自然言語処理 108－6
(1995. 7. 20)

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

松岡 正男　村田 真樹　黒橋 禎夫　長尾 眞
京都大学工学研究科 電気工学第二専攻
〒606-01 京都市左京区吉田本町
{matsuoka,mmurata,kuro,nagao}@kuee.kyoto-u.ac.jp

あらまし　テキストや談話を理解する上で，照応表現は非常に大きな役割を果たしている．本研究では，これまであまり研究されていなかったが，高品質の自然言語理解システムを実現するためにはその処理が必要不可欠である後方照応表現を取り扱った．具体的には，日本語の表層表現を手がかりとして後方照応表現の照応詞と先行詞の抽出を行った．抽出のためのルールは，まず緩やかなパターンで後方照応表現の候補文を取り出し，それらを詳細に調べることによって作成した．テストサンプルに対する実験の結果，後方照応表現の認定は適合率 47.7%，再現率 94.2%，先行詞の正解率は 71.2%であった．

キーワード　後方照応，表層表現，照応詞，先行詞

Automatic Extraction of Cataphoric Expressions Using Surface Expressions in Japanese Sentences

Masao Matsuoka　Masaki Murata　Sadao Kurohashi　Makoto Nagao
Department of Electrical Engineering II, Kyoto University
Yoshida-honmachi, Sakyo, 606-01 Japan
{matsuoka,mmurata,kuro,nagao}@kuee.kyoto-u.ac.jp

Abstract　Anaphoric expressions play important roles in understanding text or discourse. Among anaphoric expressions, though cataphoric expressions are not processed so much before, it is necessary to process them in order to establish the high quality system of natural language processing. In this research, we extracted cataphoric expressions using surface expressions in Japanese sentences. We first extracted sentences which may contain cataphoric expressions by using general clues. Then we constructed a set of rules for extracting cataphoric expressions by checking those sentences. In recognition of cataphoric expressions in held-out texts, we obtained the correct recognition scores of 47.7% and the recall scores of 94.2%. Also we obtained the correct recognition scores of 71.2%, in retrieving antecedent of the recognized cataphoric expressions.

key words　Cataphora, Surface Expression, Anaphor, Antecedent

1 はじめに

われわれは，日常生活における会話やテキストを，さほど苦労もなく無意識のうちに理解している．理解するにあたり大きな役割を果たしているのが，照応表現である．たとえば，次の例では「こんなこと」が次の文をさし示し，2つの文が意味的につながりを持つことを示している．

お爺さんがよくこんなことを言っていた． (1)
人の痛みの分かる人間になりなさい．

このように照応表現によって，談話内における形式的なつながりや意味的なつながりを把握することができる．

照応現象は文脈照応と外界照応に分類され，さらに文脈照応は前方照応と後方照応に分類される．後方照応表現は，上の例文(1)のように，さし示される表現「人の痛みの分かる人間になりなさい.」がさし示す表現「こんなこと」よりもテキスト中において後方に存在する表現のことをいう．

計算機による前方照応表現の処理については，これまで多くの研究がなされてきたが(長尾 1976, 山村 1992, 藤澤 1993 など)，後方照応表現の処理に関する研究はほとんど行われていない．しかし，高品質な機械翻訳システムや自然言語理解システムを実現しようとする場合には，後方照応表現に対しても適切な処理を行うことが必要となる．このような背景から，本研究では日本語文章における後方照応表現の照応詞と先行詞を抽出することを試みた．

抽出は表層表現を手がかりとして行った．例文(1)では，「こんな + 名詞 + を … 言う.」という文の型から，照応詞が「こんなこと」，先行詞が「照応詞を含む文の次文以降」と推測することができる．後方照応表現の実例は非常に少ないので，まず緩い手がかりによって大量のテキストからその候補を取り出し，それらを詳細に調べることによって後方照応表現の照応詞と先行詞を抽出するルールを作成した．作成したルールの有効性は実験によって検証した．

