ゼロ照応問題に関する分析のためのコーパスの構築とその利用

近藤 駿佑　　指導教員：山村 毅

1 はじめに

文章では，何回も同じ単語を使う場合にその単語を指示詞で表現したり，話の流れから書かなくても理解できるような場合にはそれを省略したりすることがある．このような表現を照応表現という．自然言語処理システムを構築していく上でこのような照応表現が何を表しているのかを明らかにすることは不可欠である．これを明らかにするシステムを構築するためには，照応現象の分析をすることが必要であるが，今までの研究では省略を含む文の少なさから完全な分析を行えているとは言えない．また，システムの構築に機械学習を用いるに当たっては大量の省略に関する正解文が不可欠である．そこで，本研究では省略が出現している文を集めたコーパスを作成し，それを用いて照応現象の調査を行う．

2 ゼロ照応解析

2.1 ゼロ照応解析の例

ゼロ照応解析とは，文の中で省略されている部分であるゼロ照応詞を明らかにする解析のことである．省略部分が指している内容を先行詞という．

太郎は朝ごはんを食べた。そして，(ϕ) シャワーを浴びた。

この文において ϕ はゼロ照応詞を表し，この ϕ が指している内容は太郎なので，先行詞は太郎ということになる．

2.2 先行研究

中岩らは，動詞の意味で先行詞の格を特定する手法を提案している．[1] そこでは，動詞を意味から 107 種類のグループに分ける「用言意味属性」に基づいてゼロ照応詞が係る動詞と先行詞が係る動詞をそれぞれ分類し，その結果とゼロ照応詞の格の組み合わせで先行詞の格を決めるルールを 116 個作成している．そして，このルールに基づいて先行詞の格を特定するという流れで解析を行っている．実験では 44 文に対して作成したルールを適用する．その結果，44 文中 36 文の解析を正しく行うことができた．先行詞の格が「ガ」格となるものが 35 文，「ヲ」格となるものが 1 文であった．しかし，調査した文の数が少ないという点と，分類したグループの組み合わせだけでも 10000 通りを超えるにもかかわらず 100 ほどしかルールが整備されていないというのが問題点である．

3 コーパス

ゼロ照応現象の分析や，機械学習によるゼロ照応解析を行うために，省略を含む文を集めたコーパス必要がある．そこで，ゼロ照応現象に関するコーパスを構築する．2005 年毎日新聞の記事から抽出した 3800 文を構文解析システム KNP に入力し，その結果を用いてゼロ照応詞を検出する．そのあと，KNP の照応解析システムで特定した先行詞を人手で確認し，間違っていた場合は修正する．そしてその照応に関する情報を付与してコーパスを構築するという流れである．コーパスの中のデータの記述の例を図 1 に示す．1 文目の例の場合，ゼロ照応文に対してゼロ照応詞が係る動詞「浴びる」，照応詞の格「ガ」，先行詞が係る動詞「食べる」，先行詞の格「ハ」，先行詞の要素「僕」が照応に関連する情報であることを表している．

僕はご飯を食べた。そのあとシャワーを浴びた。

浴びる　ガ　食べる　ハ　僕

早大は法大に勝った。連勝を 3 に伸ばした。

伸ばす　ガ　勝つ　ハ　早大

図 1　コーパスにおけるデータの記述

4 ゼロ照応現象の調査

3 で作成したコーパスを用いて，ゼロ照応に関してどのような傾向があるか，また現在言われている傾向が正しいものであるかの調査を行う．

4.1 ゼロ照応詞の格と先行詞の格の関係

4.1.1 表層格の関係

ゼロ照応詞の検出は，主に動詞の格フレームを用いて行われるため，ゼロ照応現象と格は深く関係している．このことから，ゼロ照応詞の格に対する先行詞の格の種類に何らかの傾向が現れると考え，調査を行った．例を用いて調査方法を説明する．

僕はご飯を食べた．そして(ϕ ガ) シャワーを浴びた．

この例の場合，ゼロ照応詞の格は「ガ」格であり，対して先行詞は「ハ」格である．このように照応詞の格に対してどの格がどのぐらいの割合で出現しているかを，3 で作成したコーパスを用いて調査する．表 1 に調査結果を示す．

表 1　調査結果

		先行詞の格				
		ハ	ガ	ニ	ヲ	合計
照応詞の格	ガ	2523 70.5%	801 22.3%	58 1.6%	69 1.9%	3577 100%
	ニ	32 32.3%	25 25.2%	29 29.3%	7 7.1%	99 100%
	ヲ	26 24.3%	18 16.8%	5 4.7%	54 51.1%	107 100%

照応詞が「ガ」格の場合は先行詞として 90% 以上が「ハ」格か「ガ」格を選んでいることが分かる．特に「ハ」格が 70% を占めていることから，主語の省略の場合は主題が先行詞となりやすいといえる．

4.1.2 動詞の種類別の関係

動詞の種類から先行詞の格を決定する方法は，先行研究で述べた中岩らの方法が存在する．しかし，動詞の分類の数に対してルールの数が少ないという点と実験に用いた文章が 44 文と少ない点の 2 つの問題点から，この方法が必ずしも有用であると

