

書式構造的特徴を用いたゼロ代名詞照応解析¹

中岩 浩巳* 上門 孝宏† 野沢 弥生‡ Francis BOND*

*NTT コミュニケーション科学研究所

†NTT ソフトウェア株式会社

‡NTT アドバンステクノロジー株式会社

1 はじめに

自然言語では通常、相手（読み手もしくは聞き手）に容易に判断できる要素は、文章上表現しない場合が多い。この現象は、機械翻訳システムや対話処理システム等の自然言語処理システムにおいて、大きな問題となる。例えば、機械翻訳システムにおいては、原言語では陽に示されていない要素が、目的言語で必須要素になる場合、陽に示されていない要素の同定が必要となる。特に日英機械翻訳システムにおいては、日本語の格要素が省略される傾向が強いのにに対し、英語ではそれらの格要素が訳出上必須要素となることが多いため、省略された格要素（ゼロ代名詞と呼ばれる）の照応解析技術は重要となる。

近年、電子化された文章の重要性が認識されるのに伴い、電子化の対象となる媒体（本、報告書等）に含まれるすべての情報（文だけではなく、章や節のような文書の構造を示す情報や、表、レイアウト等の情報も含む）を電子化するための記述言語（マークアップ言語）が提案されてきている。また、これら HTML や SGML のようなマークアップ言語の情報を含んだ文書をコンピュータによって、作成支援する環境が充実してきた。したがって、文だけではなく、章や節などの文書構造に関するメタ情報を含んだ文書が流通するようになってきた (Bond, Nakaiwa, and Ikehara 1995)。そこで、これら文書構造情報を含む文中に現れるゼロ代名詞の照応解析技術への期待が従来以上に高まっている。

従来からゼロ代名詞の照応解析に関しては、様々な手法が提案されている (Kameyama 1986; Walker, Iida, and Cote 1990; Yoshimoto 1988; 堂坂 1994; 林 千葉 1990)。また、翻訳対象分野を限定しない実用的機械翻訳システムへの応用を特に考慮した手法も提案されている (中岩 池原 1993; 中岩, 白井, 池原 1995; Nakaiwa and Ikehara 1995)。しかし、これらの手法が解析対象としているゼロ代名詞は、文書構造情報の活用を考慮していないため、文書構造情報付きの文書中のゼロ代名詞を解析するのに適しているとはいえない。

本稿では、従来の照応解析手法でも用いられる用言意味属性、様相表現、及び接続表現による意味的、語用論的制約に加えて、その文書の書式構造的特徴も用いた日本語ゼロ代名詞の照応解析手法について提案する。

2 文章構造に依存するゼロ代名詞照応現象

書式構造的特徴を持つ文書におけるゼロ代名詞の傾向をつかむために、仕様書（国際調達用）1856文におけ

るゼロ代名詞照応現象の調査を行った。本仕様書文を対象としたのは、以下の理由による。

- 章立てされていること（文書構造情報がある）
- 書き手の意図が明確に記述されていること

本仕様書文に対して、仕様書文が持つ構造的な特徴と、照応解析が必要となるゼロ代名詞とその照応要素の出現傾向を調査した。

2.1 仕様書文の構造的な特徴

ここでは、本調査の結果得られた仕様書文における構造的な特徴について特徴別に述べる。

(1) 構造的役割と意味的役割

まず、仕様書やマニュアルなどの構造的な特徴として、意味的役割があげられる。例えば、目次や索引などである。さらに、各文の構造的役割があげられる。例えば、章立ての番号が付与されている文は「標題」の役割として機能し、同じ章の内容を簡潔に述べている。また、記号「ア」等が付与されている文は「箇条書き」として機能し、個別の項目について述べている。これらの構造的役割は、照応関係を認定する際の条件として使用することができる。仕様書（国際調達用）における意味的役割と構造的役割を表1に示す。

