自動構築した格フレーム辞書と先行詞の位置選好順序を用いた省略解析

河原 大輔[†]　　黒橋 禎夫[†]

本稿では、日本語文章中における格要素の省略 (ゼロ代名詞) を検出し、その先行詞を同定する手法を提案する。本手法は、自動構築した格フレーム辞書に基づく格解析によってゼロ代名詞を検出し、同辞書による正確な選択制限を用いてゼロ代名詞の先行詞を同定する。また、先行詞はゼロ代名詞から近いところに存在しやすいという傾向を正確にモデル化するために、文・文章中の構造を考慮した先行詞の位置選好順序をコーパスから学習し、これを解析で利用する。格フレーム辞書、先行詞の位置選好順序、さらに機械学習を統合した省略解析システムを作成し、100 記事の大規模解析実験を行った結果、ゼロ代名詞検出が適合率 87.1%、再現率 74.8%、ゼロ代名詞の先行詞同定が 61.8%の精度であった。

キーワード： 格フレーム, ゼロ代名詞, 省略解析

Zero Pronoun Resolution based on Automatically Constructed Case Frames and Structural Preference of Antecedents

Daisuke Kawahara[†]　　Sadao Kurohashi[†]

This paper describes a method to detect and resolve zero pronouns in Japanese text. We detect zero pronouns by case analysis based on automatically constructed case frames, and rank candidate antecedents of a zero pronoun based on similarity to examples in the case frames. We also introduce an order of antecedent location preference to precisely capture the tendency that a zero pronoun has its antecedent in its close position. Large experimental results on 100 articles indicate that the precision and recall of zero pronoun detection are 87.1% and 74.8% respectively and the accuracy of antecedent estimation is 61.8%.

KeyWords: *case frame, zero pronoun, anaphora resolution*

1　はじめに

省略・照応を高精度に解析する技術は、自動要約、機械翻訳、質問応答などの言語処理アプリケーションを高度化するために必要である。日本語における照応詞は省略されることがほとんどであるので、本論文では、省略された照応詞 (ゼロ代名詞と呼ばれる) に注目し、その解析手法を提案する。省略解析における大きな手がかりとしては、次の 2 つが考えられる。

[†] 東京大学 大学院情報理工学系研究科, Graduate School of Information Science and Technology, The University of Tokyo

選択制限　先行詞は、ゼロ代名詞の文脈、特に関係する述語からの意味的制約をうける。
近距離特性　先行詞は、ゼロ代名詞から近いところにある傾向がある。

選択制限のような意味的手がかりとしては、これまで、人、具体物、抽象物のような粗い意味属性しか利用されていなかった。これは、それ以上に詳細に記述されたリソースが存在しなかったからである。我々は、大規模な格フレーム辞書を自動的に構築しており、そこには個々の単語レベルで詳細な選択制限が記述されている。本研究では、この格フレーム辞書を選択制限に用いる。

また、格フレーム辞書はゼロ代名詞の検出にも不可欠である。格フレームには、用言のとりうる格が記述されているので、入力文の述語項構造と格フレームとのマッチングの結果、格フレームに対応づけられていない格があれば、それがゼロ代名詞であると認識できる。従来の研究では、ゼロ代名詞の検出ができたと仮定し研究対象としていないか、人手で作成した格フレーム辞書を用いてゼロ代名詞を検出している。これと比べて、本研究では、自動構築した大規模格フレーム辞書を用いており、一般の文章に対して実用的に使える省略解析システムを作成することができる。

省略解析のもうひとつの大きな手がかりは近距離特性であり、これを解析に反映しようという試みがこれまでもなされてきた (Aone and Bennett 1995; 吉野圭一, 竹内和広, 松本裕治 2001; 関和広, 藤井敦, 石川徹也 2002)。それらは、ゼロ代名詞と先行詞候補間の距離を機械学習の素性や確率モデルのパラメータとするものである。しかし、これらの手法の大きな問題は、距離を flat な尺度でしか計っていないということである。つまり、ゼロ代名詞と先行詞候補の間の文数や単語数を距離としており、文・文章に当然存在する構造を考慮するようなことを行っていない。本研究では、近距離特性を正確にモデル化するために、まず、従属節、主節、埋め込み文などといった文・文章中の構造を考慮して、ゼロ代名詞に対する先行詞候補の位置を分類する。そして、どの位置にある候補が先行詞となりやすいかを学習コーパスから取得し、それを順に並べることによって先行詞の位置選好順序を得るということを行う。

本研究で提案する省略解析手法は、格フレーム辞書と先行詞の位置選好順序に加えて、先行詞の決定に関与するその他の要因を考慮するために機械学習による分類器を用いる。本手法は、この 3 つを統合的に用いるもので、先行詞の位置選好順に候補を調べ、分類器が正例と分類し、かつ、格フレームによる選択制限を満たす候補を先行詞として決定する。

2　関連研究

省略・照応解析の研究は、人手で作成した規則による手法とコーパスを用いた統計的手法に大別できる。

人手で作成した規則による省略・照応解析手法 (中岩浩巳 池原悟 1993, 1996; 村田真樹 長尾眞 1997) は、照応詞と先行詞候補間の統語的および意味的な制約・選好に着目したルールを作成し、それを適用することにより省略・照応解析を行っている。これらは高い精度を実現している

が、中岩らは新聞記事のリード文や 1 文単位で独立した文、村田らは物語文や新聞記事のコラムなどを対象としているため、一般の文章で利用するには規則の修正が必要であると思われる。

