので，これらも抽出対象から除外した．この結果，25,640 の動詞について，合計約 255 万組の共起関係 $\langle n_i, c, v \rangle$ (2.5 節参照) が抽出された．

照応解析の評価尺度として，式 (8) に示す正解率と被覆率を用いた．

$$\begin{aligned} \text{正解率} &= \frac{\text{正しく照応解析されたゼロ代名詞数}}{\text{結果を出力したゼロ代名詞数}} \\ \text{被覆率} &= \frac{\text{結果を出力したゼロ代名詞数}}{\text{評価の対象としたゼロ代名詞数}} \end{aligned} \tag{8}$$

ここで,「評価の対象としたゼロ代名詞数」は自動検出に成功した前方照応のゼロ代名詞数を指す．また,「結果を出力したゼロ代名詞数」は通常「評価の対象としたゼロ代名詞数」と同一であり，2.6 節で述べた確信度を用いてシステム出力を制限した場合だけ減少する．

3.2 実験結果

本研究で提案する確率モデルの有効性を示すため，以下の異なる手法（モデル）を比較評価した．なお「組合せ 2」が本研究の提案手法に相当する．

- 統語モデルのみ利用（「統語」）
- 式 (4) による意味モデルのみ利用（「意味 1」）
  ここでは学習用の 29 記事から意味モデルを生成し，照応解析に利用した．
- 式 (5) による共起情報を用いた意味モデルのみ利用（「意味 2」）
  ここでは上記 29 記事は利用せず，毎日新聞から抽出した共起情報 $\langle n_i, c, v \rangle$ のみを照応解析に利用した．
- 統語モデルと意味モデル 1 の組合せを利用（「組合せ 1」）
- 意味モデルと統語モデル 2 の組合せを利用（「組合せ 2」）
- 人手規則に基づくモデルを利用（「規則」）

評価のベースラインとして，人手規則に基づくモデル（上記「規則」）を用意した．これは，京都大学テキストコーパスから抽出した社説記事 10 記事を訓練データとして，約 2 人月で人手作成したモデルであり，主に，a) 指示対象候補に後接する助詞，b) ゼロ代名詞と指示対象候補間の距離（文数），c) ゼロ代名詞と指示対象候補間の意味的整合性に関する規則を利用し，ゼロ代名詞の照応解析を行なう．

各モデルの照応解析結果を表 4 に示す．ここで「1 位」「1-2 位」「1-3 位」は，確率スコア $P(a_i|\phi)$ の値に基づいて該当する上位の結果だけを出力した場合の正解数（正解率）である．例えば,「1-3 位」の場合，上位 3 位中に正解の指示対象が含まれれば「正解」と判定した．また,「正解の平均順位」は入力テキストから抽出された指示対象候補中での正解候補の平均順位を表す．ここで，抽出候補数はモデルによらず社説記事で平均 25.1 個，報道記事で平均 27.3 個であるため，無作為に選択すると正解の平均順位はそれぞれ 12.5 位，13.6 位となる．また，太字の数字は最も正解率の高かったモデルを記事種ごとに示している．以下，表 4 の結果について検討する．

まず，照応関係を付与したコーパスを用いて推定を行なう「意味 1」と，動詞と格要素の共起情報に基づく「意味 2」の結果を比較すると，記事種によらず後者が良い結果を示した．この

表 4: 照応解析の実験結果

記事種	モデル	正解数（正解率 (%)） 1 位	1-2 位	1-3 位	正解の平均順位
社説	統語	173 (34.7)	247 (49.6)	300 (60.2)	4.8
	意味 1	124 (24.9)	195 (39.2)	248 (49.8)	7.2
	意味 2	145 (29.1)	214 (43.0)	250 (50.2)	6.0
	組合せ 1	186 (37.3)	260 (52.2)	307 (61.6)	4.6
	組合せ 2	**198 (39.8)**	**274 (55.2)**	**311 (62.4)**	**4.5**
	規則	180 (36.1)	259 (52.0)	295 (59.2)	5.1
報道	統語	187 (52.7)	222 (62.5)	248 (69.9)	4.1
	意味 1	93 (26.2)	145 (40.8)	186 (52.4)	6.3
	意味 2	114 (32.1)	186 (52.4)	221 (62.3)	5.0
	組合せ 1	173 (48.7)	226 (63.7)	252 (71.0)	4.0
	組合せ 2	**192 (54.0)**	**235 (66.2)**	**268 (75.5)**	**3.2**
	規則	131 (36.9)	185 (52.1)	222 (62.5)	5.0

理由として次の三点が考えられる.