2 照応表現の分類

ある言語表現が，その近辺に存在する言語表現が表す内容ないしは対象をさし示す場合，これらの表現は照応関係にあるとされる．この場合，さし示す方の表現を**照応詞**，さし示される方の表現を**先行詞**，照応詞と先行詞の組を**照応表現**とよぶ．照応現象は図1のように分類される(山梨 1992)．

2.1 文脈照応と外界照応

照応現象は，文脈照応と外界照応に分類される．

図 1: 照応現象の分類

文脈照応

先生が宿題を出した． (2)
彼はもうそれを仕上げてしまった．

例文(2)では，照応詞「それ」に対する先行詞「宿題」[1]が文章中に存在する．このように，先行詞が言語内のテキスト中に見られる照応現象のことを文脈照応という．

外界照応

A : あれ．眼鏡をどこに置いたかな． (3)
B : そこに，あるよ．

例文(3)では，照応詞「そこ」に対する先行詞が文章中に存在しない．このように，先行詞が言語内のテキスト中に見られず，発話における言語外の場面中に認められる照応現象のことを外界照応という．

2.2 前方照応と後方照応

文脈照応は，さらに前方照応と後方照応に分類される．

前方照応

例文(2)では，照応詞「それ」に対する先行詞「宿題」がテキスト内において照応詞「それ」の前方に存在する．このように，先行詞がテキスト中において照応詞の前方に存在する照応現象のことを前方照応という．

後方照応

多くの人がそう信じているように， (4)
彼がきっと当選する．

例文(4)では，照応詞「そう」に対する先行詞「彼がきっと当選する」がテキスト内において照応詞「それ」の後方に存在する．このように，先行詞がテキスト中において照応詞の後方に存在する照応現象ことを後方照応という．

[1]以下，各例に対して照応詞には下線を，先行詞には二重下線を引く．

3 後方照応表現の自動抽出

3.1 本研究の方針

ある種の言語現象に対する計算機処理の方式を考える場合，その言語現象の現れる実例を大量に収集し，それらを詳細に調べることによって計算機処理のための手がかりを得る，という方法が堅実で有効な手段と考えられる．しかし，後方照応を扱おうとする場合には，コーパス中にまれにしかその現象が存在しないという問題がある．そこで，次のような方法で研究を行った．

1. 研究の足がかりとして山梨(1992)，馬場(1992)，木村(1983)から後方照応表現の例を抜粋し，長尾(1992)から後方照応表現の実例を人手で収集した．
2. それらの例文から後方照応表現が含まれる可能性のある緩やかな手がかりを見つけ出し，その手がかりをパターン化して大量の文章から後方照応表現の候補の文章(照応詞を含む文の前後10文，これを候補文章とよぶ)を取り出した．
3. 取り出した候補文章を詳細に調べることにより，後方照応表現を抽出するための詳細なルールを作成した．
4. 作成したルールの有効性を実験によって検証した．

3.2 後方照応表現の候補文章を取り出すパターン

後方照応表現を取り出すために，緩やかな手がかりを見つけ出しパターン化した．この全パターンを以下に示す[2]．パターンは先行詞により大きく3つに分類し，さらに照応詞により細分類した．

Pattern1 先行詞が名詞である場合

Pattern1-1 照応詞がソ系の指示詞[3]であり，照応詞を含む節が従属節であり，主節に助詞「は」が存在する表現
(例文) 意味を知らないでそれを使うと，俗語は誤解を招くことがある．

Pattern1-2 照応詞がソ系の指示詞であり，ある名詞に連体修飾節が係り，その連体修飾節の中に照応詞が存在する表現
(例文) 雑誌にその一章が発表された彼の論文は，多くの人にインパクトを与えた．

Pattern2 先行詞が照応詞を含む文の主節である場合

Pattern2-1 照応詞がコ系の指示詞であり，照応詞に助詞「は」が接続し，照応詞を含む節が接続助詞「が，けれども」で従属節になっている表現
(例文) このことは公にはされていませんが，あれは本当は事故死ではなくて自殺なのです．

Pattern2-2 照応詞が指示詞「そう」であり，「そう」が係る文節に判定詞があり，「そう」を含む節が接続助詞「が，けれども」または助動詞「ようだ，みたいだ」で従属節になっている表現
(例文) 筆者自身もそうだが，同僚の多くは住宅ローンをまだ返せないでいる．

