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

村田 真樹[†] 長尾 真[†]

日本語文章における代名詞などの代用表現の指す対象が何であるかを把握することは、対話システムや高品質の機械翻訳システムを実現するために必要である。そこで、本研究では用例、表層表現、主題・焦点などの情報を用いて指示詞・代名詞・ゼロ代名詞などの指示対象を推定する。従来の研究では、代名詞などの指示対象の推定の際には意味的制約として意味素性が用いられてきたが、本研究では対照実験を通じて用例を意味素性と同様に用いることができることを示す。また、連体詞形態指示詞の推定に意味的制約として「A の B」の用例を用いるなどの新しい手法を提案する。指示対象を推定する枠組は、以下のとおりである。指示対象の推定に必要な情報をすべて規則にする。この規則により指示対象の候補をあげながら、その候補に得点を与える。得点の合計点が最も高い候補を指示対象とする。この枠組では規則を柔軟に書くことができるという利点がある。この枠組で実際に実験を行なった結果、指示詞・代名詞・ゼロ代名詞の指示対象を学習サンプルにおいて 87% の正解率で、テストサンプルにおいて 78% の正解率で、推定することができた。

キーワード: 機械翻訳, 指示対象, 用例, 表層表現, 対照実験

An Estimate of Referents of Pronouns in Japanese Sentences using Examples and Surface Expressions

MASAKI MURATA [†] and MAKOTO NAGAO[†]

It is necessary to clarify referents of pronouns in machine translation and conversational processing. We present a method of estimating referents of demonstrative pronouns, personal pronouns, and zero pronouns in Japanese sentences using examples, surface expressions, topics and focuses. In conventional work, semantic markers have been used for semantic constraints. On the other hand, we use examples for semantic constraints and show that examples are as useful as semantic markers through control experiments. We also propose many new methods for estimating referents of pronouns. For example, we use examples of the form “A of B” for estimating referents of demonstrative adjectives. The framework of estimating referents is as follows. We make the rules from the informations that are necessary for estimating referents of pronouns. By these rules, we list possible referents of a pronoun and give them points. We estimate that the possible referent with the highest score is the referent. This framework has the advantage of writing rules flexibly. When we experimented in this framework, we obtained a precision rate of 87% in the estimation of referent of demonstrative pronouns, personal pronouns, and zero pronouns on training sentences, and obtained a precision rate of 78% on held-out test sentences.

KeyWords: *Machine translation, Referent, Example, Surface expression, Control experiment*

1 はじめに

日本語文章における代名詞などの代用表現を含む名詞の指示対象が何であるかを把握することは、対話システムや高品質の機械翻訳システムを実現するために必要である。そこで、我々は用例、表層表現、主題・焦点などの情報を用いて名詞の指示対象を推定する研究を行なった。普通の名詞の指示対象の推定方法はすでに文献(村田・長尾 1996)で述べた。本稿では指示詞・代名詞・ゼロ代名詞の指示対象の推定方法について説明する。

代名詞などの指示対象を推定する研究として過去にさまざまなものがあるが(長尾, 辻井, 田中 1976)(Kameyama 1986) (山村毅, 大西昇, 杉江昇 1992)(Takada and Doi 1994)(中岩浩巳・池原悟 1995), これらの研究に対して本研究の新しさは主に次のようなものである。

- 従来の研究では代名詞などの指示対象の推定の際に意味的制約として意味素性が用いられてきたが、本研究では対照実験を通じて用例を意味素性と同様に用いることができることを示す。一般に意味素性付きの格フレームの方が用例付きの格フレームよりも作成コストがかかるので、用例を意味素性と同様に用いることができることがわかるだけでも有益である。
- 連体詞形態指示詞の推定には意味的制約として「A の B」の用例を用いる。
- 「この」が代行指示になりにくいという性質を利用して解析を行なう。
- 指示詞による後方照応を扱っている。
- 物語文中の会話文章の話し手と聞き手を推定することで、その会話文章中の代名詞の指示対象を推定する。

論文の構成は以下の通りである。2 節では、本研究の指示対象を推定する枠組について説明する。次に、その枠組で用いる規則について、3 節, 4 節, 5 節で指示詞, 代名詞, ゼロ代名詞の順に説明する。6 節では、これらの規則を実際に用いて行なった実験とその考察を述べる。7 節で本研究の結論を述べる。

2 指示対象を推定する枠組

本研究での指示詞・代名詞・ゼロ代名詞を含む名詞の指示対象の推定は、手がかりとなる複数の情報をそれぞれ規則にし、これらの規則を用いて指示対象の候補をあげながらその候補に得点を加えていき、合計点が最も高い候補を指示対象とすることによって実現した。これは、照応解析のように複雑な問題では複数の情報が絡み合っており、複数の情報を総合的に判断することにより解析を行なうためである。規則に応じて候補に得点を足していく操作は、その候補が指示対象であるという確信度が高まっていくことに対応している。

まず、解析する文章を (Kurohashi and Nagao 1994) の方法によって構文解析・格解析する。

† 京都大学工学部 電子通信工学教室, Department of Electronics and Communication, Kyoto University

条件部 $\Rightarrow$ { 提案 提案 .. } 提案 $:=$ (指示対象の候補 得点)
--

図 1 列挙判定規則の表現

条件部 $\Rightarrow$ (得点)

図 2 判定規則の表現

その結果に対して文頭から順に文節ごとにすべての規則を適用して指示対象を推定する。規則には、指示対象の候補をあげながら候補の良さを判定する列挙判定規則とその列挙された複数の候補すべてに対して適用する判定規則の二種類がある。列挙判定規則を図 1 に、判定規則を図 2 に示す。

図中の「条件部」には、(i) 文章中のあらゆる語とその分類語彙表 (国立国語研究所 1964) の分類番号と、(ii) IPAL の格フレーム (情報処理振興事業協会技術センター 1987) の情報と、(iii) 名詞の指示性の情報と、(iv) 構文解析・格解析の結果の情報などを条件として書く。「指示対象の候補」には指示対象の候補とする単語の位置を書く。「得点」は指示対象としての適切さの度合を表す。