一方、機械学習や確率モデルを用いたコーパスベースの手法が提案されている。機械学習による手法は、照応詞と先行詞候補の間の文数や単語数を距離尺度として機械学習の素性のひとつにしている (Aone and Bennett 1995; 吉野圭一他 2001)。これは前節で述べたように、文・文章中の構造を考慮していないことが問題である。また、これらの手法は、ある範囲の候補の中から、機械学習によって作成した分類器の出力スコアがもっともよいものを先行詞として選択している。しかし、分類器の出力スコアは、候補を独立に見た分類クラスの信頼度であり、候補間の比較に用いるのは直接的でなく根拠が薄いと思われる。

確率モデルによる省略解析手法 (関和広他 2002) では、ゼロ代名詞と先行詞との間の文数をパラメータに組み込んでいる。しかし、パラメータ数の爆発を避けるために、距離属性が他の属性と独立であることの仮定を用いており、この妥当性は不明である。また、距離を文数のみで表しているので、同一文内照応の場合は距離による選好が働かないという欠点がある。

日本語以外の言語でも照応解析は盛んに行われており、コーパスベースの手法が一定の成功を収めている (Ge, Hale, and Charniak 1998; Soon, Ng, and Lim 2001; Müller, Rapp, and Strube 2002; Ng and Cardie 2002)。しかし、これらの手法もまた、距離尺度として、照応詞と先行詞候補の間の文数や単語数などしか用いていない。照応詞と先行詞候補の間の位置関係を構造的に捉えた手法として、Hobbs によるものがある (Hobbs 1978; Ge et al. 1998)。この手法は、構文木中を横断しながら先行詞を探索するモデルである。このモデルは、1 文内の構造によって探索順序を制御しているが、複数文にわたって探索する場合の優先順序は扱っていない。

3　格フレーム辞書の自動構築とそれに基づく省略解析

本研究では、ゼロ代名詞の検出と、ゼロ代名詞の先行詞が満たすべき選択制限に、自動的に構築した格フレーム辞書 (Kawahara and Kurohashi 2002) を用いる。本章では、まず、この格フレーム辞書の構築方法について概略を述べる。次に、格フレーム辞書を用いた格解析、そして、その格解析結果を用いたゼロ代名詞検出手法について説明する。

3.1　格フレーム辞書の自動構築

大規模テキストから格フレーム辞書を自動構築する際の最大の問題は、用言の用法の曖昧性である。つまり、同じ表記の用言でも複数の意味、用法をもち、とりうる格や用例が異なる。例えば、「トラックに 荷物を 積む」と「経験を 積む」は、用言は「積む」で同じであるが用法が異なっている。用法が異なる格フレームを別々につくるために、我々は、格フレーム収集の単位を用言とその直前の格要素の組とした。「積む」の例では、「荷物を積む」「経験を積む」を単位として格フレームを収集する。さらに、「荷物を積む」「物資を 積む」などかなり類似してい

る格フレームをマージするためにクラスタリングを行う。

この格フレーム辞書構築の手順を以下に示す。

1. テキストを構文解析する。
2. 構文解析結果から信頼度の高い述語項構造を抽出する。
3. 抽出した述語項構造を用言とその直前の格要素ごとにまとめ、(最初の) 格フレームをつくる。以後、用言の直前の格要素を「直前格要素」、その格を「直前格」と呼ぶ。
4. 3でつくった格フレームをシソーラスに基づいてクラスタリングし、類似しているものをマージする。
5. 格フレームごとに必須格を選択する。直前格の用例数に対して、閾値以上の用例をもつ格を必須格とする。ただし、ガ格は常に必須格とする。

4のクラスタリングでは、シソーラスとして日本語語彙大系 (NTT 1997) を用い、2用例間の類似度を定義し利用している。2つの用例 e_1, e_2 間の類似度 $sim(e_1, e_2)$ は以下のように定義する。

$$sim(e_1, e_2) = max_{x \in s_1, y \in s_2}\, sim(x, y) \tag{1}$$

$$sim(x, y) = \frac{2L}{l_x + l_y}$$

ここで、x, y は意味属性であり、s_1, s_2 はそれぞれ e_1, e_2 の日本語語彙大系における意味属性の集合である (日本語語彙大系では、単語に複数の意味属性が与えられている場合が多い)。$sim(x, y)$ は意味属性 x, y 間の類似度であり、l_x, l_y は x, y のシソーラスの根からの階層の深さ、L は x と y の意味属性で一致している階層の深さを表す。この類似度 $sim(x, y)$ は 0 から 1 の値をとる。

本研究では、新聞20年分のテキストから自動構築した格フレーム辞書を用いる。この辞書には、約23,000個の用言が含まれており、1用言あたりの格フレーム数は約14.5個である。構築した格フレームの例を表1に示す。表において、<主体>は意味マーカの主体を表し、それをもつ格は人、組織といった主体的要素をとることを示す。格ごとにそれぞれの用例がシソーラス上で主体属性以下にあるかどうかをチェックし、半数以上の用例が主体に属す場合に<主体>を付与している。

3.2　格フレーム辞書を用いた省略解析

本論文で提案する省略解析手法の概略は以下のとおりである。

1. 入力文を構文解析する[1]。
2. 入力文中の各用言について、文頭から順番に以下の処理を行う。
   2.1. 入力用言の述語項構造に合致する格フレームを選択する。
   2.2. 格フレームと入力側の格要素との対応をとる。

