「の」型名詞句に対する形容詞の係り先解析

4R－3

森内 昭雄‡ 中井 慎司† 池原 悟† 大西 真理子†
‡ワールドファッション エス・イー †鳥取大学工学部

1 はじめに

機械翻訳における問題点の一つに、言語表現の構造の曖昧性の問題がある。日本語では、特に名詞句の構造が曖昧である。従来の文法的情報の範囲のみによる構文解析では、名詞句の構造を一意に決定することが困難であった。そこで本研究では、形容詞＋「の型名詞句」(形容詞＋名詞＋の＋名詞)における形容詞と名詞との意味的結合力に着目した形容詞の係り先決定法を提案する。

2 名詞句とその解析

2.1 名詞句の曖昧さ

本論文では、「形容詞＋名詞A＋の＋名詞B」の型の名詞句を対象とする。この名詞句の曖昧性は、

【形容詞→名詞A】 例：白い 門 の 下 (白い→門)

【形容詞→名詞B】 例：長い 北国 の 夜 (長い→夜)

の二つの場合のみと考えられる。本研究の目的は、形容詞がA、Bどちらの名詞に係るかを判定することである。この判定をするため、「形容詞＋名詞」の頻度統計をとり、その頻度より形容詞と名詞の結合力を計算し、結合力の強い方を係り先とする。

2.2 単語意味属性による名詞のグループ化

日本語には多くの種類の名詞が存在するため、個々の形容詞と名詞の組を考えた場合、大量のデータが必要である。そこで本論文ではこのデータ数を削減するため、単語の意味属性体系[1]を用いた係り先決定法を提案する。

2.3 一般名詞意味属性体系

一般名詞意味属性体系とは、概念化された対象と単語との対応関係を木構造(約2,700ノード、12段)の形式に整理分類したものである。この一般名詞意味属性を係り先の名詞に付与することにより、名詞をグループ化できる。このため直接名詞を使用する場合に比べて汎用的な規則を生成できる。

3 名詞意味属性による係り先決定法

3.1 基本的な手法

本研究では以下の手順で形容詞の係り先を決定する。

(1) 名詞句(形容詞＋名詞)の名詞を意味属性に置き換え、その頻度統計をとる。
(2) 形容詞＋「の型名詞句」(形容詞＋名詞A＋の＋名詞B)の名詞を意味属性に置き換える。
(3) (2)で得られた結果を「形容詞＋名詞意味属性α」、「形容詞＋名詞意味属性β」に分ける。
(4) (1)の頻度統計より、それぞれの頻度(α、βの頻度)を調べる。頻度の大きい方を係り先に決定する。

3.2 問題点とその解決策

本手法には以下の問題点がある。

(1) 頻度統計がスパースになる。
(2) 共起頻度が正しい意味的結合力を示すとは限らない。
(3) 形容詞の直近に対する優先度を考慮していない。

そこで、これら問題点の解決策として、(1)～(3)それぞれに、汎化、例外処理の導入、形容詞の直近に対する優先度を提案する。

【1:汎化】

一般名詞意味属性体系の属性は約2,700種類あるため、名詞句の頻度統計がスパースになる。そこで、以下の通りに意味属性を汎化する。意味属性ツリー上のα、βの共通の親ノード(深さd)を基準にして、その子ノード(深さd＋n)のうち、αとβを配下に属する意味属性ノードの共起頻度の和をそれぞれα、βの頻度とする。なお、d+n≧3とする。

【2:例外処理の導入】

「大きい木の そば」、「遠い 過去 の話」など、場所や時を表す抽象名詞が含まれる場合、基本的手法(3.1節)では正しい意味的結合力が得られない場合が多い。そこで以下の例外処理を導入する。

- 「形容詞＋名詞A＋名詞B」において、AかBの意味属性が「場所」、「非層日」ならば、係り先をAにする。

Dependency Analysis for Japanese Noun Phrases with Adjectives
‡Akio MORIUTI:World Fashion S・E
†Shinji NAKAI,Satoru IKEHARA,Mariko OHNISHI:Tottori University

【3:形容詞の直近に対する優先度】
人間が文を記述するとき、形容詞の係り先は直後に記述することが多い。そのため＜直近優先の規則＞が成り立つと思われる。そこで、直近の優先度を重み t(≧1)で表し、βの頻度がαの頻度の t 倍以上あるときのみ、係り先をBにするとし、正解率が最大になるような t の値を定める。