(a) 大規模なコーパスを学習に用いたことで，データスパースネスを解消することができた.

(b) 動詞と格の組合せで意味素性を表すことにより，IPAL 動詞辞書における意味素性分類の粒度の粗さを補うことができた.

(c) 「意味 1」では，IPAL 動詞辞書を用いて格フレームを取得できない動詞に関してゼロ代名詞の意味素性を決定できない．これに対して,「意味 2」は動詞 v と格 c に基づく意味素性 s の近似によって意味モデル $P(n_i|s)$ を推定できる.

ここで (c) は，IPAL 動詞辞書から格フレームを得ることができない場合にだけ影響する．よって，IPAL 動詞辞書から格フレームを得ることができた場合だけを対象に照応解析実験を行なえば，上記 (a) (b) に挙げた理由を裏付けることができる（ただし，それぞれの寄与の程度は測定できない）．この実験を行なって「意味 1」と「意味 2」の結果を比較したところ，1 位の正解率が社説記事で 24.8%から 30.5%に，報道記事で 27.2%から 33.2%にそれぞれ向上した．すなわち，ゼロ代名詞に意味素性が与えられている場合だけを対象としても「意味 2」は「意味 1」より高い精度が得られ，上記の理由 (a) (b) の正当性が示された.

一方,「統語」モデルは「意味 2」と比較してさらに 5〜20 ポイント程度高い正解率が得られた．この結果から，本手法で利用した意味的属性に比べ，助詞や距離などの統語的属性が照応関係の特定により有効であることが分かる.

さらに，両属性を組み合わせた場合（組合せ 1，組合せ 2）は，統語的・意味的属性をそれぞれ単独で用いるよりもほとんどの場合に正解率が向上した．唯一「組合せ 1」を用いて報道記事を照応解析した場合，1 位の正解率が 52.7%から 48.7%に低下しているものの，正しい指示対象の「平均順位」は 4.1 位から 4.0 位に向上しており，全体としては正しい指示対象が候補群の中で上位に移動している．一方，共起情報を用いた意味モデルと統語モデルを組み合わせた場合（組合せ 2），社説記事，報道記事とも唯一解の正解率がそれぞれ，39.8%，54.0%で最大

となった．これらの結果から，統語的・意味的属性が相補的に機能し，両属性を複合的に利用することがゼロ代名詞の照応解析に有効であることが分かる．

続いて「組合せ 2」と人手規則によるモデル「規則」の結果を比較すると，前者の方が社説記事で約 2〜3 ポイント，報道記事では 10 ポイント以上高い正解率が得られた．報道記事に関する正解率の向上がより大きいのは，照応解析規則作成時の訓練データとして社説記事を利用したため，作成した規則やその適用順序が報道記事ではうまく機能しなかったためである．このように「組合せ 2」を用いた場合は，両記事種において「規則」以上の性能を示し，本手法が処理対象テキストの分野の変化に対しても頑健であることが分かる．

照応関係を付与したコーパスは一般に高価であるため，統計的手法では学習データ量と正解率の関係が重要である．本手法は統語モデルの推定に人手で作成したコーパスを利用しているため，まず統語モデルの推定に用いる学習データ量と照応解析の正解率の関係を調査した．図 6 は，モデル「組合せ 2」において学習データ量を 0〜29 記事まで変化させたときの 1 位の正解率の変化を示したものである．学習データ量が 0 記事の場合は，共起情報を用いた意味モデルのみで照応解析を行なう（「意味 2」に相当）．

