名詞間係り受け解析に必要な単語意味属性の組の最適化

4R－1

中井 慎司† 伊藤 真樹‡ 池原 悟† 白井 諭*
†鳥取大学大学院工学研究科 ‡アネルバ（株）
*NTTコミュニケーション科学研究所

1 はじめに

最近、結合価パターン辞書（日本語語彙体系 [1]）の開発により、日本語動詞の意味は、かなり精度良く解析できるようになってきた。しかし、名詞の意味解析では、依然としてよい方法が知られていない。本論文では、「の」型名詞句を取り上げ、格助詞「の」を介して結合された名詞間の意味的関係を記述するのに必要最低限の単語意味属性の組を明らかにする。

具体的には、任意の2つの名詞の意味的係り受け関係が名詞の意味属性間のマトリクス（正方行列）で表現できると仮定する。そして、名詞の意味属性として結合価パターン辞書の単語意味属性（約2,700種類）を使用し、意味的係り受け解析の精度を低下させないで、使用する意味属性をどこまで絞り込めるかを明らかにする。

2 名詞の意味的関係

2.1 名詞間の意味的係り受け関係の表現

今回使用する意味属性体系 [1] は動詞の意味の解析用に開発されたが、名詞の意味を詳細に分類（約2,700種類）しているため名詞の意味解析にも適用できる可能性がある。そこで、本論文では、名詞の中でも典型的な名詞句である「の」型名詞句を取り上げ、名詞間の意味的係り受け規則をこの意味属性体系を使用したマトリクス（係り名詞を行、受け名詞を列）で表現する。

2.2 使用する意味属性数の最適化

マトリクス上、名詞間の意味的係り受けの有無は、二者択一的に○×で表現される。経験的に、その値を人間の思考実験で正確に決定することはほとんど不可能*である。そこで、本論文ではコーパスを使用した標本統計によって決定することを考える。しかし、結合価パターンの記述に使用された単語意味属性、約2,700種類のすべてのペアについて、信頼できる統計量を得ることは現実的に不可能である。また、名詞間の意味的関係はより少数の意味属性で表現できる可能性がある。そこで、標本統計によって得られた係り受け関係データを用い、意味属性の数を絞り込むことにより、名詞句解析に必要な意味属性の数とその組を決定する。

2.3 係り名詞と受け名詞の意味属性について

名詞間の係り受け関係を表現するのに適した行列は、必ずしも正方行列になるとは限らない。形態素解析の文法接続表の例†から類推すると、係り名詞を表現する単語意味属性の数に比べて、受け側の名詞を表現する意味属性の数はより少なくてすむ可能性がある。しかし、ここでは、まず両者の数及び種類を等しいと仮定して、最小限の意味属性の組を求めることにする。従って、以下で求められる意味属性の組は、最終的に目標とするマトリクスの長辺を構成する意味属性の組と考えることができる。

3 最適な意味属性の組の推定法

以下に示す手順によって、最適な意味属性の組の数を推定する。

1. 初期マトリクスの設定

「名詞Aの名詞B」型名詞句の名詞A及び名詞Bをそれぞれ意味属性に置き換える。次に、初期値を任意に選びn×nの共起マトリクスを作成し頻度統計をとる。

2. 共起マトリクスの拡大

ある任意の意味属性に着目し、その意味属性をその一段配下の意味属性に置き換え、n'×n'の共起マトリクスを作成する。以後、この新しい共起マトリクスを拡大共起マトリクスと呼ぶ。

Optimization of Semantic Attributes for Japanese Noun Phrase Structure Analysis
†Shinji NAKAI,Satoru IKEHARA: Tottori University
‡Masaki ITOH: Anelva Corporation
*Satoshi SHIRAI:NTT Communication Science Laboratories

*実際に、アナリストによって単語意味属性とそれに属する名詞の例を取り出して、「の」型名詞句の表現が可能かどうかを判断する方法を試みたが、思考を重ねるたびに判断基準が不明確になるため、人為的な方法は諦めざるをえなかった。

†形態素解析（ALT-JAWS,NTT）で使用される文法的属性間の接続表の例では、前方単語と後方単語の接続関係が、約500×約100のマトリクスで表現されている。