4 実験結果

4.1 手順

本研究の実験の手順を以下に示す。

(1) 小説100冊の中から収集した5,773個の名詞句(形容詞＋名詞)の名詞に意味属性を付与し、各形容詞と単語意味属性の組に対して頻度統計をとる。
(2) 同じ小説より新たに収集した735個の名詞句(形容詞＋名詞＋の＋名詞)の名詞に意味属性を付与し、形容詞の係り先を決定する。

標本の信頼度を考慮して、本手法を適用することで求められたαとβの頻度の和が10以上の場合(350個)と20以上の場合(202個)の形容詞＋「の型名詞句」のみを使用する。

4.2 人手おける形容詞の係り先決定基準

名詞には本来、複数の意味属性を持つものが多い。一般名詞意味属性体系には最大12段あり、深いノードほど精密な意味を表している。そこで本研究では、複数ある意味属性の中で、その名詞が使われている名詞句の内容に最も的確で、かつ最もノードが深い属性を一意に決定し、それを付与した。

正解は三人が個別に決定した結果(A:Aに係る,B:Bに係る,C:どちらに係っても良い)をもとに、以下の二つの基準で決定した。

基準1：三人一致した係り先を正解とし、不一致の場合はCとする。
基準2：多数決により係り先を決定する。三人とも不一致の場合はCとする。

4.3 実験結果

人手による係り先判定結果を表1に示す。

表1: 人手による係り先判定結果 (単位：%)

頻度の和	標本数	基準1			基準2		
		A	B	C	A	B	C
10以上	350個	49.1	10.3	40.6	69.4	17.1	13.4
20以上	202個	50.0	10.0	40.0	69.8	15.8	14.4

A：Aに係る、B：Bに係る、C：どちらに係っても良い

例外処理が有無の実験結果の比較を表2に示す。

表2: 係り先判定の正解率 (単位：%)

頻度の和	例外処理有り		例外処理無し	
	基準1	基準2	基準1	基準2
10以上	88.9	81.4	93.4	88.0
20以上	88.6	80.7	91.6	85.1

また、例外処理後さらに頻度を加重した場合の正解率の変化のグラフを図1に示す。

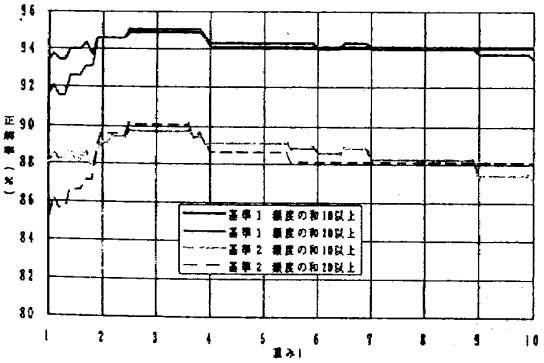

図1 ：t の変化による正解率の推移

5 考察

表1より、形容詞が名詞Aに係っている名詞句が多い。これより、もし全てがAに係っている(all A方式)としても、ある程度の正解率が得られると考えられる。(実験結果：表3参照)。表2より、例外処理を導入すれば、いずれの場合も正解率は実験結果の方が高いことが分かる。また図1より、優先度 t が2.5～3.5のとき正解率が最大になると分かる。

表3: all A 方式 (単位：%)

頻度の和	基準1	基準2
10以上	89.7	82.9
20以上	90.0	84.2

6 結論

「の型名詞句」に対して、形容詞と名詞との意味的情報を含む結合力に着目した形容詞の係り先の決定法を提案した。例外処理を導入することで正解率は向上する。さらに優先度 t を2.5～3.5とするとき正解率が最大になり、95.0％の精度で形容詞の係り先が決定できることが分かった。これにより、本手法の有効性が示された。

今後は特に以下についての検討が必要である。

(1) 一つの意味属性では名詞の意味的情報を正しく表現できなかった名詞句
(2) 形容詞と名詞との関係だけでは正しい係り先が得られなかった名詞句

参考文献

[1] 池原、宮崎、白井、横尾、中岩、小倉、大山、林 (1997): 日本語語彙体系、岩波書店