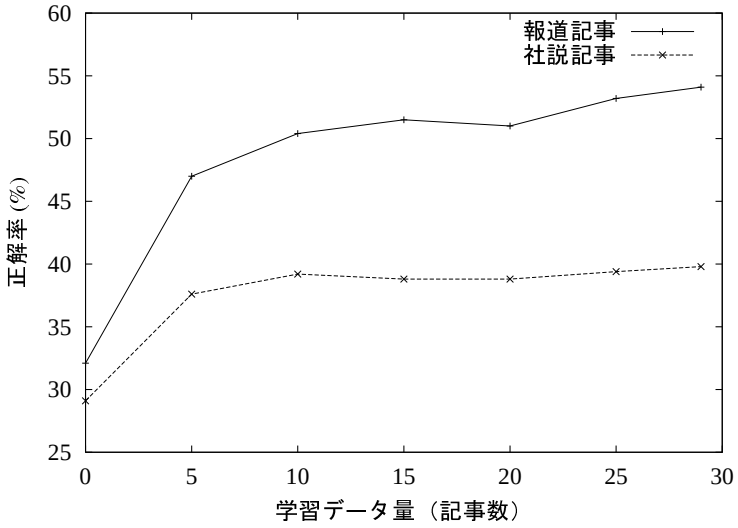

図 6: 統語モデルの学習データ量と正解率の関係（「組合せ 2」の結果）

図 6 を見ると，記事種によらずに学習データ量の増加とともに正解率が向上し，特に 0〜10 記事程度の学習データを用いた場合の立ち上がりが顕著であった．

続いて,「組合せ 2」において統語モデルの学習データ量を 29 記事に固定し，意味モデルの推定に用いる新聞記事の量を 1〜6 年分まで変化させた場合の正解率の変化を図 7 に示す．社説記事，報道記事ともに，共起情報の獲得に利用する新聞記事の量とともに緩やかに正解率が向上した．新聞記事 5〜6 年を使った場合でも正解率が微増していることから，さらに新聞記事の

量を増やすことで正解率の向上が期待できる．なお，意味モデルの学習に利用する新聞記事には形態素・構文情報や照応関係を人手で与える必要がないため，学習データの追加に伴うコストは低い点に注意を要する．

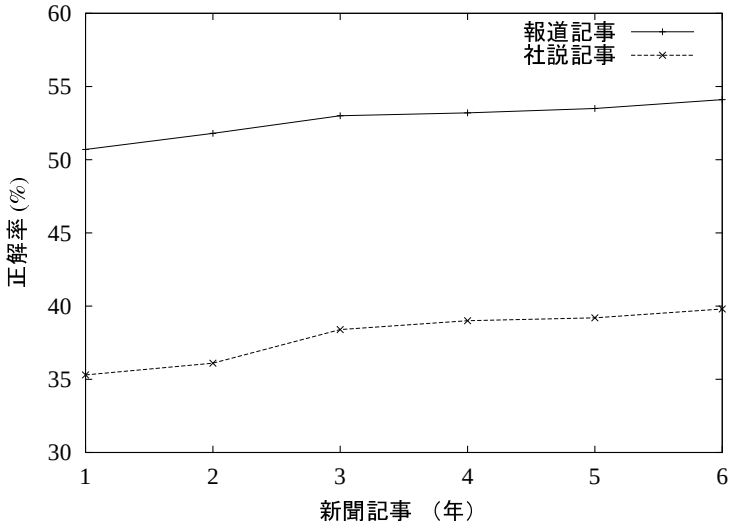

図 7: 意味モデルの学習データ量と正解率の関係（「組合せ 2」の結果）

最後に，確信度 $C(\phi)$ に関する評価実験を行なった（2.6 節参照）．すなわち，各ゼロ代名詞について確信度が閾値以上の場合だけ結果を出力し，閾値を変化させながら照応解析の被覆率と正解率の関係を調査した．その結果を図 8 に示す．なお，式 (7) 中の定数 t は，本実験では経験的に 0.5 とした．記事種によらず，被覆率の低下にともなって正解率が向上し，被覆率 10%以下で共に 70%以上の正解率が得られた．この傾向は報道記事の場合，特に顕著だった．この結果より，本研究で提案した確信度が照応解析処理の正解率向上に有効であることが確かめられた．