はいいがたい．そこで，動詞の種類ごとに先行詞としてどの格がどのくらいの割合で出現するかの傾向を調べることで，この方法が有用であるかを明らかにする．本調査では，動詞のグループ分けに概念辞書を利用することで，「用言意味属性」を用いた 107 種類から，「現象」「行為」「移動」「変化」「状態」の 5 種類に変更した．そして，ゼロ照応詞が係る動詞のグループ，先行詞が係る動詞のグループ，ゼロ照応詞の格の組み合わせごとに先行詞の格にどのような傾向が現れるかを調査した．ただし，組み合わせの事例数が 50 文以上あるものに限定している．この結果，ゼロ照応詞が「ガ」格のものに限定されてしまった．結果の一部を表 2 に示す．

表 2　照応詞が係る動詞が「行為」の場合

		先行詞の格				
		ハ	ガ	ニ	ヲ	合計
先行詞が係る動詞	「行為」	318 63.7%	145 29.5%	11 2.2%	11 2.2%	499 100%
	「移動」	248 75.3%	57 17.3%	5 1.5%	4 1.2%	329 100%
	「現象」	72 62.6%	38 33.0%	1 0.9%	3 2.6%	115 100%
	「変化」	39 59.5%	26 37.7%	1 1.4%	3 4.3%	69 100%

どの組み合わせでも，目立った違いというのは見られなかった．このことから照応詞「ガ」格のものに関しては動詞の意味による先行詞の格の違いはあまり見られず，有用ではないことが分かった．

4.2　ルールによる先行詞の特定

ルールを用いた照応解析には，代名詞の中身を特定する CogNIAC の方法が存在する．CogNIAC は，先行詞候補を対象文章中の名詞全てとし，まず代名詞の性や数から先行詞候補を全て削除する．例えば「He」であれば先行詞候補から「単数の男」以外のものを削減する．そのうえで，先行詞候補の場所や個数，照応詞と先行詞候補の距離などを用いて先行詞を限定する 8 つのルールを適用し，先行詞が一つに決まった段階でそれを先行詞としている．本研究では，この方法をゼロ照応用に変更し，ゼロ照応解析が可能であるかを調査する．ゼロ照応詞では代名詞のように先行詞候補を性と数で限定することはできないため，先行詞候補の削減に動詞の格フレーム情報を用いる．

太郎は次郎にゲームを渡した．(φガ) とても喜んだ．

この例の場合，「渡す」の「ガ」格がゼロ照応詞である．

先行詞候補削減前:　{ 太郎，次郎，ゲーム }
先行詞候補削減後:　{ 太郎，次郎 }

「渡す」の「ガ」格には意味を考慮すると基本的に人物が入るので，「ゲーム」が削減される．このように先行詞候補の削減を行う．そして，CogNIAC のルールを元に作成した 6 つのルールを順番に適用し，先行詞が一つに決まった段階でそれを先行詞とする．6 つのルールを表 3 に示す．

ルールを適用した結果を表 4 に示す．全体の正解率は 81.1% と特に高い数字であったので，この 6 つのルールは有用であることが分かった．また，ルール 4 とルール 5 の正解率の高さから，前の文の主語が先行詞となる文が多いことが分かり，1 度先行詞となった名詞はその後も先行詞となりやすいこともわかる．

表 3　ルールの詳細

ルール 1	先行詞候補が対象文章中に 1 つしかない場合， それを先行詞とする
ルール 2	ゼロ照応詞が出現する文とその一つ前の文の中に先行詞候補が 一つしか存在しない場合はその候補を先行詞とする
ルール 3	ゼロ照応詞が出現する文に先行詞候補が一つしかない場合は その候補を先行詞とする
ルール 4	ゼロ照応詞が主語であり，その 1 文前の主語が先行詞候補に 含まれるのであれば，それを先行詞とする
ルール 5	3 つ以上文があり，ゼロ照応詞を含む文が連続する場合，1 つ目の ゼロ照応詞の格がその文の格順序（ハ格＞ガ格＞ニ格＞ヲ格） の中で最も高い場合は 2 つ目のゼロ照応詞の先行詞を 1 つ目のゼロ照応詞と同じにする
ルール 6	先行詞候補の中でゼロ照応詞に最も近いものを先行詞とする

表 4　ルール適用結果

ルール	正解数	不正解数	正解率
ルール 1	1553	0	100%
ルール 2	314	129	70.9%
ルール 3	278	154	64.4%
ルール 4	432	67	86.6%
ルール 5	223	62	78.2%
ルール 6	226	294	43.5%
合計	3026	706	81.1%

5　まとめ

本研究では，照応現象の分析や機械学習の際の学習データに使われる文を増やすために，ゼロ照応に関するコーパスの作成を行い，またそれを用いてゼロ照応現象についての調査を行った．コーパスには毎日新聞のデータを用いて 3800 文のゼロ照応現象を含む文を記述した．そして，これを用いてゼロ照応現象について調査をした結果，90% 以上が先行詞に「ハ」格や「ガ」格を選ぶことが分かった．そして，照応詞が「ガ」格のものに関して，動詞の意味を用いた先行詞の格の特定は有用ではないことが示された．また，1 度先行詞となった名詞はその後も先行詞となりやすいこともわかった．今後の課題としては，今後の機械学習を用いた照応解析の精度を高めていくために学習データを増やす必要があることから，コーパスの文章を 10000 文，20000 文と増やしていくことなどがあげられる．

参考文献

[1] 中岩浩巳，日英機械翻訳におけるゼロ代名詞照応解析に関する研究，2002