

統計的手法を用いた日本語文の照応・省略における特徴解析

3 R-10

小林伸嘉

牧野武則

東邦大学大学院理学研究科

1. はじめに

一般の自然言語において容易に判断できる要素は省略され、既に文章中に現れた要素は指示詞などを用いて照応をとることが多い。

しかし、現実問題として、分かりきっているものすべてを省略したのでは文として成り立たず、同様に指示詞だらけの文というのも理解しづらい。かといって文章に必要な情報を全て埋め込むのでは、何度も同じ事を繰り返し述べることになると同時に、文そのものも長く、これも理解しづらいものになってしまう。

現在の照応・省略文生成の研究においては、どのように省略・照応を作り出すか、といったところに研究の主眼が置かれており、どの程度の省略・照応が適切であるのか、といった点にはあまり触れられない場合が多い。[1]

本研究では文中に現れる照応・省略現象に着目し、どのように照応・省略がなされているかについて統計・分類を行うと同時に、照応・省略のなされないに文に対し、その理由を明確なものとすることによって現象の法則性を見出す。

2. 解析データ

日本語の省略現象における顕著な例として、提題のは格が各動詞に対するが格のゼロ代名詞として照応をとるケースが挙げられる。[2]

そこで、今回の検証では日本電子化辞書研究所のEDRコーパスから朝日新聞の記事データをサンプル文として、提題を伴うもの1368文から絞り込みを行い、それらの文の表層情報データと、EDR単語辞書より取り出した動詞の格情報を省略語検出の際の参考として用いた。

3. 省略表現

本研究での照応・省略現象の検出には、文中の動詞を検出し、その動詞の要求格情報を名詞句と照らし合わせていく方法を取った。このうち、

Statistic analysis of anaphora and ellipsis in Japanese sentence.

Nobuyoshi Kobayashi Takenori Makino

Department of Information Sciences,

Faculty of Science, Toho University

2-2-1 Miyama, Funabashi, Chiba, Japan 274

指示詞が用いられているものや格情報に対応する句が存在しないものが、照応・省略されている疑いのあるもの、ということになる。

しかし、実際には動詞が取り得るすべての表層格フレームを埋めることのできる文章はごく僅かであり、文章外照応や背景知識、一般常識等も含めると、厳密には殆どの文に何かしらの照応・省略がなされていることになってしまう。

今回は文中に現れる照応・省略のうち、提題への照応についてを扱う。提題は「が格」に継承されてかかるケースが多いが、既に「が格」を持つ動詞には一文一格の原則上継承されない。この原則をもとに、「が格」を持った動詞を含むものについて絞り込みを行ったところ、86文が該当し、それらをまず、

(1) 通常通りの表層格の係り受けで照応されるもの。

(2) 埋め込みや受け身により、表層格の交代が起こっているもの。

に分類し、さらにこれらにおける

(1) 「は格」に照応をとる格の種別。

(2) 「が格」を持つ動詞とは格の結びつき。

に対して統計を行った。

4. 照応・省略されない表現

一度文中に出てきた語は省略されるのが普通であるが、時にそれがなされないケースがある。同じ語でも別の対象を表す場合が例として挙げられるが、ここではまず単純に同じ名詞が文中に複数現れている文について抽出を行った。

このような文は112文あり、これらの名詞間の関係について

(1) 異なる語から修飾を受けたり、熟語を構成しているもの。

例：「経済格差」と「所得格差」

(2) 先行詞に対して部分省略された語が照応詞となっているもの。

例：「1次試験」と「1次」

(3) 数量や時制が異なっているもの。

例：「35大学」と「32大学」

(4) 同種の概念を持つ語が文中に複数あるため、照応・省略によって係り先が不安定になるもの。

例：「日本とアメリカ」と「日本」

(5)単純に同じ要素が複数現れているもの。
のように分類し、それぞれの関係について出現回数を統計した。

5. 結果

3・4節で述べた統計を行った結果、次のようなデータを得た。表1は「は格」に対しゼロ代名詞化の照応をとった格を、文中に現れた動詞ごとに統計を取ったものである。縦軸は、「は格」から数えて何個目の動詞の要素であるかを距離としている。尚、受け身や埋め込み文に関しては表層格の交代を施した上での計測を行っている。

表2は既に「が格」を持つ文において、「は格」がどのように照応をとっているかを表したものである。

表3は同じ名詞が現れたものから、その原因と頻度を4節に基づいた上で統計を取ったものである。

動詞の 距離	表層格の交代なし			表層格の交代あり		
	が格	を格	に格	が格	を格	で格
1	46	8	1	8	13	1
2	16	0	2	2	2	0
3	6	3	1	1	3	0
4	7	0	0	0	1	0

表1:は格に照応をとる格の統計

	表層格の交代 なし	表層格の交代 あり
は格からの照応なし	59	15
を格で照応	4	5
に格で照応	3	0

表2:が格を持つ動詞とは格の結びつき

名詞間の関係	出現回数
修飾・熟語の構成による問題	71
部分的な省略表現	7
数量・時制の問題	11
係り先不安定	13
同一要素	10

表3:照応・省略されない表現における名詞間の関係

6. 考察

今回の統計・分類の過程において、特徴的ないくつかのパターンについて述べる。

まず、通常の文では「は格」が各動詞に対しが格として共有されるケースが多いが、受け身においては表層格の交代によりを格としての共有が多くなる。また、埋め込み文においては「は

格」が「を格」として共有される以外に、埋め込みおよび引用文内の動詞間で「が格」が共有される文が60%程見られた。

また、距離に関しては原則として「は格」に近い動詞が照応されやすいことを裏付けている。しかし、特に日本語においては後方のメイン動詞に係り受けようとする特性があるため、単純に省略の重みとして使うのではなく文の長さも考慮に入れた上での処理が必要となる。

表2ではすでに「が格」を持った動詞はそのほとんどが独立し、共有要素を持たないことを示しているが、まれに「を格」や「に格」で共有されるケースが見受けられる。特に表層格の交代が行われる場合においては「を格」での共有が多くなるので、解析や生成においては柔軟な対応が必要となる。

照応・省略表現の生成においては、従来行われてきた研究によってある程度の解決が行われており、実際に3節における(1)や(2)においては共起情報を用いるなどの方法で解決の見通しがつく。また、(3)や(4)では照応判定規則[2]の適用によって解決できる。(5)に関しては、本来これらは照応・省略されても問題のないものであるにも関わらず、照応表現のなされなかった文であり、自然な照応・省略表現をもつ文の生成を目標とするならば、このような文が実際の文章にも存在することを考慮に入れた上で生成を行わなければならない。

今後の課題として、提題以外の照応・省略現象における統計と特徴解析、統計データをもとにした照応・省略表現の解析・生成モデル化とシステムの構築を行う。

参考文献

- [1]大石 一昭、唐沢 博：日本語の複文・重文における照応表現の生成、情報処理学会第54回研究報告 6B-4(1997)
- [2]村田真樹、長尾真：用例や表層表現を用いた日本語文掌中の指示詞・代名詞・ゼロ代名詞の指示対象の推定、言語処理学会論文誌 Volume 4 Number 1(1995)
- [3]中岩 浩巳、池原 悟：語用論敵・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、言語処理学会論文誌 Volume 3 Number 4(1996)
- [4]藤崎 博也、田島 研、大野 澄雄：日本語テキストにおける省略・照応の分析とその補完方法の検討、情報処理学会第53回研究報告 2L-8(1996)