3.3　考察

本手法で提案した確率モデルでは，動詞に多義性がないと仮定し，ゼロ代名詞の意味素性 s を動詞 v と格 c で近似することで意味モデルを推定している（2.5 節参照）．また，3.2 節の評価実験では，動詞と格要素の共起情報を用いた意味モデル（「意味 2」）は共起情報を用いないモデル（「意味 1」）よりも正解率が高いものの，統語モデルと比べると一貫して正解率が低かった．そこで本節では，動詞と格による意味素性の近似が照応解析に与える影響を調査するため，意味モデルのパラメータ値に注目して照応解析の誤り例を分析した．

提案手法（「組合せ 2」）を用いて照応解析を行なった場合に，正しい指示対象（以下，正解候補と呼ぶ）が最上位に順位付けられたゼロ代名詞は，社説記事と報道記事を合わせて 390 件

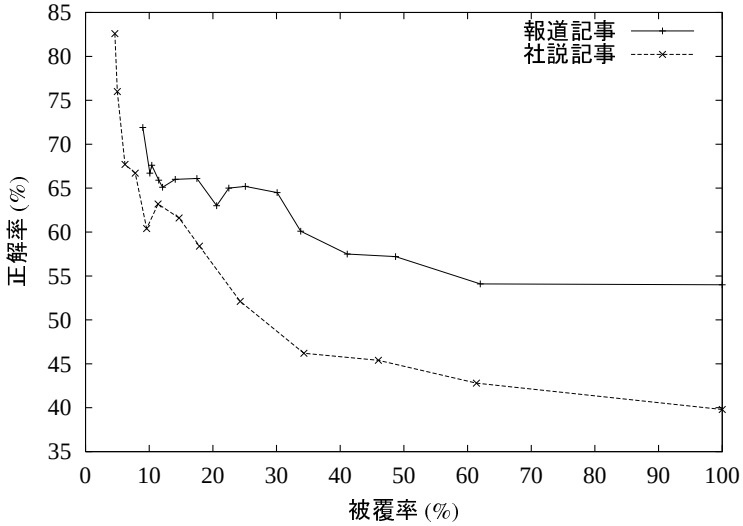

図 8: 確信度に関する閾値を変化させた場合の被覆率と正解率の関係

であった（表 4 参照）．ここでは，それ以外の 463 $(= 853 - 390)$ 件を誤りと見なし，そこから無作為に抽出した 40 件について誤りの原因を調べた．結果を表 5 に示す.

表 5: 照応解析誤りの内訳

誤りの原因	件数
正解候補が動詞の格要素として意味的に整合するのに共起頻度が低い	19 (47.5%)
動詞の格要素として意味的に整合する指示対象候補が複数ある	13 (32.5%)
多義動詞（語義の弁別が必要）	3 (7.5%)
正解候補の意味クラスがない（分類語彙表に記載されていない）	3 (7.5%)
正解候補が動詞の格要素として意味的に整合しない	2 (5.0%)
計	40

分析の結果，照応解析を誤る主要な原因は次の二点にあり，両者を合わせて分析した事例の 80%を占めた．一つは，正解候補が動詞の格要素として意味的に整合するにもかかわらず，学習データ中での両者の共起頻度が比較的少ないために，意味モデルのパラメータ値が期待される値よりも低くなる場合である．もう一つは，正解候補以外にも意味的整合性が同程度の候補が存在し，かつそれらの候補の統語モデルのパラメータ値が正解候補の統語モデルのパラメータ値よりも大きいために，正解候補の最終的な確率スコア $P(a_i|\phi)$ が相対的に小さくなる場合である．

これに対して，動詞の多義性を考慮しなかったために照応関係の特定を誤った例，すなわち語義の弁別を必要とした例は 40 件中 3 件 (7.5%) であった．このように，動詞の多義性を無視したことによる悪影響は比較的少なく，動詞と格による意味素性の近似が日本語ゼロ代名詞の

照応解析に有効に働くことが実験的に示された.