表 1: 意味的役割と構造的役割

意味的役割	構造的役割	役割を付与する対象
タイトル	タイトル	文章の題名
目次	目次	目次
	目次の項目	目次を構成する要素
本論	標題	章立ての番号が付与されているもの ex. 5 説明事項 5.1 AV制作設備 (1), ア等の項番号が付与されているもの ex. (1)収録スタジオ
	箇条書き	(1), ア等で始まる文
	箇条の補足	箇条書きの文に続く文で、「また」「なお」「さらに」等接続詞で始まる文
	本文	文字から始まる文（箇条書き、箇条の補足を除く）
表	表 No.	表の番号
	表のタイトル	表の題名
	表の項目	各カラムのタイトルについての内容
	表の本文	表の項目についての概要
	表の補足	表の項目を脚注的に補足説明したもの

(2) 階層構造

次に、文章中に存在する特徴として、階層構造があげられる。ある情報が文章中のどの範囲において有効であ

¹Resolving Zero Pronouns using Textual Structure

るかが認定できれば、照応要素の有効範囲を決定する際の重要な情報となる。仕様書（国際調達用）を分析した結果、章立ての番号を相対的に比較して階層構造（レベル）を認定すれば、情報の有効範囲を特定できることが分かった。

ex: 章立て番号の関係

6 ⊃ 6.1 ⊃ 6.1.1 ⊃ (1) ⊃ 1 ⊃ A
→ 相対的にレベルを見る

2.2 照応現象の特徴

仕様書（国際調達用）での照応現象は照応要素の出現位置によって3つに分類される。(1)同一文内に存在する場合（以下、文内照応と呼ぶ）、(2)同一文内には存在しないが、文章中のどこかに存在する場合（文間照応）、(3)文章中に存在しない場合（文外照応）である。照応要素の出現傾向とゼロ代名詞の件数を表2に示す。

(1) 文内照応

同一文内に照応要素が存在するタイプのものは、41件あった。このうち、多く見られた現象は標題が異なる語に言い換られて照応要素とされている場合である。照応要素が「は格」、すなわち助詞「は」を用いるもの等があげられる。

(2) 文間照応

文章中に照応要素が存在する場合について、「が格」がゼロ代名詞化され、照応要素が標題である場合が301件と、全体を通して最も多かった。これは、標題の情報が有効である範囲を1つのレベルとしてとらえた場合に、標題で一度明示されたものは、その後、その同じレベルに登場した文ではゼロ代名詞化されている、という傾向があるため、顕著に見られた現象だと思われる。これらは下位のレベルで見られることから、レベルがある程度深くなると、標題と、標題と同じレベルに登場する文（役割は「箇条書き」）との関係が強くなると言える。また、照応要素が同じレベルの標題である場合が293件に対して、1つ上のレベルの標題が7件、2つ上のレベルの標題が1件であることから、ゼロ代名詞が存在する文に近い標題が照応要素として優先されると言える。照応要素が「箇条書き」である場合、ゼロ代名詞が存在する文が「また」「なお」「さらに」の接続詞で始まり、直前の文（役割は「箇条書き」）を補足説明しているものが多く見られた。表は仕様書（国際調達用）中に5箇所あった。全て二項関係のものであり、「表の本文中」にゼロ代名詞が存在し、照応要素が「表の項目」である場合であった。その他の現象は存在しなかった。表は文章中のどこにでも現れ得るが、他の文章とは比較的関わりのない表の枠組みだけで閉じられた特殊な構造であると言える。

(3) 文外照応

文章中に照応要素が現れないタイプの場合、仕様書（国際調達用）では話し手（依頼者）・聞き手（作者）と対象が決まっており、誰が誰にどんな情報を伝えたいかが、比較的明確であるためゼロ代名詞化される傾向にあった。照応要素が聞き手（You）のものは「～て下さい」（二人称依頼・命令）、「～こと」（仕様義務）等の表現を伴う場合が多く、また、照応要素が話し手（I）の場合は行動を表す用語に「～ます」（断定・丁寧）の