[1] 格・省略解析と同時に構文構造を決めることもできるが、本研究では最初に構文構造を決定した。

表 1: 格フレームの例

	格	用例		格	用例
決定 (1)	ガ	<主体>, 政府, 委員会, ···	擁立 (1)	ガ	<主体>, 派, 政党, ···
	ヲ	方針, 案, 計画, ···		ヲ	<主体>, 候補, 候補者
	デ	会, 総会, 閣議, ···		ニ	<主体>, 選挙区, 選, ···
	時間	<時間>	擁立 (2)	ガ	<主体>
決定 (2)	ガ	<主体>, 地裁, 家裁, ···		ヲ	<主体>, 議員, 外相, ···
	ヲ	処分, 措置, 扱い, ···		ニ	<主体>, 候補, 後継, ···
⋮	⋮	⋮	⋮	⋮	⋮

2.3. 格フレーム中で対応づけられていない格をゼロ代名詞と認識する。

2.4. ゼロ代名詞の先行詞を同定する。

手順 2.1、2.2 の処理は格解析であり、どちらにも以下のような難しさがある。

格フレーム辞書には、用言ごとに複数の格フレームが存在するので、入力文の用言の用法にもっとも合致する格フレームを選択する。格フレームの選択を行うためには、入力の述語項構造が用言の用法を決定するのに十分な情報をもつ必要がある。十分な情報がなく、格フレームを選択できない場合は、入力用言のすべての格フレームについて、ゼロ代名詞の先行詞同定までの処理を行い、最後にもっともスコアが高かった格フレームに決定する (6 章参照)。

格要素の格フレームの格スロットへの対応付けは、基本的には一致している格同士を対応付ければよいが、係助詞句や非連体修飾詞の場合には格が明示されていないので問題となる。

以下では、手順 2.1 から 2.3 の処理を説明する。2.4 の先行詞同定処理については 6 章で詳細に述べる。

格フレームの選択

前節で述べたように、用言の用法の決定に対して、用言の直前格要素が重要な役割を果たす。特に、直前格がヲ格、ニ格の場合はその傾向が強い。また、直前格要素が<主体>の場合、例えば「<主体>が　求める」という表現からは、用言の用法が決まらず、格フレームを選択することができない。これらの点を考慮して、格フレームの選択を行う条件を以下のように設定する。

1. 入力側の対象用言が直前格要素 C をもつ。
2. 直前格要素 C と直前格 cm が以下のいずれかの条件を満たす。
   - cm がヲ格、ニ格のいずれかである。
   - cm がヲ格、ニ格以外で、C が意味属性<主体>をもたない。
3. cm をもつ格フレームが存在し、cm の用例群と C の類似度が閾値以上ある。

...

石原知事が再選を目指して、知事選への立候補を表明した。

自民党は支持する方針を決定したが、民主党は独自候補を 擁立する ことを検討している。

...

図 1: 記事の例

条件 3 を満たす格フレームのなかで、もっとも類似度が高い格フレームを選択する。もっとも類似度が高い格フレームが複数存在するときは、格フレームを選択できない場合と同様に、それらの格フレームそれぞれについて後続の処理を行った後に格フレームを決定する。ここで用いる類似度は、直前格要素と格フレームの直前格の各用例との類似度のうちもっとも高いものとする。用例間の類似度は (1) 式を用いて計算する。

例として、図 1 の 2 文目の「擁立する」を考える。「擁立」に対して表 1 のような格フレームがある。入力側の表現「候補を　擁立する」は上記の条件 1, 2 を満たし、格フレーム「擁立 (1)」が条件 3 を満たすので、格フレーム「擁立 (1)」が選択される。

入力側格要素と格フレームの格との対応付け

選択された格フレームについて、入力側の格要素と格フレームの格との対応づけを行う (Kurohashi and Nagao 1994)。格要素に格助詞が付属している場合は、その格助詞の格に対応する格フレーム側の格に対応づける。被連体修飾詞や係助詞句のように、文中から格がわからない場合は、次表の格それぞれに対応させ、対応づけ全体のスコアがもっともよい対応を選択し、格を決定する。対応づけ全体のスコアは、各格の対応の類似度を足したものとする。各格の対応の類似度は「格フレームの選択」で述べた類似度と同様である。

係助詞句	：ガ, ヲ, ガ２[2]
被連体修飾詞	：ガ, ヲ, 外の関係

ゼロ代名詞の検出

格解析が終わったときに、格フレームに入力文の格要素と対応づけられていない格があり、それがガ格、ヲ格、ニ格のいずれかであれば、ゼロ代名詞であると認識する。図 1 の「候補を擁立する」では、格フレーム「擁立 (1)」が選択されており、格フレームのガ格とニ格に対応する入力側の格要素がないことがわかる。従って、システムはガ格とニ格をゼロ代名詞として検出する。

[2] 二重主語構文の外のガ格を「ガ２」格と呼ぶ。

表 2: ゼロ代名詞の格分布 (上位 8 個まで)

格	回数　(割合)	格	回数　(割合)
ガ	4,295 (65.1%)	ガ 2	118　(1.8%)
ニ	1,026 (15.5%)	外の関係	80　(1.2%)
ヲ	460　(7.0%)	ト	71　(1.1%)
デ	163　(2.5%)	カラ	47　(0.7%)

表 3: 先行詞の出現位置分布

位置	回数　(割合)	位置	回数　(割合)
同一文	3,354 (67.3%)	2 文前	301　(6.0%)
1 文前	990 (19.9%)	3 文前以前	341　(6.8%)

検出されたゼロ代名詞に対する先行詞同定処理は 6 章で述べる。

4　位置カテゴリの設定とその順序づけ

格フレームの選択制限によって、先行詞候補は、先行詞として適格なもののみに絞ることができるが、それでも複数残る場合が多い。一般に、ゼロ代名詞の先行詞は、ゼロ代名詞から距離が近いところにある傾向 (近距離特性) があるので、この傾向を先行詞同定に利用することを考える。従来の研究では、ゼロ代名詞と先行詞の間にある文数や単語数など flat な距離でその近さを捉えており、従属節、主節、埋め込み文など文・文章中の構造を用いていない。本研究では、近距離特性を正確にモデル化するために、まず、文・文章中の構造を考慮して、照応詞に対する先行詞候補の位置を分類する。そして、どの位置にある候補が先行詞となりやすいかを学習コーパスから取得し、それを順に並べることによって先行詞の位置選好順序を得る。

