

構成要素の統語・意味的制約と用例を利用した 1 E-3 「の/と型」名詞句構造解析法

金内 哲也

宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

日本語文には、いくつかの名詞を助詞「の」や「と」で結合した名詞句が数多く出現する。これらの名詞句の構造解析は、その意味の多様性から日本語処理における困難で重要な課題となっている。

本稿では、名詞の統語・意味的制約と用例を用いた「 N_1 の N_2 の N_3 」型名詞句の構造解析法を拡張し、 n 個の名詞が「の」または「と」で結合された名詞句の構造解析法を提案し、その有効性について論じる。

2 「の型」名詞句の構造解析

3個の名詞が「の」で結ばれた「 N_1 の N_2 の N_3 」型名詞句の構造解析については、文献[1]において正解率90.5%を達成している。[1]で用いられている接続強度と用例による評価点に、名詞間距離による重み付けをすることにより、4名詞以上の「の型」名詞句についても適用可能とした。

2.1 接続強度

名詞句を構成する名詞にも、その種類によって「の」の左側に来やすい名詞と右側に来やすい名詞がある。名詞を具体名詞、抽象名詞など16種類に分類し、それぞれに左側接続強度と右側接続強度を設定した(表1)。

接続強度を用いて「 N_1 の N_2 」を評価する場合、

Japanese Noun Phrase Structure Analysis using
Syntax, Semantic Constraint and Noun Phrase
Corpus

Tetsuya Kaneuchi, Masahiro Miyazaki

Niigata University

表 1: 接続強度の例

	左側接続強度	右側接続強度
具体名詞	10	2
関係名詞	12	6
サ変名詞	12	4

(N_1 の右側接続強度 + N_2 の左側接続強度) が評価点の基準となる。

2.2 名詞句用例データベース

名詞句の構造の評価用に、「 N_1 の N_2 」型の名詞句約75,000例からなる用例データベースを用意した。これらは主に EDR 共起辞書から獲得したものである。解析対象となる名詞句を構成する名詞を2つずつ取り出して、それぞれデータベース中の用例との類似度評価をする。

表2における一致条件は

X 字面; 意味カテゴリ; 品詞の一致

Y 意味カテゴリ; 品詞の一致

Z 品詞のみの一致

の3種類とする。

表 2: 一致条件と類似度

前方 \ 後方	X	Y	Z
X	6.0	5.6	0.2
Y	5.8	×	×
Z	0.4	×	×

2.3 4名詞以上の名詞句への対応

前述の接続強度と名詞句用例データベースによる評価結果を4名詞以上からなる「の型」名詞句にも適用するため、名詞句内における名詞間距離(隣接した名詞間で1)による重み付けを行なった。試行錯誤の結果、名詞間距離による重みの値は0.8、[1]より用例評価の重みは0.2となり、名詞 N_A と名詞 N_B の接続強度による評価点を C_{AB} 、用例による評価点を E_{AB} 、名詞間距離を D_{AB} とした場合の評価式を $0.8^{D_{AB}-1}(C_{AB}+0.2E_{AB})$ と設定した。簡単な評価実験によれば、この評価式により4名詞からなる「の型」名詞句解析において82.5%の正解率を得た。

3 「と」を含む名詞句への拡張

名詞句を構成する助詞として、「の」の他に「と」を含む名詞句の係り受け構造を解析する場合、名詞句内で並列構造を作る「と」に着目して名詞句を分割すると、ある名詞は「の」によって、「と」で分割された各部の名詞の中の1個ずつに係ることができると考えられる。従って、それぞれの部分で最も評価点が高い名詞に係り先候補とし、「と」を挟んで係るのは評価点があらかじめ設定した閾値を越えた場合のみとした。この閾値の具体的な値は現在検討中である。

例えば図1において、「歌手ら」は「質」「高さ」のうち評価点が高い方(この場合は「質」)にそのまま係ることができるが、「と」を挟む「顔触れ」「豊富さ」には、閾値を越えた場合のみ係ることができる。

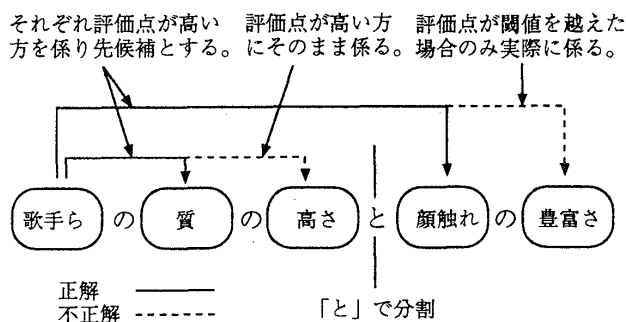


図1: 「歌手ら」の係り先判定処理

4 「と」により並列する名詞の抽出

「 N_1 と N_2 の N_3 」という名詞句において「の」による係り受け関係を解析した段階で、「と」の直前にある N_1 の係り先が無ければ、 N_1 と並列するのは N_2 以降の「の型」名詞句の主名詞である N_3 となり、 N_1 が N_3 に係るようであれば、 N_1 とその係り先の直前の名詞である N_2 が並列すると考えることができる。

例えば「本と新聞の社説」という名詞句では「本」の係り先が無いので、「本」と「社説」が並列する名詞となり、「テレビとビデオの電源」では「テレビ」が「電源」に係るため、「テレビ」と「ビデオ」が並列となる。

5 おわりに

本稿では、接続強度と用例を用いた「 N_1 の N_2 の N_3 型」名詞句の構造解析法を拡張し、 n 個の名詞が「の」または「と」で結ばれた「の/と型」名詞句の構造解析法を提案した。

「の/と型」名詞句を解析する際の、「と」を越えて係るための閾値の具体的な値は現在検討中である。また、今回の方式では「束ね」の名詞[2]や名詞の飽和度[3]といった要素を考慮していないため、これらを組み込んだ新たな評価法の導入を考える必要がある。

謝辞

「EDR 日本語共起辞書」の使用を許可された日本電子化辞書研究所、単語意味属性体系データの使用を許可されたNTTコミュニケーション科学研究所の関係各位に深謝いたします。

参考文献

- [1] 江尻, 宮崎: 名詞間の接続強度と「の」型名詞句の用例を利用した日本語名詞句構造解析法, 情報処理学会第56回全国大会講演論文集(2), 1Q-2, pp193-194(1998.3)
- [2] 田村, 田中: 意味解析に基づく並列名詞句の構造解析, 自然言語処理 59-2(1987)
- [3] 西山佑司: NP_1 の NP_2 と" NP_1 of NP_2 ", 日本語学, vol.12(1993)