表現を伴うものが多いことから、用言の意味的屬性の他に様相表現の情報を用いることによって、照応関係を決定することが可能である。

このように、照応要素の出現傾向から推察すると、文章中に照応要素が存在するタイプのものは、文章構造の情報が有効に活用できると考えられる。文の構造だけでなく、文の持つ役割（意味）や文のレベル（入れ子構造）の情報を用いることによって、より正確に照応関係を認定することが可能である。

3 照応解析手法

前節の調査で得られた結果を元に、書式構造的特徴を持つ文書におけるゼロ代名詞の照応解析手法について述べる。本手法は、以下の2つに分かれている。

- 書式構造的特徴をつかむための文章構造認定
- 書式構造的特徴を用いたゼロ代名詞照応解析

3.1 文章構造認定手法

まず、各文に対する構造として、以下の要素から成るリスト構造を文構造とする。

- 文管理番号
- 段落表示
- 階層レベル
- 構造的役割
- 意味的役割
- 特別な意味的役割

ここで、文管理番号と段落表示は、解析対象となる入力日本文に付与されている情報である。この段落表示とは、マークアップ言語における章や節などのメタ情報に対応する情報であり、階層レベル、構造的役割、意味的役割、特別な意味的役割と対応関係にある。また、階層レベルとは、後述の意味的構造におけるリスト構造の深さを表すもので、意味的構造のトップレベルを1として、リストが深くなるほど大きくなる。なお、構造的役割には、新しく階層構造を作るか否かの情報も含まれる。特別な意味的役割とは、後述の意味的構造において、以下の理由により、特別にまとめる必要がある構造に対して付与する意味的役割のことである。これにより、まとめられた文構造の集合を特別な意味的構造とする。

- 文章中に表などがある場合、表などの直後にある文については、表などに含まれている最後の文ではなく、表などの直前の文との比較により、文章構造を認定するため。

つぎに、同じ意味的役割を持つ文構造の集合を意味的構造とする。この意味的構造は、階層レベルに応じた深さを持つ文構造により構成されるリスト構造である。最後に、意味的構造を要素とするリスト構造を文章構造とする。実際の文章構造の認定はつぎのようになる。まず、文章構造から当該文を認定するために比較しようとする文（以下、比較文と呼ぶ）を特定する。比較文は通常当該文の直前に認定された文であるが、以下の場合には特別な意味的構造の直前に認定された文となる。

表 2: 照応要素の分類によるゼロ代名詞の件数

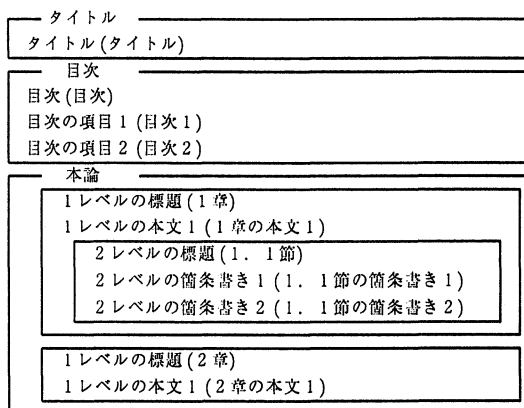
ゼロ代名詞	照応要素													小計 (件)
	文内照応				文間照応						文外照応			
	は	が	を	に	標題			箇条書き			表の項目	話し手	聞き手	
					同じレベル	1つ上のレベル	2つ上のレベル	1文前	2文前	3文前				
が	12	6	0	0	293	7	1	17	2	2	34	39	34	447
が(埋込)	0	0	0	0	0	0	0	0	0	0	0	4	6	10
を	13	2	7	1	0	0	0	0	0	0	0	0	0	23
小計	41				356						83			480