本章では、まず、学習コーパスにおける先行詞の捉え方を述べる。次に、ゼロ代名詞と先行詞候補の位置関係の分類 (**位置カテゴリ**と呼ぶ) を導入し、最後にその順序づけ、つまり先行詞の位置選好順序について述べる。

4.1　関係コーパスにおける先行詞の捉え方

本研究では、関係コーパス (Kawahara, Kurohashi, and Hasida 2002) を用いて位置カテゴリの順序の学習や解析実験などを行う。このコーパスは、新聞記事に対して文章中の用言・サ変名詞に対する格関係、名詞間の関係、および共参照をタグづけしたものである。用言に対する格関係には、ゼロ代名詞の先行詞が含まれている。

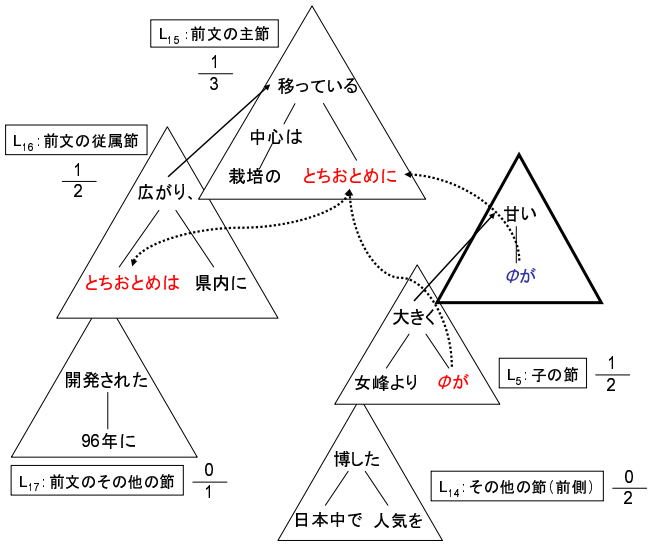

図 2: 先行詞の捉え方

本研究で対象とするのは、関係コーパス中の 379 記事、3,695 文である。用言 (動詞、形容詞、名詞+判定詞) は 11,149 個出現し、そのうちゼロ代名詞を格要素とする用言は 5,530 個あり、ゼロ代名詞は 6,602 個であった。ゼロ代名詞となっている格の分布を表 2 に示す。ゼロ代名詞のうち先行詞が記事中に存在するものは 4,986 個、先行詞が記事中に存在しないものは 1,616 個であった。記事中に存在しない先行詞は不特定の人々である場合が多かった。

先行詞が記事中に存在するゼロ代名詞に対して、コーパスに直接タグづけされているもの以外にも先行詞が存在することがある。それは、タグづけされている先行詞と共参照タグで関係づけられているものや、同じ対象を指示する他のゼロ代名詞であり、それらも先行詞として捉える。例えば、図 2 において、「甘い」のガ格はゼロ代名詞であり、その先行詞は前文主節の「とちおとめ」であるとタグづけされている。この「とちおとめ」と、前文従属節にある「とちおとめ」は共参照関係にあるので、前文従属節の「とちおとめ」も先行詞とみなす。また、対象文の「大きく」のガ格もゼロ代名詞であり、これも「とちおとめ」を指しているため、このゼロ代名詞も先行詞とみなす。

先行詞が出現する文の位置を同一文、1 文前、2 文前、···、n 文前、··· のように分類したときの分布を表 3 に示す。この表において、先行詞が複数存在する場合は、n の値がもっとも小さい先行詞の位置としている。次節で導入する位置カテゴリは、先行詞が対象文から 2 文前までの間にある場合を対象としている。これによって 93.2%のゼロ代名詞がカバーされる。

以下で述べる位置カテゴリの順序づけの学習、5 章で述べる分類器の学習は、関係コーパス中の 279 記事を学習コーパスとして用いる。7 章の実験では、残りの 100 記事をテストコーパスとして用いる。

表 4: 設定した位置カテゴリ

対象文		
L_1 ：「V_z の親用言」の格要素		主節
L_2 ：「V_z の親用言」の格要素		
L_3 ：「V_z の親用言」の格要素	並列	主節
L_4 ：「V_z の親用言」の格要素	並列	
L_5 ：「V_z の子用言」の格要素		
L_6 ：「V_z の子用言」の格要素	並列	
L_7 ：「V_z の親体言の親用言」の格要素		主節
L_8 ：「V_z の親体言の親用言」の格要素		
L_9 ：「V_z の親用言の親用言」の格要素		主節
L_{10}：「V_z の親用言の親用言」の格要素		
L_{11}：「主節用言」の格要素		主節
L_{12}：「主節に係る従属節用言」		
L_{13}：その他の節の格要素 (V_z より後)		
L_{14}：その他の節の格要素 (V_z より前)		
1 文前		
L_{15}：「主節用言」の格要素		主節
L_{16}：「主節に係る従属節用言」の格要素		
L_{17}：その他の節の格要素		
2 文前		
L_{18}：「主節用言」の格要素		主節
L_{19}：「主節に係る従属節用言」の格要素		
L_{20}：その他の節の格要素		

