2L－8

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

藤崎 博也　田島 研　大野 澄雄

東京理科大学　基礎工学部

1. はじめに

人間同士の言語によるコミュニケーションでは、発信者は情報を言語に変換する際に省略・照応を多用するが、受信者は発信者と共有している知識を用いて、これらを補完し内容を正しく理解することができる。このような処理を機械に正しく行わせることは、現在の自然言語処理技術では非常に困難で、未解決の大きな課題である。本論文では、日本語を対象とし、機械に省略・照応を正しく処理させるための基礎として、種々のテキストにおける省略・照応を、補完に必要な知識に基づいて分類し、各々について頻度や補完に必要な参照の範囲を調べた結果について述べる。

2. 分析対象としたテキスト

本研究で用いたテキストは、NHK-AM放送の「気象通報」の冒頭の天気概況(180例、1291文、総文字数48,600文字)、NHK-FM放送の昼のニュース(2日分、136文、総文字数9,673文字)、日本音響学会研究用連続音声データベースに収録されている模擬対話「観光・旅行案内」(4対話、467文、総文字9,389文字)を書き起こしたテキスト3種類を用いた。それぞれのテキストには表1のような特徴があり、1文字あたりの平均文字数は天気概況が37.6文字、ニュースが71.1文字、模擬対話が20.1文字である。

表1. テキストの特徴

テキストの種類	話題 限定	話題 非限定	相手 特定	相手 不特定	方向性 一方向	方向性 双方向
天気概況	○			○	○	
ニュース		○		○	○	
模擬対話		○	○			○

Analysis of ellipsis and anaphora in Japanese text from the viewpoint of their restoration
Hiroya Fujisaki, Ken Tajima and Sumio Ohno
Science University of Tokyo
2641 Yamazaki, Noda, 278, Japan

3. 日本語における省略現象の分析

本論文では、省略を1)文脈に基づくもの、2)背景知識に基づくもの、3)一般常識に基づくものの3種類に分ける。それを更に細分化したものを表2に示す。なお、ここでは省略が起きている場所を省略箇所、省略された語句のことを省略対象と呼ぶこととする。

表2. 省略の分類

1 文脈に基づく省略
(1-1) 省略対象が相手発話内にある場合
(1-2) 省略対象が自己発話内にある場合
(1-3) 後方に省略対象がある場合
2 背景知識に基づく省略
(2-1) 話者間で対話の前提となる知識に基づく省略
3 一般常識に基づく省略
(3-1) 一般的な常識から、容易に推測できる事項の省略
(3-2) 動詞の特性による省略
(3-3) 助述表現や尊敬語、謙譲語による省略
(3-4) 代用表現による省略

先に述べた3種類のテキストにおける省略箇所を検出し、表2に従って省略の分類を行い、それぞれの頻度を求めた結果を表3に示す。

表3. テキスト別による省略の種類と頻度

テキストの種類	省略の種類	1文あたりの省略の数
天気概況	文脈	0.536
	背景知識	0.001
	一般常識	0.046
ニュース	文脈	0.338
	背景知識	0.022
	一般常識	0.125
模擬対話	文脈	0.317
	背景知識	0.229
	一般常識	0.366

この表の数値は、表1に示したテキストの特徴の差を具体的に表わしており、特に天気概況と模擬対話とが極めて対照的であることを示している。

次に、省略箇所と省略対象の位置との距離の頻度分布を求めた結果を表4に示す。ここで、距離とは

省略箇所と最も近い省略対象の位置との間に介在する文の数で表わし、両者が同一文中にある場合を距離0とする。この距離は、補完の際に必要な参照の範囲を示している。

表4. 省略における距離の頻度分布 [%]

距離	0	1	2	3以上
天気概況	3.2	92.9	3.9	0.0
ニュース	8.6	68.6	17.1	5.7
模擬対話	2.1	50.0	15.5	32.4

ここでも天気概況と模擬対話は対照的であることが顕著に現われている。

4. 日本語における照応現象の分析

ここでは、照応を1) 直接照応、2) 間接照応、3) テキスト外照応の3種類に分ける。以下では、「これ」「それ」「あの」「そう」などの指示代名詞や連体詞などを照応語と呼び、照応語が指している語句を照応対象と呼ぶこととする。

省略の場合と同様に、3種類のテキストにおける照応の検出と分類を行い、それぞれの頻度を求めた結果を表5に示す。

表5. テキスト別による照応の種類と頻度

テキストの種類	照応の種類	1文あたりの照応の数
天気概況	直接照応	0.102
	間接照応	0.116
	テキスト外照応	0.0
ニュース	直接照応	0.103
	間接照応	0.191
	テキスト外照応	0.074
模擬対話	直接照応	0.105
	間接照応	0.058
	テキスト外照応	0.002

1つの文が比較的長いニュースでは間接照応が多く、逆に1文の短い模擬対話では少ない。間接照応の頻度では、ニュースと模擬対話が対照的である。

次に、省略の場合と同様に、照応語と照応対象の位置との距離を定義し、その頻度分布を求めた結果を表6に示す。

表6. 照応における距離の頻度分布 [%]

距離	0	1	2	3以上
天気概況	0.7	51.8	41.0	6.5
ニュース	0.0	69.0	10.3	20.7
模擬対話	5.3	51.3	15.8	27.6

この表の数値も、ニュースと模擬対話とが対照的な性質を持っていることを反映している。

5. 考察

省略・照応の対象の補完については、Sidner[2] は焦点を探すことが探索範囲をせばめる役割があるとしているが、計算機上で省略補完を行う場合には、まず文章の表層構造から一意に決まる省略対象を補完しておくことが良いと考えられる。例えば、「お願いします。」という文は、表2の(3-4)代用表現による省略であるが、省略対象は主語「私が」であることが一意に決定できる。最初にこれらを補完することにより、省略対象の探索範囲を縮小し、処理の効率を向上させることができる。

6. おわりに

本論文では、天気概況、ニュース、模擬対話という3種類のテキストを分析対象として、省略・照応について分析した。その結果、テキストによっては、省略・照応がそれぞれ偏って起きていることが明らかとなり、さらに省略・照応箇所とそれぞれの補完対象との距離にも特徴があることが明らかとなった。

今後は、省略・照応の生成・理解処理を計算機上で実現し、これらを含む自然言語処理システムを構築する予定である。

参考文献

[1] 藤崎 博也, 亀田 弘之 : "天気予報を対象とする言語表現と知識表現との相互変換," 信学論 (D), J67-D, 8, pp. 924-931 (1984).

[2] Sidner, C. L. : "Focussing for interpretation of pronouns," American Journal of Computational Linguistics, 7(4), pp. 217-231 (1981).