Pattern3 先行詞が照応詞を含む文の次文以降である場合

Pattern3-1 照応詞が指示詞「これ」または「"この"+名詞」であり，照応詞に助詞「は」が接続し，照応詞を含む節が判定詞の終止形で締めくくられている主節である表現
(例文) これは今から2年前の話である．
彼は東京に住んでいた．…

Pattern3-2 照応詞が「"この"以外の名詞修飾形態指示詞+名詞」であり，照応詞に接続する助詞が「は」以外である表現
(例文) 「ボケが軽くなった」と報告が続いたあとで，こんな話になった．
「それでも家族は …．」

Pattern3-3 照応詞が述語修飾形態指示詞である表現
(例文) イギリスの社会福祉施設運営基準は，こう記している．
「ひとつの部屋に …．」

Pattern3-4 照応詞が名詞「次」を含む表現
(例文) 命名に関する最新のニュースは次の話だ．
生まれた男の子に …．

Pattern3-5 照応詞が名詞「以下」または「後述」である表現
(例文) 以下は，中国を旅行した同僚記者の取材メモによる．
北京内燃機総廠という …．

3.3 後方照応表現を抽出するためのルール

前節のパターンによって取り出された候補文章を詳細に調べることにより，後方照応表現を抽出するための詳細なルールを作成した．3.2節のパターンの中の一部につい

[2] ただし本研究ではゼロ代名詞による後方照応表現は扱っていない．
[3] ソ系の指示詞とは「そ」ではじまる指示詞のことをいう．コ系の指示詞の場合も同様である．

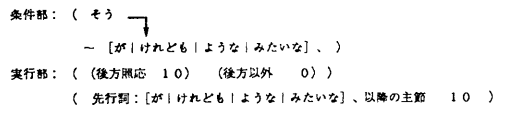

図 2: ルールの一例

ては，そのままの形でルールとした．ルールは以下の形式をとる．

条件部：(入力文の依存構造のパターン) [4]

実行部：(後方照応，後方以外 それぞれに対する得点)

(先行詞の候補　その候補に対する得点)

ルールの例を図 2に示す．図 2の例では，入力文が「そう～ [が | けれども | ような | みたいな] [5],」という文型の場合，後方照応に 10 点，後方以外に 0 点を加算し，先行詞の候補として，「“そう ～ [が | けれども | ような | みたいな],” 以下の主節」に 10 点を与えることを意味する[6]．ルールの一部を以下に示す[7]．

Rule1 Pattern1-2 and 連体修飾節が係る名詞が形式名詞である表現
((後方照応 0)(後方以外 10))
(先行詞:-,-) [8]

Rule2 Pattern2 and 主節に引用の助詞「と」がある表現
((後方照応 10)(後方以外 0))
(先行詞:引用の助詞「と」以降を除いた主節 20)

Rule3 Pattern2-2 and 「“…もあるが” | “…ではあるが”,」という表現
((後方照応 0)(後方以外 10))
(先行詞:-,-)

Rule4 Pattern3 and 照応詞を含む文の次文の文頭に “「” がある表現
((後方照応 10)(後方以外 0))
(先行詞:“「” から “」” まで 20)

Rule5 Pattern3-4 and 照応詞が「[次 | つぎ][の (数詞) つ | の (数詞) 点] であり，照応詞を含む文の次文に「第一に」がある表現
((後方照応 10)(後方以外 0))
(先行詞:次文の先頭から,「第 (数詞) に」がある文の文末まで 30)

Rule6 Pattern3-5 の表現
((後方照応 10)(後方以外 0))
(先行詞:照応詞を含む文の次文 1 文)

3.4 後方照応表現を抽出するシステム

前節のルールによって後方照応表現を抽出するシステムは，以下のように動作する．

1. 依存構造の形式をとる入力文章[9]に対してすべてのルールを適用させ，条件部が満足すれば後方照応，後方以外それぞれの得点をそのつど加算していく．また，新たな先行詞の候補があれば，その候補と得点を追加する．