指示対象の推定は条件を満足した規則により与えられる得点の合計点で行なう。まずすべての列挙判定規則を適用し得点のついた指示対象の候補を列挙する。このとき同じ候補を列挙する規則が複数あれば得点を加算する。次に列挙された指示対象の各候補に対してすべての判定規則を適用して、各候補ごとに得点を合計する。最も合計点の高い指示対象の候補を指示対象と判定する。最も合計点の高い指示対象の候補が複数個ある場合は、一番初めに出された¹ 指示対象の候補を指示対象とする。

代名詞などを解析するために列挙判定規則および判定規則をそれぞれ指示詞については 51 個と 11 個、代名詞については 4 個と 6 個、ゼロ代名詞については 24 個と 4 個作成した。以降、これらのうち主要なものを指示詞、代名詞、ゼロ代名詞の順に説明する。

3 指示詞の指示対象を推定するための規則

指示詞の解析のための規則は、(正保 1981)(林 1983)(高橋 1990)(金水・田窪 1992) などの文献を参考にしたり、実際の文章を調査することによって作成した。指示詞には名詞形態指示詞、連体詞形態指示詞、副詞形態指示詞の三つがある。以下にそれぞれの指示詞の解析をするための規則の説明を行なう。

¹ 規則の適用順序に従う。ただし、3 節以降で説明する規則の順と適用順序は異なる。

表 1 主題の重み

表層表現	例	重み
ガ格の指示詞・代名詞・ゼロ代名詞	(太郎が) した.	21
名詞 は/には	太郎はした.	20

表 2 焦点の重み

表層表現 (「は」がつかないもので)	例	重み
ガ格以外の指示詞・代名詞・ゼロ代名詞	(太郎に) した.	16
名詞 が/も/だ/なら/こそ	太郎がした.	15
名詞 を/に/, /.	太郎にした.	14
名詞 へ/で/から/より	学校へ行く.	13

3.1 名詞形態指示詞の指示対象を推定するための規則

名詞を指示対象の候補とする規則

列举判定規則 1

名詞形態指示詞か「その/この/あの」の場合

{(同一文中か前文の重みが w で主題と焦点を合わせて数えて n 個前の主題 $w - n - 2$)

(同一文中か前文の重みが w で主題と焦点を合わせて数えて n 個前の焦点 $w - n + 4$)}

この括弧式は、図 1 の規則の提案のリストを表わす。

主題と焦点の定義と重みは表 1, 表 2 のとおりである。

指示詞の場合は、ゼロ代名詞の場合と異なり、既知情報の主題と照応するよりも未知情報の焦点と照応しやすいと考え、係数-2, +4 をつけることによって、主題よりも焦点の方が指示対象になりやすくしている。

主題・焦点の重みや指示詞と指示対象の候補の間の距離に応じて、それぞれの候補の得点、すなわち、その候補の指示対象としての確からしさが変わる。

事態 (用言) を指示対象の候補とする規則

列举判定規則 2

「それ/あれ/これ」や連体詞形態指示詞の場合

{(前文, もしくは, 指示詞の前方の同一文内に逆接接続助詞か条件形を含む用言がある場合はその用言 15)}

表 3 名詞形態指示詞の場合に与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	0	0	-10	-10	-10	-10	-10	-10

表 4 分類語彙表の分類番号の変更

意味素性	分類語彙表の 分類番号	変更後の 分類番号
ANI(動物)	156	511
HUM(人間)	12[0-4]	52[0-4]
ORG(組織・機関)	125,126,127,128	535,536,537,538
PLA(植物)	155	611
PAR(生物の部分)	157	621
NAT(自然物)	152	631
PRO(生産物・道具)	14[0-9]	64[0-9]
LOC(空間・方角)	117,125,126	651,652,653
PHE(現象名詞)	150,151	711,712
ACT(動作・作用)	13[3-8]	81[3-8]
MEN(精神)	130	821
CHA(性質)	11[2-58],158	83[2-58],839
REL(関係)	111	841
LIN(言語作品)	131,132	851,852
その他	110	861
TIM(時間)	116	a11
QUA(数量)	119	b11

125,126 については二つの分類番号が与えられる。

前文の事態を指示する例文として以下のものがある。

天狗達は間もなくやってきて、前の晩のように歌ったり踊ったりし
始めました。

(1)

おじいさんはそれを見て、こんな風に歌い始めました。

この例の「それ」の指示対象は「天狗達が歌ったり踊ったりし始めた」という事態を指示している。

また、前文の事態でなく、同一文内の逆接の接続助詞の存在する用言の表す事態を指示する場合として、次の例の「それ」がある。

おじいさんは一所懸命に歌い、そして踊りましたが、それは言葉で
は言い表せないほど下手糞でありました。

(2)

指示詞は人を指しにくいという性質を利用した規則

判定規則 1

照応詞が名詞形態指示詞の場合で、指示対象の候補となった名詞が意味素性 HUM を満足

する時、-10 点を与える。このとき指示対象の候補となった名詞の意味素性は名詞意味素性辞書 (渡辺, 黒橋, 長尾 1992) のものを用いる。

判定規則 2

照応詞が名詞形態指示詞の場合、指示対象の候補となった名詞の分類語彙表の分類番号と以下の人間を代表する分類語彙表の番号 {5200003010 5201002060 5202001020 5202006115 5241002150 5244002100} との類似レベルの最も大きいものにより得点を与える。与える得点は表 3 のとおりである。このとき用いる類似度計算には分類語彙表の分類番号を表 4 に従って変換したものをを用いる。この変更は、分類語彙表の分類番号の付け方が意味的に妥当でないところがあったためである。また、一桁目の数字に種類を設けたので、もとの分類番号よりも細かい分類となっている。

これらの規則は名詞形態指示詞で人を指すことがないという性質を用いることによって指示対象の候補を削減するためのものである。例えば、次の例文中の「それ」の指示対象は「コンピューター」であるが、「それ」の近くにある「コンピューター」以外の名詞は人を表すものではなく、「それ」は人を指さないの、指示対象が「コンピューター」であることがわかる。

太郎は最新のコンピューターを買いました。

ジョンに早速それを見せました。

(3)