3. 共起頻度を用いた「名詞A＋の＋名詞B＋の＋名詞C」の係り受け解析

先ほど作成した拡大共起マトリクスを用い、(AのB)と(AのC)の頻度より、「名詞A＋の＋名詞B＋の＋名詞C」の係り受け解析を行う。係り先は、以下の式(1)に従い決定した。

$$f(A\ の\ B) * w \geq f(A\ の\ C)\ ならば\ ((A\ の\ B)\ の\ C)$$
$$f(A\ の\ B) * w < f(A\ の\ C)\ ならば\ (A\ の\ (B\ の\ C)) \quad (1)$$

f(A の B) : (A の B) の頻度

f(A の C) : (A の C) の頻度 , w : 重み

この時、名詞Aは名詞Cより名詞Bに係りやすいことを考慮にいれ適当な重み w（2 ～ 2.5）を (A の B) の頻度に掛けた。そして、重みを変化させながら係り受け解析を行い、その時の最高の正解率をこの拡大共起マトリクスを使用した係り受け解析の正解率とした。

4. 手順2と手順3を繰り返す

拡大共起マトリクスを使用したときの係り受け正解率とその前の共起マトリクスを使用したときの係り受け正解率とを比較し、正解率が上がっていればその拡大共起マトリクスを採用し、手順2へ行く。正解率が下がっているならば元の共起マトリクスに戻し、手順2へ行く。手順2と3を繰り返し、どの意味属性を拡大しても係り受け正解率が上がらなくなった場合終了とする。

4 実験

4.1 実験データ

使用する名詞句データは新潮文庫の小説100冊より抽出した。「名詞A＋の＋名詞B」型名詞句は抽出された約17万件のうち、約4,200件を使用し、それを共起分析用データとした。また「名詞A＋の＋名詞B＋の＋名詞C」型名詞句は抽出された約12,000件のうち、約1,300件を使用し、それを係り受け解析用データとした。

4.2 実験結果

図1に意味属性体系の深さ（その深さにある意味属性を使用する）と係り受け正解率の関係を示す。この図より、正解率のピークは深さ3（意味属性数:21）から深さ5（意味属性数:256）の間にあることが分かる。

そこで、深さ3にある意味属性21種類を初期値とし、3章に示した手順に従って21×21の共起マトリクスを拡大していく。図2に意味属性の数と係り受け解析正解率の関係を示す。この図より意味属性の組の数を81個にした時、係り受け正解率が最大になることが分かる。よって名詞句解析に使用する「AのB」の共起マトリクスは81×81のマトリクスが最適であることが分かる。

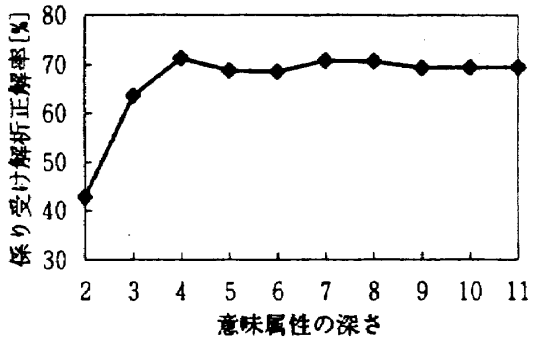

図1: 意味属性体系の深さと係り受け正解率

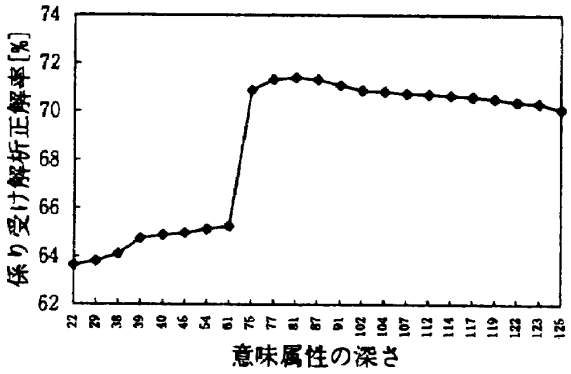

図2: 意味属性の組の数と係り受け正解率

なお、81×81（=6,561）の共起マトリクスのうち実際に共起関係にある「AのB」の数は全体の約1割にあたる665組（頻度1以上）であった。

5 おわりに

本稿では、「の」型名詞句の名詞間の意味的関係を記述するために必要最小限の意味属性の組の数を推定する手法を提案し、名詞間の共起関係を共起マトリクスにまとめた。実験により、81×81の共起マトリクスが名詞句解析に最適という結果が得られた。

今後は、受け側の名詞の意味属性を圧縮しn（名詞A）×m（名詞B,n>m）の共起マトリクスを作成し、再度比較してみる必要がある。また、作成した共起マトリクスに意味関係を付与する予定である。

参考文献

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