2. 最終的に後方照応の得点が後方以外の得点よりも同点ないしは高ければ (ただし，一つもルールが適用されなければ後方以外とする)，後方照応表現の照応詞を出力し，先行詞の候補の中で一番得点の高い先行詞を出力する．

4 実験と考察

4.1 実験と結果

実験のテキスト (学習サンプルとよぶ) として，社説 (1985 年，1986 年，合計 11,375 文)，天声人語 (1985 年，1986 年，合計 10,494 文) を用いた．

まず，3.2節で示したパターンによって候補文章を取り出し，その中で実際に後方照応表現であるあるものに対して人手でマークを付けた．このように解析の正解/不正解を自動的に調べられるようにした上で，以降に述べる各評価の値がよくなるようにルールの追加，削除を行い，実験を繰り返した．

次に，このように作成したルールの有効性を調べるために，別のテキスト (テストサンプルとよぶ) に対して実験を行った．テストサンプルは，社説 (1987 年 ～ 1989 年，合計 19,247 文)，天声人語 (1987 年 ～ 1989 年，合計 18,909 文) を用いた．

実験の結果は次の 2 つの点から評価した．まず，後方照応表現の照応詞が正しく取り出されたかどうかを調べた (表 1)．評価としては適合率と再現率を採用した．ただし，再現率については，本来の値を求めようとすればテキスト全文 (学習サンプル 21,869 文，テストサンプル 38,156 文) を調べる必要があり，これは現実的に困難であったの

[4] 依存構造木に対するパターン照合機能は Murata.M(1993) を用いた．そこでは，依存構造木とその構成要素である文節 (単語の並び) に対するパターンを，正規表現，論理和，論理積，否定などにより指定できる

[5] 記号 “|” は “または” を意味する．

[6] 後方照応表現かどうかの決定と先行詞の決定は，この順に別々に行う．これは，ある種の表現ではそれが後方照応表現であることは推測できるが，先行詞が何であるかについては情報がないというような場合があるからである．

[7] すべてのルールについては，松岡 (1994) を参照．

[8] 記号 “(先行詞:-,-)” は，新たに先行詞の候補とその得点を追加しないことを意味する．

[9] 形態素解析 (松本 1992) と構文解析 (黒橋 1994) の処理を行った．

表 1: 後方照応表現の認定に関する実験結果

学習サンプル						
	社説		天声人語		合計	
	適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)
Pattern1-1	0.0 (0/ 2)	— (0/ 0)	0.0 (0/ 3)	— (0/ 0)	0.0 (0/ 5)	— (0/ 0)
Pattern1-2	0.0 (0/ 21)	— (0/ 0)	7.7 (1/ 13)	100.0 (1/ 1)	2.9 (1/ 34)	100.0 (1/ 1)
Pattern2-1	25.0 (1/ 4)	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (1/ 1)	40.0 (2/ 5)	100.0 (2/ 2)
Pattern2-2	100.0 (6/ 6)	100.0 (6/ 6)	100.0 (2/ 2)	66.7 (2/ 3)	100.0 (8/ 8)	88.9 (8/ 9)
Pattern3-1	0.0 (0/ 9)	— (0/ 0)	0.0 (0/ 10)	— (0/ 0)	0.0 (0/ 19)	— (0/ 0)
Pattern3-2	12.5 (2/ 16)	100.0 (2/ 2)	67.6 (25/ 37)	89.3 (25/28)	50.9 (27/ 53)	90.0 (27/ 30)
Pattern3-3	90.9 (10/ 11)	100.0 (10/10)	95.2 (40/ 42)	95.2 (40/42)	94.3 (50/ 53)	96.2 (50/ 52)
Pattern3-4	95.2 (20/ 21)	100.0 (20/20)	100.0 (8/ 8)	100.0 (8/ 8)	96.6 (28/ 29)	100.0 (28/ 28)
Pattern3-5	0.0 (0/ 0)	0.0 (0/ 0)	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (1/ 1)
合計	43.3 (39/ 90)	100.0 (39/39)	66.7 (78/117)	92.9 (78/84)	56.5 (117/207)	95.1 (117/123)