4.2 位置カテゴリの設定

近距離特性を正確にモデル化するために、まず、ゼロ代名詞に対する先行詞候補の位置を分類した位置カテゴリを導入する。これは、従属節、主節、埋め込み文など文・文章中の節がもつ構造に着目して表 4 のように設定した。表 4 において、V_z はゼロ代名詞をもつ用言を示す。ここで、表 4 において「」で囲まれた用言を V_a と呼ぶことにする (V_a の格要素が先行詞候補である)。“並列” とは、V_z と V_a が並列関係にあることを示し、“主節” とは V_a がその文の主節用言であることを示す。位置カテゴリはそれぞれ独立であり、“主節用言” は、主節属性をもつほかの位置カテゴリに含まれない主節の用言を表す。

例えば、図 2 の「甘い」については、ガ格がゼロ代名詞である。「移っている」は前文の主節であるので、その格要素である「中心」「栽培」「とちおとめ」は L_{15}、「大きく」は子用言な

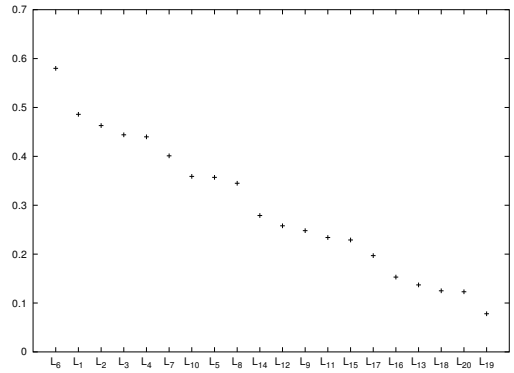

図 3: 位置カテゴリの順序 (ガ格)

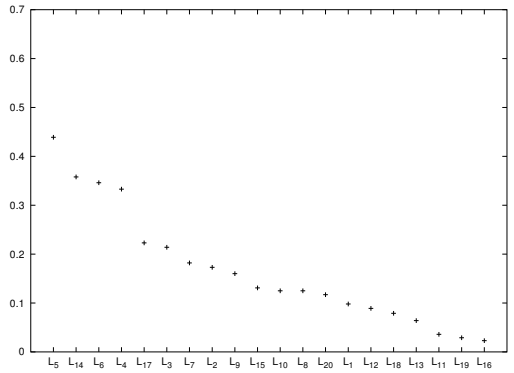

図 4: 位置カテゴリの順序 (ヲ格)

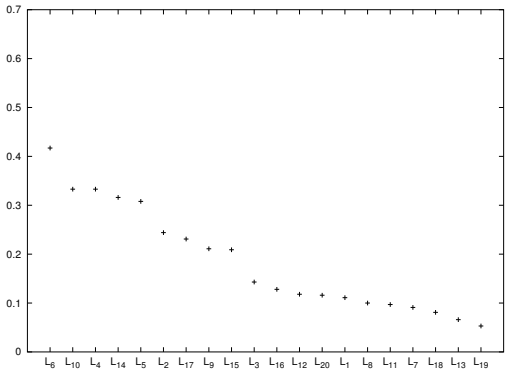

図 5: 位置カテゴリの順序 (ニ格)

ので「とちおとめ」を指すゼロ代名詞と「女峰」は L_5 などとなる。

4.3　位置カテゴリの順序づけ

それぞれの位置カテゴリが、どれくらい先行詞をとりやすいかを関係コーパスから取得し、その順に位置カテゴリを並べることによって、先行詞の位置選好順序を得る。位置カテゴリ L の先行詞のとりやすさを次式のようにスコア化する。

$$\frac{\text{先行詞が } L \text{ にある回数}}{L \text{ にある先行詞候補の数}}$$

例えば、図 2 の「甘い」のゼロ代名詞に対して、前文の主節 L_{15} に含まれる先行詞候補は「中心」「栽培」「とちおとめ」の 3 つであり、「とちおとめ」は先行詞であるので、この位置カテゴリのスコアは 1/3 となる。子用言の節 L_5 については、候補が「とちおとめ」を指示するゼロ代名詞と「女峰」の 2 つなので、スコアは 1/2 となる。

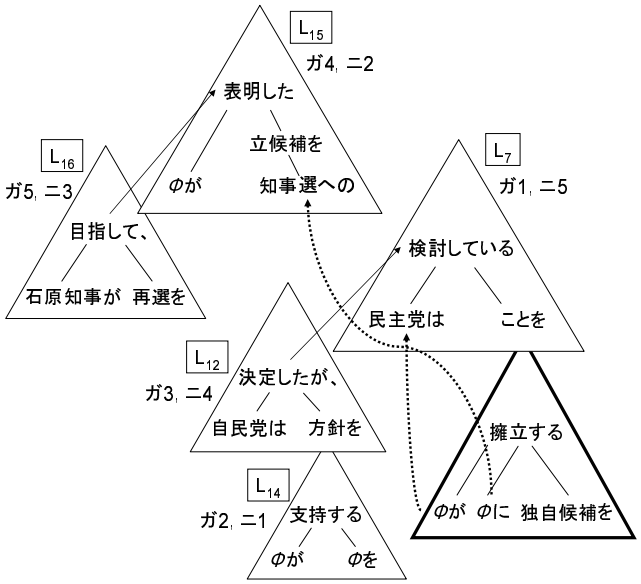

図 6: V_z = 「擁立する」のときの位置カテゴリ

学習コーパス 279 記事を用いて、スコアをゼロ代名詞の格ごとに集計し、スコア順に位置カテゴリを並べた。ガ格、ヲ格、ニ格の順序をそれぞれ、図 3、4、5 に示す。ガ格のゼロ代名詞は、主節の親用言の格要素を先行詞としやすく、ヲ格のゼロ代名詞は、子用言の格要素を先行詞とすることが多いことなどがわかる。