ゼロ代名詞の照応解析処理精度をさらに向上させるためには，前述の照応解析を誤る主要な二つの原因への対処が重要である．以下，それぞれの原因による誤り例を示す．なお，ゼロ代名詞の出現位置を「ϕ」，ゼロ代名詞を含む動詞を下線，誤って特定された指示対象（誤）と正解候補（正）を太字で示す.

（例 2）　…だが、いつかやって来る地震に備えて、さまざまな手を打つことができる。例えば要注意の活断層の近くでは、新たな開発の規制や建築物の補強、耐震基準の強化などを行うことが重要な課題になる。**自治体**$_{(誤)}$ が詳細な活断層地図を公表し、だれもが自分の家や会社の置かれた状況を理解できるようにすべきではないか。米国カリフォルニア州では活断層周辺の開発を規制し、成果を上げている。世界一の地震国である日本も見習うべきだろう。これまで地震予知や活断層、構造物の安全性などの**専門家**$_{(正)}$ がそれぞれ別々に研究を進めてきた。(ϕ ガ）お互いにデータを 公開し 合うことも少なかった。…

例 2 は，指示対象が意味的に整合するにもかかわらず意味モデルのパラメータ値が低くなる例である． このゼロ代名詞の正しい指示対象は「専門家」であるのに対し，確率スコア $P(a_i|\phi)$ を最大化する候補は「自治体」であった．これは，いずれの語も意味的に容認できるにもかかわらず,「専門家 (意味クラス $n = 12340$)」の意味モデルのパラメータ値 (0.0002) が「自治体 ($n = 12700$)」のパラメータ値 (0.0373) より著しく小さいことに原因がある．これは，今回の実験に用いた学習データにおいては，$\langle 12340, \text{ガ}, \text{公開する} \rangle$ の出現頻度が $\langle 12700, \text{ガ}, \text{公開する} \rangle$ の頻度と比べて極端に低かったことを示している．このように，学習に用いたコーパス中の出現頻度が必ずしも直観的な正しさを表さないという現象は，コーパスを用いた統計的な手法における典型的な問題であり，今後さらに検討しなければならない.

なお，人手作成の規則に基づく手法（「規則」）を用いた場合も，例 2 の指示対象は正しく特定できなかった．これは,「公開し」（基本形 : 公開する）が IPAL 動詞辞書に未記載で意味素性を取得できず，正しい指示対象「専門家」との意味的な整合性を判定できなかったためである.

次に，正しい指示対象以外にも同程度の意味的整合性を持つ候補が存在し，照応解析を誤る例を示す.

（例 3）　…**英国**$_{(誤)}$ はこの間、調整に遅れたため、重化学工業の発展は、ドイツ、米国に先を越され、七つの海を支配した大英帝国の没落に直結するのだ。**日本**$_{(正)}$ も大戦略を立てねば、二十一世紀には、(ϕ ガ）繁栄を 見ないまま 没落する。…

例 3 では，正解候補「日本」に対して「英国」が指示対象として誤って特定される．それぞれの候補の意味モデルのパラメータ値は,「日本」「英国」とも 0.0142 である．また直観的にも，動詞「見ないまま」の格要素としての意味的整合性だけで両者を区別することはできない．このように，意味モデルによって正解を識別することができない場合は，統語モデルの貢献によって高

い確率スコアを与える必要がある．しかし，統語モデルのパラメータ値は「日本」が 0.0056，「英国」が 0.0134 であり，現在のモデルで利用している属性のみで正解候補「日本」に高い確率スコアを与えることは難しい．新たな属性として，例えば重文の接続助詞（例 3 では「(〜立てね）ば」）(中岩, 池原 1996)，前後の動詞の意味的な関連性（例：A が戦略を立てない → A が繁栄を見ない）(中岩, 池原 1993) なども検討する必要がある．