テストサンプル						
	社説		天声人語		合計	
	適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)
Pattern1-1	0.0 (0/ 5)	— (0/ 0)	0.0 (0/ 5)	— (0/ 0)	0.0 (0/ 10)	— (0/ 0)
Pattern1-2	0.0 (0/ 39)	— (0/ 0)	0.0 (0/ 21)	— (0/ 0)	0.0 (0/ 60)	— (0/ 0)
Pattern2-1	0.0 (0/ 1)	— (0/ 0)	0.0 (0/ 2)	— (0/ 0)	0.0 (0/ 3)	— (0/ 0)
Pattern2-2	100.0 (7/ 7)	87.5 (7/ 8)	100.0 (4/ 4)	100.0 (4/ 4)	100.0 (11/ 11)	91.7 (11/ 12)
Pattern3-1	0.0 (0/ 10)	— (0/ 0)	7.1 (1/ 14)	100.0 (1/ 1)	4.2 (1/ 24)	100.0 (1/ 1)
Pattern3-2	27.3 (9/ 33)	81.8 (9/11)	49.0 (24/ 49)	92.3 (24/26)	40.2 (33/ 82)	89.2 (33/ 37)
Pattern3-3	50.0 (8/ 16)	100.0 (8/ 8)	87.7 (50/ 57)	92.6 (50/54)	79.5 (58/ 73)	93.6 (58/ 62)
Pattern3-4	100.0 (34/ 34)	100.0 (34/34)	100.0 (7/ 7)	100.0 (7/ 7)	100.0 (41/ 41)	100.0 (41/ 41)
Pattern3-5	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (1/ 1)	100.0 (2/ 2)	100.0 (2/ 2)
合計	40.4 (59/146)	95.2 (59/62)	54.4 (87/160)	93.5 (87/93)	47.7 (146/306)	94.2 (146/155)

適合率 = (システムが後方照応表現と認定しかつ正解である数 / システムが後方照応表現と認定した数)
再現率 = (システムが後方照応表現と認定しかつ正解である数 / 正解の数)

で，3.2節のパターンで取り出される候補文章のみを人手で調べ，その中で実際に後方照応表現であった文の数を母数とした．これらの結果，学習サンプルにおける適合率は56.5%，再現率は95.1%，テストサンプルにおける適合率は47.7%，再現率は94.2%であった．

次に，後方照応表現の先行詞が正しく抽出されているかどうかを調べた(表2)． 調査した対象は，システムが後方照応表現と認定しかつ実際に後方照応表現であったものである．評価として正解率を採用した．学習サンプルにおける正解率は78.6%，テストサンプルにおける正解率は71.2%であった．

4.2 考察

本研究では，後方照応表現をできるだけもれなく抽出するということを目標としてルールを作成した．よって，後方照応表現の認定に関する実験結果では，再現率については高いが適合率ついては低いデータが得られた．適合率については，ルール別にかなりの差があった．適合率が高いものは，Pattern2-2,Pattern3-4,Pattern3-5 に関するルールで，これらの表現は後方照応となる文型が決まっているために高い適合率が得られた．一方，適合率が低いものの例は，Pattern1,Pattern2-1,Pattern3-1,Pattern3-2,Pattern3-3 に関するルールであった．これらのルールは前方照応とも後方照応ともとれる表現であり，頻度的には前方照応になる表現の方が圧倒的に多いことから，適合率はこのような低い値になった．誤りの例としては，次のようなものがある．

(例文)

国鉄の「分割民営化」の具体案をつくる国鉄改革推進本部が発足した．その基礎となる再建監理委の意見を読んで，いちばん疑問に思ったのは貨物のことだ．

この例では Pattern1-2 が適用されるが (Pattern1-2 はそのままの形でルールとなっている)，3.3節の Rule1 は適用されず，結果として「意見」を先行詞とする後方照応表現と認定された．しかし，実際には先行詞は直前の文であり，前方照応表現である．このルールの場合，先行詞の候補「意見」を「その」に代入してみて,「意見の基礎」という表現が可能であるかどうかを意味的に判断する必要が