図 1 の記事中の「擁立する」のガ格とニ格のゼロ代名詞に対する位置カテゴリの順序を図 6 に示す。ガ格では L_7 が 1 位になるが、ニ格では L_{14} が 1 位となり、格ごとに位置カテゴリの順序が異なっている。また、L_{12} のように、構造を考慮しない距離尺度でゼロ代名詞に近い位置であっても、位置カテゴリ順の上位にくるわけではないことがわかる。

5　分類器の利用

本研究では、先行詞の決定に関与する様々な要因を考慮するために機械学習による分類器を用いる。この分類器は、先行詞候補が先行詞として適格かどうかを判定する 2 値分類器である。これは、表 5 に示した素性を用いて作成する。素性は、類似度や位置カテゴリなど、先行詞候補とゼロ代名詞の両方に関係するもの、格や意味属性など、先行詞候補とゼロ代名詞のどちらか一方にのみ関係するものからなる。これらは自動的に生成されるのでノイズが含まれる。類似度 *similarity* は、格フレームの選択において格フレームが決定できない場合には計算できない。このような場合には、対象となっている用言のすべての格フレームに含まれる用例群と先行詞候補との類似度をとることにする。

表 5: 分類器に用いる素性

ゼロ代名詞と先行詞候補の両方に関わる素性	
similarity	先行詞候補と格フレームの用例との類似度　(0 - 1)
loc_category	先行詞候補の位置カテゴリ ($L_1, \cdots, L_{20}$)
before_ana	先行詞候補がゼロ代名詞より前にある　(yes, no)
depend_over_ana	先行詞候補がゼロ代名詞を越えて係る　(yes, no)
先行詞候補の素性	
ante_CM	先行詞候補の格　(が, を, に, $\cdots$)
ante_pred_nm	先行詞候補が係る用言が連体修飾節を構成している　(yes, no)
ante_fs	先行詞候補がその記事の先頭文にある　(yes, no)
ante_depend_mc	先行詞候補がその文の主節に係る　(yes, no)
ante_tm	先行詞候補が後に係助詞「は」を伴っている　(yes, no)
ante_agent	先行詞候補が意味属性<主体>に属する　(yes, no)
ante_pred_level	先行詞候補の係る用言の節の強さ　(0, 1, $\cdots$, 6)
ゼロ代名詞の素性	
ana_CM	ゼロ代名詞の格　(が, を, に)
ana_pred_nm	ゼロ代名詞をもつ用言が連体修飾節を構成している　(yes, no)
ana_pred_voice	ゼロ代名詞をもつ用言の態　(能動, 受動, 使役)
ana_pred_type	ゼロ代名詞をもつ用言の種類　(動詞, 形容詞, 名詞+判定詞)
ana_head_verbal	ゼロ代名詞をもつ用言がサ変名詞である　(yes, no)
ana_cf_agent	格フレームの用例が意味属性<主体>に属する　(yes, no)

分類器としては Support Vector Machines (SVM) (Vapnik 1995) を用いる。訓練データは関係コーパスを用いて作成する。正例としては、ゼロ代名詞からもっとも近く、ゼロ化していない先行詞が存在する文と解析対象の文との間にある先行詞とし、負例は、その間にある先行詞以外の候補とする。もっとも近い先行詞が解析対象の文の 2 文前までに存在しないゼロ代名詞については、訓練対象から除くことにする。図 2 の場合は、「甘い」のゼロ代名詞に対して、「大きく」のゼロ代名詞と前文の 2 つの「とちおとめ」を正例とし、それ以外の名詞を負例とする。

6　先行詞同定処理

本研究で作成した省略解析システム (3.2 節) の先行詞同定処理は、位置カテゴリの順序、格フレーム辞書、分類器を統合的に用いる。本手法は、格ごとに設定された位置カテゴリの順序に従って先行詞候補を調べ、格フレームの用例との類似度が閾値を越え、かつ、分類器によって正例と分類される候補を先行詞に決定する。先行詞としての適格さを計るために分類器のみ

を用いる手法も考えられるが、本研究では格フレームとの類似度を重要視し、分類器と類似度の and 条件とした。これは、次節で述べる実験によっても有効性が示されている。

本手法の先行詞探索範囲は、対象のゼロ代名詞がある文、1 文前、2 文前の 3 文とし、そこに含まれる形式名詞、副詞的名詞以外の名詞を先行詞の候補とする。本手法は、文頭に近い用言から文末に向かって順番に解析しており、すでに解析されたゼロ代名詞も候補に含める。先行詞探索範囲内に条件を満たす先行詞候補がみつからない場合において、格フレームに意味マーカ<主体>があれば、先行詞として「不特定:人」を出力する。これは文章中に存在しない不特定の人々を指すという意味である。候補と格フレームの用例との類似度は、格フレームの対象としている格に属する用例群と候補との類似度で、「格フレームの選択」の節で述べたものと同じである。

手順 2.1「格フレームの選択」において、格フレームが選択されなかった場合は、それぞれの格フレームについて先行詞同定までの処理を行う。省略の指示対象も含む入力述語項構造と格フレームとの対応づけ全体のスコアを格フレームのスコアとし、それがもっとも高かった格フレームに決定する。対象となっている用言の省略解析の結果は、その格フレームを用いた場合のものとする。

類似度の閾値は 0.60 に設定した。ヲ格、ニ格に関しては、位置カテゴリのうち、スコアが 0.20 より低いものは使わないことにした。これは、ある程度よりスコアが低い位置カテゴリは、ゼロ代名詞を余分に認識させることにつながり、適合率を悪化させるためである。これらの閾値は、学習コーパスを用いて交差検定を行うことにより決定した。