(ゼロ代名詞の出現数は480箇所423文)

- 当該文の直前に認定された文が特別な意味的役割を持ち、当該文が特別な意味的役割を持っていない場合

以下のような文章構造であるとして、

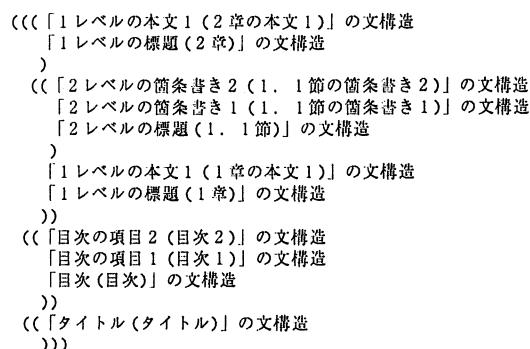
つぎに当該文と比較文とを比較して、当該文と比較文とが異なる意味的構造に属する場合には、当該文を新しい意味的構造の始まりとして、当該文の文構造のみを要素とする意味的構造を文章構造に追加する。当該文と比較文とが同じ意味的構造に属する場合には、当該文の構造的作用と比較文の構造的作用の関係により、以下のように認定する。



- 当該文が新しく階層構造を作る構造的作用を持つ場合は、当該文の階層レベルと比較文の階層レベルの関係により、意味的構造において、当該文の階層レベルに対応した位置に、当該文の文構造を追加する。
- 当該文が新しく階層構造を作らない構造的作用を持つ場合、比較文と同じ階層レベルの構造として、当該文の文構造を意味的構造に追加する。

以下のようなリスト構造が得られる。

本処理で出力される文章構造は、GroszとSidnerが提案したDiscourse Structure (Grosz and Sidner 1986)に類似したものであるため、文章構造としては汎用的で信頼性の高いものであるといえる。また、本節で説明した文章構造認定手法では、完全にマークアップ言語におけるメタ情報に相当する情報(段落表示)のみを用いて、文章構造を認定しているため、本手法を実行する際の処理的負荷は軽くなる。



本手法を用いた場合、以下のような文章に対しては、

```

    タイトル
    目次
    目次1
    目次2
    1章
    1章の本文1
    1.1節
    1.1節の箇条書き1
    1.1節の箇条書き2
    2章
    2章の本文1
  
```

3.2 ゼロ代名詞照応解析手法

本手法では、(1)文内照応、(2)文間照応、(3)文外照応を対象としているが、このうち、書式構造的特徴を特に用いているのは、(2)文間照応である。また、本手法で用いるゼロ代名詞照応解析ルールは、ゼロ代名詞が含まれている文(以下、省略文と呼ぶ)からゼロ代名詞を特定するための条件として、以下の2つの条件を記述する部分と、

- 省略文を特定するための書式構造的特徴に対する条件
- 省略文からゼロ代名詞を特定するための構文構造的特徴に対する条件

照応要素が含まれている文(以下、照応文と呼ぶ)から照応要素を特定する条件として、以下の2つの条件を記述する部分と、

- 照応文を特定するための書式構造的特徴に対する条件
- 照応文から照応要素を特定するための構文構造的特徴に対する条件

ルールマッチングを制御するための情報を記述する部分とから成り立っている。

(1) 文内照応

文内照応の場合、省略文からゼロ代名詞を特定するための構文構造的特徴に対する条件は、同時に照応文を特定するための書式構造的特徴に対する条件でもある。この条件とのマッチングを行うことにより、ゼロ代名詞を特定すると同時に照応要素も特定する。

(2) 文間照応

文間照応の場合、文内照応、文外照応の場合と異なり、書式構造的特徴を用いて、ルールマッチングの有効範囲を絞り込んでいる。

(a) ゼロ代名詞の特定

照応させようとする文と、省略文を特定するための書式構造的特徴に対する条件とをマッチングさせることによって省略文を特定し、ルールマッチングの有効範囲を絞り込んでいる。例えば、以下のようになる。