なお「規則」に基づく手法を用いた場合も，例 3 の指示対象を正しく特定することはできなかった．これは,「見ない」（基本形：見る）のガ格の意味素性として，IPAL 動詞辞書の格フレームに「ORG」が含まれておらず，正しい指示対象「日本」と適合しなかったことが原因である．

4　関連研究との比較

江原と金 (1996) は，日英機械翻訳の前編集において，日本語の原文（長文）を短文に分割した際に発生する主格の欠落をゼロ主語（ガ格のゼロ代名詞）と見なし，確率モデルを用いて補完する手法を提案した．彼らは，ニュース原稿 108 文を対象に評価実験を行ない，オープンテストで 80.6%の正解率を得ている．しかし，この手法は同一文内に指示対象がある場合のみが対象であり，それ以前の文脈に指示対象が現れる状況を考慮していない．ゼロ主語の照応先は同文内とは限らず，照応先としてゼロ主語以前の文脈を考慮するほど指示対象候補が増えて照応解析が難しくなる．事実，彼らの実験においてゼロ主語ごとの指示対象候補は平均 3.9 個であり，本研究に比べると極端に少ない（3.2 節参照）．

Aone と Bennett (1995) は，ゼロ代名詞とそれ以外の照応詞（固有名詞，限定詞）を対象に，決定木を用いた照応解析手法を提案した．彼らは，合弁事業に関する新聞記事を用いて評価実験を行ない，ゼロ代名詞に関して，オープンテストで 80%前後の正解率を得ている．しかし，この実験で対象となったゼロ代名詞は，会社名等の組織を照応するものに限定されている．よって，指示対象候補として会社名等のみを考慮すればよく，この制約により大幅に候補を絞り込める．

以上二つの先行研究と比較して，本研究は前方照応のゼロ代名詞全般を対象にしており，適用範囲が広い．また，以上の研究が学習データとして人手で照応関係を付与したコーパスを必要とするのに対し，本提案手法は，照応関係が付与されていないコーパスを併用することで，大規模なコーパスを容易に学習に利用できる．また本手法では，確信度を利用することで正解の確信が高いゼロ代名詞のみ選択的に結果を出力し，利用目的に応じて照応解析の正解率を向上させることができる．

5　おわりに

本論文は，確率モデルを用いて日本語ゼロ代名詞の前方照応解析を行なう手法を提案し，評価実験を通してその有効性を示した．本手法は，ゼロ代名詞と指示対象に関する属性間の依存関係に基づいて確率モデルを意味モデルと統語モデルに分解し，パラメータ推定を効率化する．

統語モデルの推定には，照応関係が付与されたコーパスを学習データとして用い，意味モデルに関しては，動詞と格要素の共起関係を利用することで，照応関係が付与されていないコーパスからの学習を可能にする．また，照応解析の精度向上のために確信度を定量化する手法を提案した．

新聞記事コーパスを用いて統語モデルと意味モデルを個別に評価したところ，統語モデルは意味モデルよりも 5〜20 ポイント程度高い正解率を示した．この結果から，本手法で利用した属性のうち，ゼロ代名詞と指示対象間の統語的な属性が照応解析の手がかりとしてより有効であることが分かった．両モデルを組み合わせて用いる提案手法では，さらに正解率が向上し，統語属性と意味属性が相補的に機能することが分かった．また，人手規則による手法と比べると，提案手法は報道・社説記事のいずれにおいても良い結果を示した．さらに，確信度を用いて選択的に指示対象を出力したところ，正解率のさらなる向上を確認できた．

今後の研究課題として以下の点が残されている．確信度を用いて照応解析を行なった場合，(確信度を用いない) 通常の手法と比較して正解率は向上するものの，高い正解率を得るためには被覆率の低下も大きい．よって，ゼロ代名詞の照応解析をより実用的な処理とするには，高い正解率を保持しつつ被覆率をさらに向上させる必要がある．また，そもそもゼロ代名詞の照応解析を行なうためには，ゼロ代名詞の出現箇所を正確に検出する必要がある．ゼロ代名詞出現箇所の特定には，連体修飾節と係り先の名詞の格関係の解析，述語と格要素の係り受け・格解析などを高精度で実現し，加えて大規模な格フレーム辞書を整備する必要がある．

参考文献

Aone, C. and Bennett, S. W. (1995). "Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies." In *Proceedings of 33th Annual Meeting of the Association for Computational Linguistics*, pp. 122–129.