「ここ」「そこ」などは場所を指示しやすいという性質を利用した規則

判定規則 3

照応詞が「ここ/そこ/あそこ」の場合、指示対象の候補となった名詞が場所を意味する意味素性 LOC を満足する時、10 点を与える。

判定規則 4

照応詞が「ここ/そこ/あそこ」の場合、指示対象の候補となった名詞の分類語彙表の分類番号と以下の場所を代表する分類語彙表の番号 {6563006010 6559005020 9113301090 9113302010 6471001030 6314020130} との類似レベルの最も大きいものにより得点を与える。与える得点は表 5 のとおりである。

例えば、次の例 (4) の「そこ」の指示対象は、場所を表す名詞「売店」であることがわかる。

太郎が公園で本を読んでいました。

コーラを買いに売店に入りました。

次郎はそこで偶然、でくわしました。

(4)

表 5 場所を指示する指示詞の場合に与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	-10	-5	0	5	10	10	10	10

「ここで」「そこで」などが接続詞として用いられる場合の規則

列挙判定規則 3

「ここで/そこで」の場合 {(前文 11)}

この規則は、「ここで/そこで」が接続詞として用いられている場合、前文を指示すると解析するための規則である。場所を指す語が近くにはない場合には、この規則によってあげられた候補が一番高い得点を持つことになり、接続詞として用いられていると解析できる。例えば、この規則により次の例文の「そこで」は接続詞として用いられていると判定できる。

歌い始めると、おじいさんは天狗が少しも怖くなくなってしまいました。
 した。(5)
そこでおじいさんは隠れていた穴から出てきてしまいました。

「そこで」を英語に翻訳する際には指示詞か接続詞であるかによって “there” か “then” に訳し分けをする必要があるが、このときこの規則が必要となる。

後方照応の場合の規則

名詞形態指示詞は、同一文内の後方照応をする場合がある。同一文内の後方照応をする場合については、他の形態指示詞も含めて文献(松岡, 村田, 黒橋, 長尾 1995)での方法から作成した規則により対処する。文献(松岡他 1995)では、名詞形態指示詞が同一文内の後方照応の他に次の文などを指示する場合も扱っているが、そのようになる場合はまれでありその規則を用いない方が精度が良いので本研究では使用していない。

その他の規則

以上で述べたものの以外に以下のような規則がある。

列挙判定規則 4

「それ/あれ/これ」や連体詞形態指示詞の場合で、その指示詞の直前の文節に用言の基本形か「～とか」などの例を列挙するような表現がある場合
 {(用言の基本形か例を列挙するような表現 40)}

列举判定規則 5

指示詞の場合

{(個体を導入 10)}

この規則は指示詞の指示対象が文章中にないときでも、システムになんらかのものを指示させるためのものである。この規則により個体を導入し、その個体を指示すると解析する。

3.2 連体詞形態指示詞の指示対象を推定するための規則

「この」「その」「あの」「こんな」「そんな」などの連体詞形態指示詞と呼ばれるものには限定指示と代行指示の二種類がある。

限定指示とは、「連体詞形態指示詞+名詞」の形で指示するものであり、例えば、次の例の下線部の「この」のようなものである。

おじいさんは天狗達の前に出ていって踊り始めました。
けれどもこのおじいさんは歌も踊りも下手糞でした。 (6)

この例では「このおじいさん」という一固まりで第一文の「おじいさん」を指示している。

また、代行指示とは、連体詞形態指示詞の部分が指示対象を持つ用法であり、「その」ならば「それの」と置き換えて考えることができる。例えば、次の例では連体詞形態指示詞の「その」が「天狗」を指示しており、代行指示である。

また、鳥の様な顔をした天狗も居ました。
その口はまるで鳥の嘴の様に尖っているのです。 (7)

以下に限定指示、代行指示の場合のための規則を示す。

限定指示の場合の規則

列举判定規則 6

「その名詞 A」の場合

ソ系の連体詞形態指示詞+名詞 A の場合

{ (名詞 A を部分文字列として含む名詞 45)

(重みが w で n 個前²の主題で名詞 A の下位語の名詞 $w - n * 2 + 10$)

(重みが w で n 個前の焦点で名詞 A の下位語の名詞 $w - n * 2 + 10$) }

語の上位下位の関係の把握は、EDR の単語辞書 ((株) 日本電子化辞書研究所 1994b) の定義文の文末の単語をその単語の上位語とする方法 (鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将 1991) で行なった。次に示すコ系に比べ、ソ系は比較的近くにあるものしか指さないで、第二項には係数 2 をつけている。

² 主題が何個前かを調べる方法は、主題だけを数えることによって行なう。主題がかかる用言の位置が今解析している文節よりも前の場合はその用言の位置にその主題があるとして数える。そうでない場合はそのままの位置で数える。

表 6 ソ系の連体詞形態指示詞の場合に与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	-10	-2	-1	0	1	2	3	4

列挙判定規則 7

コ系の連体詞形態指示詞+名詞 A の場合

{ (名詞 A を部分文字列として含む名詞 45)

(重みが w で n 個前の主題で名詞 A の下位語の名詞 $w - n + 30$)

(重みが w で n 個前の焦点で名詞 A の下位語の名詞 $w - n + 30$) }

列挙判定規則 8

ア系の連体詞形態指示詞+名詞 A の場合

{ (名詞 A を部分文字列として含む名詞 45)

(重みが w で n 個前の主題で名詞 A の下位語の名詞 $w - n * 0.4 + 30$)

(重みが w で n 個前の焦点で名詞 A の下位語の名詞 $w - n * 0.4 + 30$) }

上の三つの規則では「連体詞形態指示詞+名詞 A」の近くに「名詞 A」があれば、限定指示と解釈して「名詞 A」を指示対象の候補とする。また、「連体詞形態指示詞+名詞 A」の近くに「名詞 A」の下位語があればそれらの間に照応関係が存在すると考えられるので、それも指示対象の候補とする。これらの規則には比較的大きな得点を与えている。下位語を指示する例として以下のものがある。

おじいさんは遠くの山のむこうに見えなくなってしまうまで、遠の
いていく鶴の姿を見送るのでありました。

「あの鳥を助けてやってよいことをした」とおじいさんはひとりご
とを言いました。

(8)