- 照応させようとする文の構造的役割が「本文」で、省略文を特定するための書式構造的特徴に対する条件も「本文」であれば省略文であると判断するが、省略文を特定するための書式構造的特徴に対する条件が「箇条書き」であれば省略文ではないと判断する。

省略文が特定できれば、省略文からゼロ代名詞を特定するための構文構造的特徴に対する条件とのマッチングを行うことにより、ゼロ代名詞を特定する。省略文が特定できなければ、照応を中止する。

(b) 照応要素の特定

照応要素を特定するためには、まず照応文を特定しなければならないが、一文ごとにさかのぼって照応文を特定することは得策とはいえない。そこで、照応文を特定するための書式構造的特徴に対する条件を用いて、照応文を特定する。

まず、意味的役割により、文章構造における意味的構造の絞り込みを行う。なお、意味的役割が条件として指定されていない場合には、省略文と同じ意味的構造を対象とする。つぎに、階層レベル、特別な意味的役割などにより、意味的構造における階層構造の絞り込みを行う。なお、これらが条件として指定されていない場合には、意味的構造全体を対象とする。最後に、構造的役割により、照応文の絞り込みを行う。

照応文が特定できれば、照応文から照応要素を特定するための構文構造的特徴に対する条件とのマッチングを行うことにより、照応要素を特定する。照応文が特定できなければ照応を中止するが、条件によっては複数の文構造を対象として、OR条件での照応要素の特定も可能である。

(3) 文外照応

ゼロ代名詞照応解析手法は文内照応の場合と同様であるが、照応要素は文章外の要素である。

4 おわりに

本稿では、従来の照応解析手法でも用いられる意味的、語用論的制約に加えて、この文書の書式構造的特徴も用いた日本語ゼロ代名詞の照応解析手法について提案した。書式構造情報をもとに階層的な文章構造を構築し、意味的、語用論的制約を用いて文章構造に応じた照応要素を決定することにより、汎用的な照応解析が可能となった。今後は、本手法の性能評価実験を行うとともに、実際のマークアップ言語を付与した文章における文章構造認定処理手法の検討を行いたい。また、照応解析ルール間の優先順位を効率的かつ効果的に制御する枠組についての検討を行いたい。

参考文献

- Bond, F., Nakaiwa, H., and Ikehara, S. (1995). "Tagging an aligned Japanese/English Corpus." In *Proc of NLP '96*, pp. 325-328.
- Grosz, B. J. and Sidner, C. L. (1986). "Attention, Intentions, and the Structure of Discourse." *Computational Linguistics*, 12 (3), 175-204.
- Kameyama, M. (1986). "A Property-sharing Constraint in Centering." In *Proceedings of the 24th Annual meeting of the ACL*, pp. 200-206.
- Nakaiwa, H. and Ikehara, S. (1995). "Intrasentential Resolution of Japanese Zero Pronouns in a Machine Translation System using Semantic and Pragmatic Constraints." In *Proc of TMI '95*, pp. 96-105.
- Walker, M., Iida, M., and Cote, S. (1990). "Centering in Japanese Discourse." In *Proc of COLING '90*.
- Yoshimoto, K. (1988). "Identifying Zero Pronouns in Japanese Dialogue." In *Proc of COLING '88*, pp. 779-784.
- 堂坂浩二 (1994). "語用論的条件の解釈に基づく日本語ゼロ代名詞の指示対象同定." *情報処理学会論文誌*, 35 (5), 768-778.
- 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, 白井諭, 池原悟 (1995). "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." *電子情報通信学会, 言語理解とコミュニケーション研究会, NLC-95* (5), 33-40.
- 林良彦 千葉裕子 (1990). "日本語受動化可否判定アルゴリズムの検討." *情報処理学会論文誌*, 31 (10), 1438-1443.