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

1Q－2

江尻 秀彰　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「の」や「と」で結合した名詞句が数多く出現し、その意味も多様である。そのため、日本語の処理において、このような名詞句を解析することは、重要で困難な課題となっている。

本稿では、n個の名詞を「の」で結合した名詞句に限定し、名詞の接続強度の利用と、「名詞＋の＋名詞」型名詞句の用例を利用した名詞句構造解析を提案し、その定量的評価とその有効性について論じる。

2 接続強度について

格助詞「の」で連結される名詞において、名詞の性質によって「の」の前方に来やすい名詞、「の」の後方に来やすい名詞がある。この性質を考慮して、文献[1]を基に名詞を16種に分類したものに接続強度づけした。接続強度は、数字が大きいほど接続強度は大きく、接続しやすいということを表している。

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

Japanese Noun Phrase Stracture Analysis Using Connection Cost And Noun Phrase Corpus

Hideaki Ejiri, Masahiro Miyazaki
Niigata University

3 用例の利用

名詞句の用例としては、「名詞＋の＋名詞」型のものを用意する。これらは、主にEDR共起辞書から獲得したものであり、約75,000例を用意した。名詞句の情報としては、各名詞の字面、品詞、そして意味カテゴリ[2]を記述した。名詞句用例データベースから用例を検索する際に、その類似度を設定しておくことが大切である。そこで用例数が多いことを考慮して、一致条件を厳しくし、どちらか一方の名詞の字面が一致することを最低条件として、その類似度を設定した。ここでは、主名詞となりやすい後方の名詞と一致する場合に高い点を与えた。一致条件としては、

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. 一致条件とその類似度

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

4 評価値の計算

入力名詞句の各名詞間には、接続強度と用例データベースとの類似度の評価点を与える。そして、名詞句の構造は評価点の合計とその評価点の重みによって決定される。以下に、3名詞の名詞句入力「A の B の C」の場合の評価値の計算式を示す。

$\alpha \leqq \beta \rightarrow$ （A の B）の C

$\alpha > \beta \rightarrow$ A の（B の C）

用例による評価

$\alpha_s = S_{A,B}$

$\beta_s = S_{A,C}$

接続強度による評価

$\alpha_c = C_{A,B}$

$\beta_c = (C_{A,C} * Wc + C_{B,C})/2$

両方用いた評価

$\alpha = \alpha_s * Ws + \alpha_c$

$\beta = \beta_s * Ws + \beta_c$

$S_{A,B}$ ・・・名詞 A と名詞 B間の類似度

$C_{A,B}$ ・・・名詞 A と名詞 B間の接続強度

Ws ・・・類似度の重み

Wc ・・・間接接続の重み

解析の例

例. 最期（時詞）の 場面（関係名詞）
の 台本（抽象名詞）

用例との類似度は、「最期の場面」が「字面＋字面」で完全にマッチし6.0,「最期の台本」が「字面＋品詞」(「最期の手紙」とマッチ）で0.2となる。接続強度の評価点は、表3より、「最期の場面」1+12=13,「最期の台本」1+12=13,「場面の台本」6+12=18である。ただし、重みはそれぞれ、Ws=0.2、Wc=0.7である。

$\alpha_s = 6.0, \beta_s = 0.2$

$\alpha_c = 13, \beta_c = (13 * Wc + 18)/2 = 13.55$

$\alpha = \alpha_s * Ws + \alpha_c = 14.2$

$\beta = \beta_s * Ws + \beta_c = 13.59$

よって、

$\alpha > \beta$ → （最期の場面）の台本

となる。

表3. 接続強度

	左側接続強度	右側接続強度
時詞	6	1
関係名詞	12	6
抽象名詞	12	4

5 評価

3名詞間の名詞句1210例に対して解析を行なった。各評価による正解数と正解率の結果は次の通りである。

・接続強度のみによる解析 … 正解1060（87.6%）
・用例のみによる解析 … 正解994 （82.1%）
・両方用いた解析 … 正解1095（90.5%）

この評価により、接続強度と用例との類似度の評価の両方を用いることの有効性が示された。なお重みは試行実験をくり返した結果、Ws=0.2、Wc=0.7とした。

6 まとめ

本稿では、日本語名詞句の「の」で結ばれた名詞句に限定し、その句構造解析を行なうにあたり、接続強度と、「名詞＋の＋名詞」型名詞句の用例を利用する方法を提案し、その定量的な評価を行なうことにより本手法の有効性を確認した。

本評価では、対象を格助詞の「の」で結ばれた3名詞間の名詞句のみに限定したが、今後は助詞の「と」を含むもの、4個以上の名詞を含む名詞句にも対応できるよう解析法を拡張する必要がある。

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] 宍倉 、宮崎：構成要素の統語・意味的制約を利用した日本語名詞句解析、信学技報、NLC94-49, PP.41-48 (1995)

[2] 宮崎 、池原 、横尾 、白井：日英機械翻訳のための意味属性体系、信学技報、NLC97-12, PP.29-36 (1997)