この例では下線部の「あの鳥」がその前文にある下位語「鶴」を指示している。

ソ系の連体詞指示詞が代行指示をする場合の規則

判定規則 5

照応詞がソ系の連体詞形態指示詞の場合、それが係る名詞 B の用例「名詞 A の名詞 B」を検索し、名詞 A と指示対象の候補となった名詞の類似レベルにより得点を与える。与える得点は表 6 のとおりである。このとき用いる「名詞 A の名詞 B」の用例は EDR の共起辞書 ((株) 日本電子化辞書研究所 1994a) のものを用いる。

表 7 「名詞 A の口」の用例

名詞 A になるもの
ポリ袋 ルポライター 委員長 一同 家庭教師 幹部 関係者 灸 牛 国民 採用 市民 私 自分 周作 就職 庶民 人間 青木 赤ちゃん 先生 袋 谷村 担当者 炭がま 長 日本人 彼 彼ら 彼女 被災者 避難民 負傷者 兵 兵隊 弁護士 母

表 8 ソ系以外の連体詞形態指示詞の場合に与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	-30	-30	-30	-30	-10	-5	-2	0

この規則は代行指示の場合における意味的整合性を調べるための規則である。(代行指示の場合の指示対象の候補は 3.1 節であげた列挙判定規則 1 によりあげられる。)

代行指示の例として前にあげた例文 7 の下線部の「その」の指示対象は前文の「天狗」であるが、この規則では「天狗」と「その」が修飾している「口」の間の意味的整合性を調べるために「名詞 A の口」という用例を集め、この「名詞 A」と意味的に近い場合は指示対象として適切であると判定する。たとえば、EDR の共起辞書 ((株) 日本電子化辞書研究所 1994a) には「名詞 A の口」という用例は表 7 だけある。

ソ系以外の連体詞指示詞が代行指示をする場合の規則

判定規則 6

照応詞がソ系以外の連体詞形態指示詞の場合、それが係る名詞 B の用例「名詞 A の名詞 B」を検索し、名詞 A と指示対象の候補となった名詞の類似レベルにより得点を与える。与える得点は表 8 のとおりである。ソ系以外の連体詞形態指示詞は代行指示になりにくいという性質 (正保 1981) (山村毅他 1992) があるので、ソ系の場合よりも得点を低く設定している。

事態を指示する場合の規則

連体詞形態指示詞は名詞形態指示詞と同様に、前文の文末の用言が表す事態を指示する場合がある³。

3 事態を指示する場合でも代行指示か限定指示かの区別を行なう必要があるが、本研究ではこの区別は行なっていない。

表 9 「こんな(名詞)」が前文を指示するか後文を指示するかの調査結果

「こんな(名詞)」 につく助詞	は	は ない	に	に も	に は	で	で は	の	す ら	が	を	も	で は ない	合 計
前文を指示(個)	9	5	17	1	2	15	5	9	2	27	43	2	0	137
後文を指示(個)	0	0	0	0	0	0	0	0	0	22	26	4	1	53

つまり、人間の脳より優秀なパターン認識プログラムが作れない段階では、非常に複雑で面白そうな事象については、まずその画像を作って、そのデータを物理学者に吟味させる必要がある。(9)

1980年代の初頭に LEP 実験装置の設計が始まった時、この戦略が採用されたのだった。

この例の「この戦略」の指示対象は前文の用言「吟味させる」が表す事態である。このように事態が指示対象となる場合の推定は、代行指示として「この」が指すものや限定指示として「この戦略」が指すものとして適正な名詞が「この」の近くでない場合、事態を指示するようにすることで実現する。ただし、連体詞形態指示詞の場合も名詞形態指示詞と同様に同一文内に逆接の接続助詞の存在する用言があれば、その用言の表す事態を指示するようにしている。以上のことは 3.1 節の列举判定規則 2 によって実現される。

「こんな」＋名詞を解析する場合の規則

「こんな(名詞)」については次の文を指示対象とする場合がある。

おじいさんは急に天狗達と一緒に踊りなくなってきました。とうとうおじいさんは踊りだし、踊りながらこんな歌を歌いました。「天狗、天狗、八天狗」(10)

例えば、上の例の「こんな歌」の指示対象は『「天狗、天狗、八天狗」』である。ところが、「こんな(名詞)」という手がかりだけでは前方照応か後方照応かを判定することができない。そこで、「こんな(名詞)」につく助詞によって前方か後方かの判定を行なうために、1986, 1987 年の天声人語と社説の約 6 万文から 317 個の「こんな」を含む部分を抽出し、そのうち前文か後文を指示対象とする場合の「こんな(名詞)」190 個について前文、後文を指す個数を数えた。この結果を表 9 に示す。この表により、未知情報を表現する時に用いられやすい助詞「が」「を」などがつく場合以外は、前方照応であることがわかる。助詞「が」「を」がつく場合は、文献(松岡他 1995)の方法と同様に引用記号の“ ”“ ”がついている文が指示対象になりやすいなどの規則によって指示対象を推定する。

3.3 副詞形態指示詞の指示対象を推定するための規則

ソ系の副詞形態指示詞が前文を指示する場合の規則

列挙判定規則 9

「そう」などのソ系の副詞形態指示詞の場合

{(前文 30)}

例文として、以下のものがある。

「天狗、天狗、八天狗」

そう 歌ったのは数えてみますとそこに八匹の天狗が居たからです。

(11)

例えば、この例の「そう」の指示対象は前文の『「天狗、天狗、八天狗」』である。

ソ系の副詞形態指示詞が文内後方照応をする場合の規則

列挙判定規則 10

「そう/そうして/そのように」の場合で、それらが存在する文が逆接の接続助詞か助動詞

「ように」を持つ従属節である場合 {(主節 45)}

この規則は (松岡他 1995) の方法を利用したものである。

コ系の副詞形態指示詞が前文を指示する場合の規則

列挙判定規則 11

「こう」などのコ系の副詞形態指示詞の場合

{(前文 25)}

コ系の副詞形態指示詞が後ろの文を指示する場合の規則

列挙判定規則 12

コ系の副詞形態指示詞の解析の場合

{(後の文 26)}

コ系の副詞形態指示詞については次の文を指示対象とする後方照応をする場合がある。

天狗達は暫くおじいさんを見ていましたが、天狗達はとうとうこう
言いました。

「今日のお前は駄目だな。

(12)