表 2: 後方照応表現の先行詞の認定に関する実験結果

	学習サンプル			テストサンプル		
	社説	天声人語	合計	社説	天声人語	合計
Pattern1	— (0/ 0)	0.0 (0/ 1)	0.0(0/ 1)	— (0/ 0)	— (0/ 0)	— (0/ 0)
Pattern2	85.7 (6/ 7)	67.7 (2/ 3)	80.0(8/ 10)	71.4 (5/ 7)	75.0 (3/ 4)	72.7 (8/ 11)
Pattern3	65.6 (21/32)	85.1 (63/74)	79.2(84/106)	67.3 (35/52)	73.5 (61/83)	71.1 (96/135)
合計	69.2 (27/39)	83.3 (65/78)	78.6(92/117)	67.8 (40/59)	73.6 (64/87)	71.2 (104/146)

各枠内の数字は，正解率 [%] を示している．
正解率 = (先行詞が正しく抽出できた後方照応表現の数 [個]
/ システムが後方照応表現と認定しかつ正解である数 [個])

ある．

先行詞の認定に関する実験結果では，Pattern1 については実験データが存在しなかった．Pattern2 については，先行詞は主節部をさすとほぼ決まっているので高い正解率が得られた．Pattern3 については，引用記号 (“「”, “」”) が手がかりとなる場合が多く，また，頻度的に次文の 1 文をさす場合が多いために高い正解率が得られた．誤りの例としては，次のようなものがある．

(例文)

断片的な情報から，次のようなことがうかがえる．高層ビルの倒壊がきわめて多い．‥‥．断水，停電，ガスもれなどが発生している．これらの点は，新しい都市型の震災を考える場合の教訓になる．

この例は，Pattern3-4 の例である．この場合，引用記号はなくかつ Rule5 が適用されないので，先行詞は次の 1 文と判定される．しかし，実際には後ろの複数の文が先行詞である．このような場合は文と文のつながり，すなわち文章構造を考慮に入れる必要がある．

なお，学習サンプル，テストサンプルともに解析誤りの原因の種類は，以上の誤りを含め同じものであった．

5 おわりに

本研究では，表層表現 (主に文型の情報) を手がかりとして，日本語文章における後方照応表現の抽出を行った．学習サンプルにおける適合率は 56.5%，再現率は 95.1%，先行詞の正解率は 78.6%であった．テストサンプルにおける適合率は 47.7%，再現率は 94.2%，先行詞の正解率は 71.2%であった．

本研究では，照応詞と先行詞の整合性を検査するルールは作成していないが，今後システムの正解率を向上させるためには，意味の情報を使用することによって照応詞と先行詞の整合性を検査する必要がある．さらに，節レベル，および文レベルでのつながりの情報，すなわち文章構造の情報を取り入れることも必要である．

参考文献

木村英樹 (1983). 「こんな」と「この」の文脈照応について，日本語学, Vol.5, pp.71-83.

黒橋禎夫, 長尾 真 (1994). 並列構造の検出に基づく長い日本語文の構文解析, 言語処理学会, **1**(1).

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, Vol.17, No.1, pp.19-28.

長尾真 (1992). 人工知能と人間, 岩波新書.

馬場俊臣 (1992). 指示詞—後方照応の類型について，表現研究, Vol.55, pp.20-27.

藤澤伸二, 増山繁, 内藤昭三 (1993). 日本語文章における照応・省略現象の基本検討，情報処理学会, Vol.34, No.9, pp.1909-1918.

松岡正男, 長尾真 (1994). 日本語文章における後方照応表現の自動化抽出, 卒業論文, 京都大学工学部電気系

松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真 (1992). 日本語形態素解析システム JUMAN 使用説明書 version1.0, 京都大学工学部長尾研究室.

Murata.M and Nagao.M(1993). Determination of referential property and number of nouns in Japanese sentences for machine translation into English, TMI'93, pp.218-225.

山梨正明 (1992). 推論と照応, くろしお出版.

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，電子情報通信学会, Vol.J75-D-II, pp.371-378.