図 1 の 2 文目の「擁立」では、ガ格とニ格がゼロ代名詞として認識されている。例えば、ガ格の候補は、位置カテゴリ順に、L_7:民主党, L_{14}:自民党 (ϕ ガ), L_{14}:石原知事 (ϕ ヲ), $\cdots$ となっている。1 番目の「民主党 (類似度: 0.73)」は分類器によって正例と分類され、格フレームの用例との類似度が 0.73 と閾値を越えているので先行詞に決定される。

7　実験

関係コーパスを用いて、省略解析の実験を行った。コーパス中の各記事の長さをある程度そろえるために、各記事の先頭 10 文までを用いることにした。学習コーパス 279 記事を分類器の学習、テストコーパス 100 記事をテストに用いた。分類器としては、SVM パッケージ *TinySVM* (Kudo 2002) を用い、2 次の多項式カーネルで学習を行った。テストとしては、正解の構文構造をもつ 100 記事をシステムに入力し、実験・評価を行った。

その他の手法を比較するために、図 7 のように 7 つの設定について実験を行った。パラメータは、「先行詞の探索戦略」、「距離の尺度」、「スコア付け」の 3 つである。「先行詞の探索戦略」とは、どのように先行詞を探索し決定するかで、“best” と “closest” の 2 種類である。“best” とは、先行詞候補のなかでもっともスコアのよい候補を選択する手法で、“closest” とは、近い候補から順に調べていき、スコアがある閾値を越える候補を先行詞に決定する手法である。「距

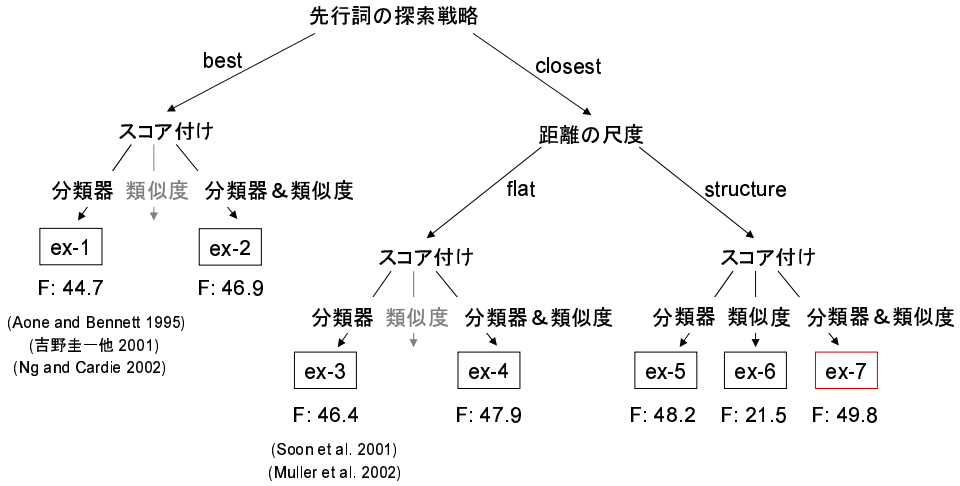

図 7: 実験設定

表 6: 省略解析結果

	適合率	再現率	F 値
ex-1	443/908 (48.8%)	443/1072 (41.3%)	44.7%
ex-2	449/841 (53.4%)	449/1072 (41.9%)	46.9%
ex-3	472/964 (49.0%)	472/1072 (44.0%)	46.4%
ex-4	480/933 (51.5%)	480/1072 (44.8%)	47.9%
ex-5	494/976 (50.6%)	494/1072 (46.1%)	48.2%
ex-6	239/1149 (20.8%)	239/1072 (22.3%)	21.5%
ex-7	496/921 (53.9%)	496/1072 (46.3%)	49.8%

離の尺度」とは、照応詞と先行詞候補間の距離の近さを計る尺度であり、“structure” と “flat” の 2 種類である。“structure” とは、文・文章を構造的にみる手法であり、“flat” とは、間にある文節数を距離とする手法である。「スコア付け」は、先行詞候補を比較するスコアとして何を用いるかで、「分類器」、「類似度」、「分類器&類似度」の 3 種類である。「分類器」は分類器の出力する信頼度をそのまま用いる手法で、「類似度」は先行詞候補と格フレームの用例との類似度を用いる手法であり、「分類器&類似度」は本手法のように分類器と類似度の両方を用いる手法である。図の ex-7 が本手法であり、ex-1 が (Aone and Bennett 1995; 吉野圭一他 2001; Ng and Cardie 2002) による手法に類似しており、ex-3 が (Soon et al. 2001; Müller et al. 2002) の手法に類似している。

実験結果を表 6 に示す。表の適合率、再現率はゼロ代名詞検出、先行詞の同定処理の双方を併せて評価したものである。先行詞同定の評価は、システムの出力する先行詞が正解コーパスの先行詞 (群) に含まれていれば正解とする。システムの出力が「不特定:人」の場合も評価に含め、正解コーパスも「不特定:人」であれば正解とする。表 6 によると、ex-7 の精度が他の手法を上回っており、本手法の有効性が示されている。また、ex-7 について、ゼロ代名詞の検出と先行詞の同定の評価を別々に行った。ゼロ代名詞検出については、適合率 87.1%、再現率 74.8%、F 値 80.5%であり、先行詞同定については 61.8%の精度であった。