どうして昨日の様に歌ったり、昨日の様に踊ったりできないのだ。

さあ、これを返してやるから家へ帰ってしまえ。」

上の例の「こう」の指示対象は次の文以降の発話である。前方照応となる場合は文の形が「こうして」「こうすれば」などのように典型的な形になるのでこのような場合は前方照応とし、これら以外の場合を後方照応とみなす。このための規則として次の規則を作成した。

列挙判定規則 13

「こう/こんなふうに」+条件節もしくは「こうして」の場合で、文末でない場合
{(前文 7)}

4 代名詞の指示対象を推定するための規則

列挙判定規則 1

一人称の代名詞の場合 {(一人称 25)}

列挙判定規則 2

二人称の代名詞の場合 {(二人称 25)}

一人称、二人称の代名詞はほとんど会話文章中にあらわれ、会話文章中の一人称(話し手)、二人称(聞き手)をあらかじめ推定しておくことで、ほぼ確実に推定することができる。会話文章の話し手や聞き手の推定は、その会話文章の発話動作を表す用言のガ格と二格をそれぞれ話し手、聞き手とすることによって行なう。会話文章の発話動作を表す用言は、その会話文章に「と言った。」などがつけばそれとし、そうでない場合は前文の文末の用言とする⁴。例えば、次の文章中の二人称の代名詞「お前さん」の指示対象は、この会話文の二人称の「おじいさん」である。

「明日、また（おじいさんが）参りますよ。」とおじいさんは約束
しました。
「もちろん、お前さんを（一匹の天狗が）疑う訳ではないのだが」
と、一匹の天狗がおじいさんに言いました。 (13)

この会話文の二人称が「おじいさん」であることは、その会話文の発話動作を表す動詞「言う」の二格が「おじいさん」であることから求まる。

列挙判定規則 3

三人称の代名詞の場合 {(一人称 -10) (二人称 -10)}

一般の代名詞については以下の三つの規則で解析される。列挙判定規則 4 により、主題・焦

⁴ 実際にはこのような方法では発話動作を表す用言の推定を誤ることがあるが、本論文で用いたテキストではこの方法ですべて正しく解析できた。

表 10 人称代名詞の場合に与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	0	0	3	7	10	10	10	10

点と代名詞と先行詞の距離を考慮して、優先順序を持った候補群をあげ、判定規則 1、判定規則 2 により人間である候補の得点を高くする。

列举判定規則 4

代名詞の場合

{(同一文中か前文の重みが w で主題と焦点を合わせて数えて n 個前の主題 $w - n - 2$)
(同一文中か前文の重みが w で主題と焦点を合わせて数えて n 個前の焦点 $w - n + 4$)}

判定規則 1

照応詞が代名詞の場合で、指示対象の候補となった名詞が意味素性 HUM を満足する時、10 点を与える。

判定規則 2

照応詞が代名詞の場合で、指示対象の候補となった名詞の分類語彙表の分類番号と以下の人間を代表する分類語彙表の番号 {5200003010 5201002060 5202001020 5202006115 5241002150 5244002100} との類似レベルの最も大きいものにより得点を与える。得点は表 10 のとおりである。

5 ゼロ代名詞の指示対象を推定するための規則

一般のゼロ代名詞の指示対象の候補をあげるための規則

列举判定規則 1

ガ格の省略の場合のデフォルト規則

{(重みが w で n 個前の主題 $w - n * 2 + 1$)
(重みが w で n 個前の焦点 $w - n + 1$)
(今解析している節と並列の節の主格 25)
(今解析している節の従属節か主節の主格 23)
(今解析している節が埋め込み文の場合で主節の主格 22)}

表 11 用言との意味関係から与える得点

類似レベル	0	1	2	3	4	5	6	一致
得点	-10	-2	1	2	2.5	3	3.5	4

列挙判定規則 2

ガ格以外の省略の場合のデフォルト規則

{(重みが w で n 個前の主題 $w - n * 2 - 3$)

(重みが w で n 個前の焦点 $w - n * 2 + 1$)}

主題・焦点の重みや指示詞と指示対象の候補の間の距離に応じて、それぞれの候補の得点が変わる。

複文の解析のための規則

列挙判定規則 3

複文での主格の不一致の条件となる接続助詞「ので」「ならば」が含まれる文で、主節(もしくは従属節)のガ格が省略されていて、従属節(もしくは主節)の主格が省略されずに存在してその主格につく助詞が「が」の場合、

{(従属節(もしくは主節)の主格 -30)}

複文でのガ格の省略の場合、列挙判定規則 1 により主節か従属節に省略されていないガ格の名詞があれば、それを他方に補完することを行なう。ただし、従属節の接続助詞によっては主節と従属節の主格が一致しないということが知られており(南 1974)(吉本 1986)(平井 1986)(中岩浩巳・池原悟 1995)、このような接続助詞の場合は主節と従属節の主格を他方に補完するということとは行なわない。列挙判定規則 3 は、このための規則である。

用言との意味関係を利用した規則

判定規則 1

指示対象の候補となった名詞が格フレームの格要素の意味素性を満足しない時、 -5 点を与える。

判定規則 2

指示対象の候補となった名詞と格フレームの格要素の用例の名詞との類似レベルにより得点を与える。与える得点は表 11 のとおりである。

おじいさんは地面に腰を下ろしました。		
やがて (おじいさんは) 眠ってしまいました。		
意味素性	HUM/ANI が	眠る。
用例	彼/犬が	眠る。

図 3 意味的整合性の調べ方の例

この二つの規則はゼロ代名詞の解析においてそのゼロ代名詞を格要素にとる用言との意味的整合性を調べるための規則である。判定規則 1 は、意味素性により意味的整合性を調べるもので、判定規則 2 は、用例を利用して意味的整合性を調べるものである。意味的整合性の調べ方を図 3 の例文のゼロ代名詞を例にして説明する。意味素性による方法では、指示対象の候補となった名詞が指示対象として妥当であるための条件は、その名詞に付けられた意味素性の一つが動詞の格フレームに記述された意味素性と同一もしくは下位の意味素性であることとする。例えば、図 3 のゼロ代名詞に対しては動詞「眠る」のガ格の意味素性が HUM/ANI⁵ であり「おじいさん」が HUM であることから、「おじいさん」は指示対象として妥当とする。