Brennan, S., Friedman, L., and Pollard, C. (1987). "A Centering Approach to Pronouns." In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155–162.

Ge, N., Hale, J., and Charniak, E. (1998). "A statistical approach to anaphora resolution." In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 161–170.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse." *Computational Linguistics*, **21** (2), 203–226.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman.

Hobbs, J. R. (1978). "Resolving Pronoun References." *Lingua*, **44**, 311–338.

Kameyama, M. (1986). "A Property-Sharing Constraint in Centering." In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 200–206.

Kilgarriff, A. (1998). "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation

Programs." In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pp. 581–588.

Kurohashi, S. and Nagao, M. (1998). "Building a Japanese Parsed Corpus while Improving the Parsing System." In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pp. 719–724.

Mitkov, R., Belguith, L., and Stys, M. (1998). "Multilingual robust anaphora resolution." In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, pp. 7–16.

Ney, H., Essen, U., and Kneser, R. (1994). "On structuring probabilistic dependences in stochastic language modeling." *Computer Speech and Language*, **8** (1), 1–38.

Okumura, M. and Tamura, K. (1996). "Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory." In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 871–876.

Soon, W. M., Ng, H. T., and Lim, C. Y. (1999). "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.

Strube, M. and Hahn, U. (1996). "Functional Centering." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 23–28.

Walker, M., Iida, M., and Cote, S. (1994). "Japanese Discourse and the Process of Centering." *Computational Linguistics*, **20** (2), 193–233.

江原暉将, 金淵培 (1996). "確率モデルによるゼロ主語の補完." 自然言語処理, **3** (4), 67–86.

黒橋禎夫, 長尾真 (1998). 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.

黒橋禎夫 (1998). 日本語構文解析システム KNP version 2.0 b6 使用説明書. 京都大学大学院情報学研究科.

国立国語研究所（編） (1964). 分類語彙表. 秀英出版.

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

情報処理振興事業協会技術センター (1996). 計算機用日本語基本名詞辞書 IPAL (Basic Nouns) 解説編.

中岩浩巳, 池原悟 (1993). "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, **34** (8), 1705–1715.

中岩浩巳, 池原悟 (1996). "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, **3** (4), 49–64.

村田真樹, 長尾真 (1997). "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, **4** (1), 87–109.

村田真樹, 長尾真 (1998). “表層表現と用例を用いた照応省略解析手法.” 言語理解とコミュニケーション研究会 NLC97-57, pp. 9–16.

山本和英, 隅田英一郎 (1999). “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, **6** (1), 3–28.

略歴

関 和広: 2000 年 3 月図書館情報大学卒業．2002 年 3 月，図書館情報大学大学院情報メディア研究科博士前期課程修了．2002 年 4 月産業技術総合研究所情報処理研究部門非常勤研究員．2002 年 9 月から Indiana University, School of Library and Information Science, Doctoral Program に進学予定．自然言語処理に興味を持つ．

藤井 敦: 1993 年 3 月東京工業大学工学部情報工学科卒業．1998 年 3 月同大学大学院博士課程修了．1998 年図書館情報大学助手，現在に至る．博士（工学）．自然言語処理，情報検索，音声言語処理の研究に従事．情報処理学会，人工知能学会，電子情報通信学会，Association for Computational Linguistics 各会員．

石川徹也: 1977 年 3 月慶應義塾大学大学院修士課程（図書館情報学）修了．富士写真フイルム（株）足柄研究所入社，図書館短期大学を経て現在，図書館情報大学教授．工学博士．情報管理システムの高度化の研究に従事．情報処理学会，人工知能学会，ACM 等各会員．