本手法の精度を、実験環境が類似している関らの手法 (関和広他 2002) と比較する。関らは、新聞の社説記事 30 件と報道記事 30 件に対して、ゼロ代名詞検出と先行詞の同定処理を行っている。我々が用いたコーパスには社説記事は含まれていなかったので、報道記事についての精度をここに挙げると、ゼロ代名詞検出は適合率 48.9%、再現率 88.2%、F 値 62.9%、先行詞同定については 54.0%の精度となっている。コーパスのサイズの違いなどがあるため一概に比較するのは難しいが、本手法の精度は関らの手法より、ゼロ代名詞検出の F 値が 17.6%、先行詞同定の精度が 7.8%よい。特に、ゼロ代名詞検出の精度は大幅に向上しており、関らの用いた人手構築の格フレーム辞書より、自動構築した格フレーム辞書が有効であったと思われる。

主な誤り原因を以下に示す。

ゼロ代名詞の検出誤り

ゼロ代名詞を余分に検出する傾向がある。

(1)　一方、ドゥダエフ政権側の首都防衛司令官は同日夕、テレビを通じ、首都防衛はうまくいっており、ロシア軍の戦車五十両を破壊したと 発表 。

この例では、「発表」の格フレームに対応づけられていないニ格があり、ニ格をゼロ代名詞と認識してしまう。これは格フレームが悪いのではなく、文脈によってその格をとる場合ととらない場合があるためである。これに対処するためには、文脈に関する素性を機械学習に入れる必要がある。

省略解析の対象を限定していることによる誤り

現在の省略解析システムは用言のみしか対象としていない。

(2)　村山富市 首相$_{\times}$ は年頭にあたり首相官邸で内閣記者会と二十八日会見し、社会党$_{\bigcirc}$ の新民主連合所属議員の離党問題について「政権に影響を及ぼすことにはならない。離党者がいても、その範囲にとどまると思う」と述べ、大量離党には 至らない との見通しを示した。

この例の「至らない」のガ格は「社会党」(○印) であるが「首相」(×印) と誤って解析される。これは、「社会党」は位置カテゴリ順では上位に来ないためである。この誤りは、省略解析の対象をサ変名詞など用言以外にも広げることによって解決できると思われる。つまり、「離党 (問題)」「(大量) 離党」のガ格に「社会党」が補われ、「(大量) 離党」は「至らない」に対して位置カテゴリの上位なので、「至らない」のガ格に「社会党」が補われる。このように用言だけでなく文章中のより多くの関係を用いれば、省略解析の精度向上が期待できる。

8　おわりに

本稿では、格フレーム辞書、先行詞の位置選好順序、機械学習の 3 つを統合した省略解析手法を提案した。本手法は、ゼロ代名詞検出とその先行詞同定の 2 つの処理からなる。ゼロ代名詞検出は、格フレーム辞書に基づく格解析によって行う。先行詞同定は、先行詞の位置選好順に候補を調べ、機械学習による分類器が正例と分類し、かつ、格フレームによる選択制限を満たす候補を先行詞として決定する。大規模解析実験を行った結果、ゼロ代名詞検出が適合率 87.1%、再現率 74.8%、ゼロ代名詞の先行詞同定が 61.8%の精度であった。特に、ゼロ代名詞検出の精度は、先行研究と比べてかなり向上しており、自動構築した格フレーム辞書の有効性が示された。今後は、解析対象にサ変名詞を入れるなど文章中の関係をさらに捉えることを目指す予定である。

参考文献

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

Ge, N., Hale, J., and Charniak, E. (1998). "A Statistical Approach to Anaphora Resolution." In *Proceedings of the 6th Workshop on Very Large Corpora*, pp. 161–170.

Hobbs, J. (1978). "Resolving Pronoun References." *Lingua*, **44**, 339–352.

Kawahara, D. and Kurohashi, S. (2002). "Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis." In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431.

Kawahara, D., Kurohashi, S., and Hasida, K. (2002). "Construction of a Japanese Relevance-tagged Corpus." In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013.

Kudo, T. (2002). *TinySVM: Support Vector Machines.* http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/.

Kurohashi, S. and Nagao, M. (1994). "A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary." In *IEICE Transactions on Information and Systems*, Vol. E77-D No.2.

Müller, C., Rapp, S., and Strube, M. (2002). "Applying Co-Training to Reference Resolution." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 352–359.

Ng, V. and Cardie, C. (2002). "Improving Machine Learning Approaches to Coreference Resolution." In *Proceedings of the 40th Annual Meeting of the Association for Computational*

Linguistics, pp. 104–111.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases." *Computational Linguistics*, **27** (4), 521–544.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

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

関和広, 藤井敦, 石川徹也 (2002). "確率モデルを用いた日本語ゼロ代名詞の照応解析." 自然言語処理, **9** (3), 63–85.

吉野圭一, 竹内和広, 松本裕治 (2001). "機械学習を用いた日本語ゼロ代名詞照応関係の同定." 言語処理学会 第 7 回年次大会発表論文集, pp. 506–509.

村田真樹 長尾眞 (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.

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), 49–65.

略歴

河原 大輔: 1997 年京都大学工学部電気工学第二学科卒業。1999 年同大学院修士課程修了。2002 年同大学院博士課程単位取得認定退学。現在、東京大学大学院情報理工学系研究科 学術研究支援員。構文解析、省略解析の研究に従事。

黒橋 禎夫: 1989 年京都大学工学部電気工学第二学科卒業。1994 年同大学院博士課程修了。京都大学工学部助手、京都大学情報学研究科講師を経て、2001 年東京大学大学院情報理工学系研究科助教授、現在に至る。自然言語処理、知識情報処理の研究に従事。

(2002 年 1 月 1 日 受付)
(2002 年 1 月 1 日 再受付)
(2002 年 1 月 1 日 採録)