また、用例による判定方法では、指示対象の候補となった名詞と、動詞の格フレームに記述された名詞の用例とが意味的に類似していればその類似度に応じてその名詞は指示対象として妥当であるとする。例えば、図 3 の省略部分に対しては動詞「眠る」のガ格の用例が「彼/犬」であり「おじいさん」と「彼」が意味的に近いことから「おじいさん」は指示対象として妥当とする。

この用言との意味関係を用いた判定方法は、指示詞や代名詞の指示対象の推定にも用いた。

同一用言の複数の格要素に同じ要素が入りにくいという性質を利用した規則

列挙判定規則 4

今求める省略要素を格要素に持つ用言の他の格要素に名詞 A がすでに入っている場合、
{(名詞 A -20)}

視点を利用した規則

列挙判定規則 5

「くれる」「くださる」が補助動詞としてつく用言のガ格の省略の場合で二格に省略がある場合は二格を先に解析し、ガ格の省略に対しては {(省略部分を埋めない -5)}

この規則は視点の理論 (Kameyama 1986) を利用したものである。「くれる」「くださる」が補助動詞としてつく用言の場合は、共感度の高い二格の解析を先に行なうことによって主

5 HUM, ANI はそれぞれ人間 (HUMAN), 動物 (ANIMAL) を表す意味素性である。"/" は“または”を表す。

題などの共感度の高い名詞が二格の格要素に入り、残ったものがガ格の格要素に入ることになる。

会話文章中のゼロ代名詞のための規則

列挙判定規則 6

会話文章中で「やる」「したい」「行く」などの一人称がガ格に入りやすい用言のガ格の省略の場合、{(一人称 5)}

列挙判定規則 7

会話文章中で「くれる」「なさる」「来る」などの二人称がガ格に入りやすい用言か命令表現か疑問表現を文末に持つ文中のガ格の省略の場合、{(一人称 -30)(二人称 25)}

列挙判定規則 8

会話文章中のガ格の省略の場合、{(一人称 15)}

会話文章中でのゼロ代名詞の解析では、文末表現などからその動詞の省略された格要素の指示対象に入るべき人称を推定できる場合がある。このような場合は、その会話文章の一人称と二人称を推定することで代名詞と同様に会話文章中でのゼロ代名詞の解析を行なうことができる⁶。例えば、次の会話文は、「言う」のガ格と二格から一人称は「天狗達」で二人称は「おじいさん」である。

天狗達はとうとうこうおじいさんに言いました。

「今日のお前は駄目だな。

さあ、これを(天狗達が)(おじいさんに)返してやるから家へ

(おじいさんが)帰ってしまえ。」

(14)

「返してやる」のガ格の省略部分の指示対象は、「返してやる」に補助動詞「やる」がついていることから一人称の「天狗達」であることがわかる。「返してやる」の二格の省略部分の指示対象も、補助動詞「やる」から二人称の「おじいさん」と判定できる。「帰ってしまえ」のガ格の省略部分の指示対象は、命令表現から二人称の「おじいさん」と判定できる。

その他の規則

列挙判定規則 9

「A を B だと言う/思う」における「B だ」のガ格の省略の場合、{(名詞 A 50)}

6 工藤(工藤育男・友清睦子 1993)は文末表現から対話コーパス中のゼロ代名詞の人称を推定しているが、本研究は物語文内の会話文章を対象としており、各発話文の話し手と聞き手を推定する必要があるという点で異なる。

ドル相場は、米新政権の経済政策に対する期待の高まりなどから 130 円台に上昇した。

このドル高は、米国と欧州各国との間の政策協調をぎくしゃくさせている。

規則	各候補の得点 (点)				
	前文	個体を導入	130 円台	高まり	ドル相場
列挙判定規則 2	15	10	17	15	15
列挙判定規則 5					
列挙判定規則 1					
判定規則 6					
合計	15	10	-13	-15	-15

図 4 指示詞「この」の指示対象の推定例

6 実験と考察

6.1 実験

指示対象の推定を行なう前に構文解析・格解析を行なうが、その際の誤りは人手で修正した。格フレームは IPAL の辞書のものを用いたが、IPAL の辞書にない用言に対しては人手で格フレームを作成した。

指示詞「この」の指示対象を推定した例を図 4 に示す。これは図中の下線部の「このドル高」の指示対象を前文全体を指すと正しく解析したことを示している。これを以下に説明する。

まず、3 節で示した指示詞の列挙判定規則 2 により前文を指示対象とする「前文」という候補があげられ、それに 15 点が与えられる。また、列挙判定規則 5 により「個体を導入」という候補があげられ、それに 10 点が与えられる。さらに、列挙判定規則 1 により、主題と焦点から「130 円台」「高まり」「ドル相場」という候補があげられ、これらの候補にはそれぞれ 17, 15, 15 点の得点が与えられる。これらの候補に対して判定規則 6 を適用してみる。判定規則 6 は「(名詞 A) の (名詞 B)」の用例を利用する規則であり、この場合は「(名詞 A) のドル高」という用例を利用する。この用例の「(名詞 A)」の部分に来る名詞は EDR の共起辞書では「最近」しかなく、この「最近」との分類語彙表での類似レベルは「130 円台」「高まり」「ドル相場」ともに低く、それぞれには表 8 より -30 点が与えられる。「前文」と「特定指示として導入」は名詞でないので、この判定規則により得点を与えられることはない。この結果、合計点の最も高い「前文」という候補が指示対象と正しく推定される。

本研究による方法で指示詞、代名詞、ゼロ代名詞の指示対象を解析した実験結果を表 12 に

表 12 本研究の実験結果

テキスト	文数	指示詞	代名詞	ゼロ代名詞	合計
学習サンプル	204	87%(41/ 47)	100%(9/ 9)	86%(177/205)	87%(227/261)
テストサンプル	184	86%(42/ 49)	82%(9/11)	76%(159/208)	78%(210/268)

各規則で与える得点は学習サンプルにおいて人手で調節した。学習サンプルは、例文 (43 文)、童話「こぶとりじいさん」全文 (93 文)(中尾 1985)、天声人語一日分 (26 文)、社説半日分 (26 文)、サイエンス (16 文) であり、テストサンプルは、童話「つるのおんがえし」前から 91 文抜粋 (中尾 1985)、天声人語二日分 (50 文)、社説半日分 (30 文)、サイエンス (13 文) である。

表 13 指示詞の実験結果の内訳

テキスト	文数	名詞形態	連体詞形態	副詞形態	すべての指示詞
学習サンプル	204	83%(15/ 18)	86%(19/ 22)	100%(7/ 7)	87% (41/ 47)
テストサンプル	184	82%(14/ 17)	88%(23/ 26)	83%(5/ 6)	86% (42/ 49)

示す。また、指示詞の実験結果について名詞形態、連体詞形態、副詞形態指示詞の内訳を表 13 に示す。指示詞の解析において、指示対象が文になる場合で、前の数文もしくは後ろの数文が指示対象となる場合は、その範囲まで推定できなくても、指示対象が前方の文か後方の文かの判定を行なうことができれば正解とした。これは指示対象の範囲の推定は、原因-結果、例提示などの文の間の関係を解析した後ですべきであると考えためである。また、ゼロ代名詞の解析精度は指示対象が存在するか否かがあらかじめわかっていると仮定して解析した時の精度である。

6.2 考察

指示詞については、正解率はテストサンプルにおいても 80%を越えたので、本システムで用いた規則は有効であることがわかる。しかし、指示詞は種類が多いのでより詳細に規則を作成することでさらに高精度に解析できる可能性がある。また、本研究では「この」が代行指示になりにくいという性質を利用したが、これを利用したために正しく解析できた例が 4 例あった。

代名詞については、今回の実験テキストにおいて一人称と二人称の代名詞しか出現しなかった。その代名詞が存在する会話文の一人称と二人称を推定することでほぼ正確に推定できた。代名詞の解析において誤った主な原因は、「言う」の二格のゼロ代名詞の解析を誤り会話文の二人称の推定を誤ったためであった。

ゼロ代名詞の解析誤りとしては、分類語彙表、意味素性辞書、格フレーム辞書に誤りがある

ために誤ったものや、統語構造や補助表現から指示対象が推測できるのに規則が不十分なために誤ったものがあった。

また、理解や推論が必要なために誤ったものとして次のものがあった。

そんな状況なのに、ワシントンで開かれる主要先進7カ国の蔵相中央銀行総裁会議（G7）について各国の通貨当局は「大きな問題はないので共同コミュニケは出ない。顔合わせ中心の会合だ」と、まるで会議の意義を薄めようとしているような言い方だ。

（中略）

米新政権は近く、財政赤字削減の具体的構想を議会に示す予定である。

（段落替え）

（通貨当局が）共同コミュニケの発表を控えるのは、為替市場に過大な期待を与えたくないためだろう。

（15）

この例の「控える」のガ格の省略部分の指示対象は「各国の通貨当局」である。しかし、システムは「米新政権」を誤って指示対象と解析した。この指示対象を正しく解析できるようにするためには、そこまでの文章から共同コミュニケの発表を控えるものが通貨当局であることを理解しておく必要がある。

6.3 対照実験

指示詞と代名詞とゼロ代名詞の解析において判定規則として用例を用いる規則と意味素性を用いる規則とを用いるが、これらの有効性を調べるために表14にあげる対照実験を行なった。ここでいう用例を用いる規則は指示詞の判定規則2,4、代名詞の判定規則2およびゼロ代名詞の判定規則2とそれに対応する指示詞・代名詞の判定規則を意味する。また、意味素性を用いる規則は指示詞の判定規則1,3、代名詞の判定規則1およびゼロ代名詞の判定規則1とそれに対応する指示詞・代名詞の判定規則を意味する。「名詞Aの名詞B」の用例を用いる規則には対応する意味素性を用いる規則がないので、この対照実験ではすべての場合でこの規則を用いた。推定精度は表14のように用例を用いる方法と意味素性を用いる方法とは同程度であった。このことにより、意味素性と同様に用例を用いることができることがわかった。また、分類語彙表においては分類番号の付け方が意味的に妥当でないところがあったため分類番号を変更したが、分類番号を変更したものの方が分類語彙表の分類番号のままのものより精度が良く、妥当な変更であることが確認できた。

実験結果を考察すると用例による方法では格フレームの記述から外れた表現に対しても有効な場合があった。例えば、「言う」のニ格には人間しか入らないように格フレームに書いてある

表 14 意味素性と用例の対照実験の結果

テキスト	用例と意味素性の両方を用いる	意味素性のみを用いる	用例のみを用いる (分類語彙表の分類番号の変更)	用例のみを用いる (分類語彙表の分類番号のまま)	意味素性と用例の両方を用いない
指示詞					
学習サンプル	87% (41/47)	83% (39/47)	87% (41/47)	83% (39/47)	79% (37/47)
テストサンプル	86% (42/49)	88% (43/49)	88% (43/49)	84% (41/49)	86% (42/49)
代名詞					
学習サンプル	100% (9/ 9)	100% (9/ 9)	100% (9/ 9)	100% (9/ 9)	89% (8/ 9)
テストサンプル	82% (9/11)	64% (7/11)	82% (9/11)	55% (6/11)	64% (7/11)
ゼロ代名詞					
学習サンプル	86%(177/205)	83%(171/205)	86%(176/205)	82%(169/205)	66%(135/205)
テストサンプル	76%(159/208)	76%(158/208)	79%(164/208)	75%(155/208)	63%(131/208)

ので、意味素性による方法では次の例の二格の省略部分には鶴を補うことができない。

おじいさんは鶴をはなしてやりながら（鶴に）言いました。 (16)

しかし、用例による方法では人間と動物は表 4 により類似レベルが 1 で減点が小さく鶴を二格に補うことができる。

6.4 各規則の貢献度

本研究では様々な規則を使用した。それぞれの規則の正解への貢献度の考察を行なった。

ゼロ代名詞の解析では表層の手がかりが少ないので、動詞との意味的整合性の情報が重要になる。一方、指示詞の場合は、表層の手がかりが多く、また、指示詞であるということから人を指しにくいとわかり、動詞との意味的整合性の情報はあまり重要でなくなる。また、指示詞は、それぞれ詳細に規則化する必要がある、すべての規則が必須かつ重要である。一人称、二人称の代名詞は話し手と聞き手を把握して解析する規則が有効である。

また、会話文章において話し手と聞き手を把握することで代名詞やゼロ代名詞の指示対象を推定する規則を作成していたが、これらの規則は会話文章がよく現れる童話で有効であった。

7 おわりに

本論文では指示詞・代名詞・ゼロ代名詞の指示対象を推定するために既存の手法の整理、および、新しい手法の提案を行なった。その結果を用いて実際に解析を行なったところ、指示詞・代名詞・ゼロ代名詞の指示対象を学習サンプルにおいて87%の正解率で、テストサンプルにおいて78%の正解率で、推定することができた。また、対照実験を通じて意味的制約として意味素性と同様に用例を用いることができることがわかった。

謝辞

本研究において有意義な議論をいただいた黒橋禎夫氏に感謝いたします。最後に、本研究および実験に関して援助して下さった松岡正男氏をはじめとする長尾研究室の皆様感謝します。

参考文献

- 林四郎 (1983). “代名詞が指すもの、その指し方.” 運用 I, 朝倉日本語新講座 5, pp. 1-45. 朝倉書店.
- 平井誠 (1986). “省略語と事象間関係から見た日本語文の意味解析と文脈解析の枠組みについて.” 言語理解とコミュニケーション研究会 NLC86-8, 7-14.
- 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) 説明書.”.
- Kameyama, M. (1986). “A Property-sharing Constraint in Centering.” In *24th Annual Meeting of the Association for Computational Linguistics*, pp. 200-206.
- 金水敏・田窪行則 (1992). 指示詞. ひつじ書房.
- 国立国語研究所 (1964). 分類語彙表. 秀英出版.
- 工藤育男・友清睦子 (1993). “日本語の述部の特性を用いた省略の補完機構について.” 信学論, **J76-D-II** (3), 624-635.
- Kurohashi, S. and Nagao, M. (1994). “A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary.” *IEICE Transactions on Information and Systems*, **E77-D** (2), 227-239.
- 村田真樹・長尾真 (1996). “名詞の指示性を利用した日本語文章における名詞の指示対象の推定.” 言語処理学会論文誌, **3** (1).
- 松岡正男, 村田真樹, 黒橋禎夫, 長尾真 (1995). “表層表現を利用した日本語文章における後方照応表現の自動抽出.” 情報処理学会自然言語処理研究会 95-NL-108.
- 南不二男 (1974). 現代日本語の構造. 大修館書店.
- 長尾真, 辻井潤一, 田中一敏 (1976). “意味および文脈情報を用いた日本語文の解析—文脈を考慮した処理.” 情報処理学会誌, **17** (1), 19-28.

- 中岩浩巳・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.” 言語理解とコミュニケーション研究会 NLC95-5, 33-40.
- 中尾清秋 (1985). こぶとりじいさん 他 鶴の恩がえし、きき耳ずきん, 英訳「日本むかしばなし」シリーズ, 7 巻. 日本英語教育協会.
- (株) 日本電子化辞書研究所 (1994a). “EDR 電子化辞書 日本語共起辞書評価版第 2.1 版.”.
- (株) 日本電子化辞書研究所 (1994b). “EDR 電子化辞書 日本語単語辞書評価版第 2.1 版.”.
- 正保勇 (1981). 「コソア」の体系. 国立国語研究所.
- Takada, S. and Doi, N. (1994). “Centering in Japanese: A Step Towards Better Interpretation of Pronouns and Zero-Pronouns.” In *COLING '94*, pp. 1151-1156.
- 高橋太郎他 (1990). “特集 指示詞.” 日本語学, vol. 9. 明治書院.
- 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将 (1991). “国語辞典を用いたシソーラスの作成について.” 自然言語処理研究会 91-NL-83.
- 渡辺靖彦, 黒橋禎夫, 長尾真 (1992). “IPAL 辞書と分類語彙表を用いた単語意味辞書の作成.” 情報処理学会第 45 回全国大会予稿集, 6F-8.
- 山村毅, 大西昇, 杉江昇 (1992). “日本語指示詞の前方照応現象の分類.” 信学論, **J75-D-II** (2), 371-378.
- 吉本啓 (1986). “談話処理における日本語ゼロ代名詞の扱いについて.” 自然言語処理研究会 86-NL-56.

略歴

村田 真樹: 1993 年京都大学工学部電気工学第二学科卒業. 1995 年同大学院修士課程修了. 同年, 同大学院博士課程進学, 現在に至る. 自然言語処理, 機械翻訳の研究に従事.

長尾 真: 1959 年京都大学工学部電子工学科卒業. 工学博士. 京都大学工学部助手, 助教授を経て, 1973 年より京都大学工学部教授. 国立民族学博物館教授を兼任 (1976.2 ~ 1994.3). 京都大学大型計算機センター長 (1986.4 - 1990.3), 日本認知科学会会長 (1989.1 - 1990.12), パターン認識国際学会副会長 (1982 - 1984), 日本機械翻訳協会初代会長 (1991.3 - 1996.6), 機械翻訳国際連盟初代会長 (1991.7 - 1993.7), 電子情報通信学会副会長 (1993.5 - 1995.4), 情報処理学会副会長 (1994.5 - 1996.4). 京都大学附属図書館長 (1995 -). パターン認識, 画像処理, 機械翻訳, 自然言語処理等の分野を並行して研究.

(1996 年 3 月 7 日 受付)

(1996 年 6 月 3 日 再受付)

(1996 年 9 月 19 日 採録)