名詞間接続強度と用例の係り受け情報を用いた日本語「の」型名詞句解析

武本 裕　　宮崎 正弘

新潟大学大学院自然科学研究科

1　はじめに

単純名詞Nを格助詞「の」で結合した「NのNのN の…」の形式を持つ名詞句は「の」型名詞句と呼ばれるが、この構造はN≧３の場合係り受けに曖昧性を生じる。２名詞で構成される「の」型名詞句に関しては、名詞の種類により「の」の左側に出現しやすいものと右側に出現しやすいものがある。これらの可能性をコーパス等を用いて数値化したものを接続強度とする。

本手法では、この接続強度に加えて例外的な構造をもつものへの対処として用例を用いた解析を統合することで解析精度の向上を図った。品詞レベルの分類に基づいた接続強度に加えて、語彙レベルの傾向を捉える。

本稿では上記手法による３名詞で構成された「の」型名詞句の構造解析法を提案し、その有効性を示した。

2　「の」型名詞句

複数の名詞を助詞「の」で結合した名詞句は一般に「の」型名詞句と呼ばれている。３名詞以上からなる名詞句の場合には係り受け構造に曖昧性が生じる。本稿では、３名詞からなる「N_Aの N_Bの N_C」の形式の名詞句を対象とした構造解析を行なう。この名詞句の構造には、N_A が N_B に係る (B 係り) 場合と N_A が N_C に係る場合の２通りの可能性がある。図１にそれぞれの構造を示す。

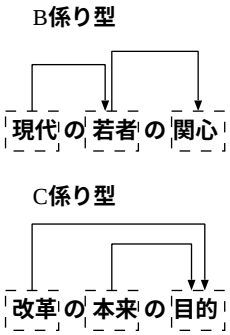

図 1: B 係り型と C 係り型

3　接続強度を用いた構造解析

3.1　接続強度

名詞句を構成する名詞には、「の」の右側に来やすいものと左側に来やすいものとがある。

それぞれの傾向の強さを右側接続強度、左側接続強度として設定する。

接続強度の値は、文献 [5] の値を用いた。接続強度を表 1 に示す。ここでは、名詞を 13 種類に分類して接続強度を設定している。品詞種別をもとにいくつかのグループに分類している。これらの値は、コーパスから得られた統計情報をもとにして、さらに実験により調整された値である。

表 1: 接続強度

品詞	右側接続強度	左側接続強度
普通名詞	8	11
サ変名詞	6	12
動作名詞	5	10
状態名詞	10	9
形容詞転生名詞	4	6
形容動詞転生名詞	3	5
連体詞性名詞	6	2
数詞	8	6
時詞	7	5
副詞型名詞	6	4
固有名詞	5	3
形式名詞	7	10
代名詞	12	10

3.2 接続強度を用いた判定法

接続強度を用いた判定法では以下の式を用いる。この判定法は文献 [5] に基づく。

B 係りの評価点：$P_{AB} = (R_A + L_B) * 1.00$

C 係りの評価点：$P_{AC} = (R_A + L_C) * 0.85$

$P_{AB} \geq P_{AC}$ のとき B 係り型、$P_{AB} < P_{AC}$ のとき C 係り型と判定する。

この接続強度を用いた判定法では、B 係りの正解率は高いが C 係りの正解率はやや低い。

4 用例による補強

接続強度の値は品詞レベルでの分類に基づいて設定されている。品詞レベルでは捉えられない、個々の語彙に依存した傾向を補完することによってさらなる解析正解率の向上を狙う。具体的には以下の手法を考案し判定プログラムを試作した。

4.1 ２名詞からなる「の」型名詞句の用例とのマッチングによる判定

２名詞からなる「の」型名詞句の用例を収集しデータベース化する。名詞句「N_Aの N_Bの N_C」の係り受け判定では、「N_Aの N_B」と「N_Aの N_C」をそれぞれデータベースから検索する。このとき、「字面→名詞意味カテゴリ→品詞」と段階的にマッチングを行なう。段階が進むにつれて「類似度」の値は低くなる。ここでは、表 2 の「類似度」を用いる。

表 2: 類似度

左＼右	名詞	主名詞	意味	品詞
名詞	10	8	5	1.5
主名詞	8.5	7	3.5	0.5
意味	5.5	4	-	-
品詞	2	1	-	-

「N_Aの N_B」と「N_Aの N_C」でより類似度が高いものがより適切であると考え、「N_Aの N_B」が高ければ B 係り、さもなくば C 係りと判定する (図 2)。

用例とのマッチングの類似度が低い場合、信頼性が低いことがあるため、単独での利用は困難で組み合わせて補助的に利用することにする。

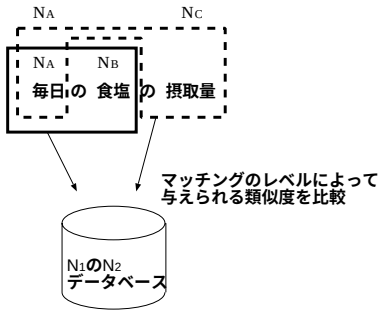

図 2: ２名詞「の」型名詞句用例による判定

4.2 N_B に着目した C 係り判定

３名詞からなる「A の B の C」のタイプの名詞句に関して、名詞 B に着目する。特に接続強度による判定で正解率の低い C 係りの精度向上を狙う。B 係りの場合には「N_A→ N_B」と「N_B→ N_C」、C 係りの場合には「N_A→ N_C」と「N_B→ N_C」の係り受けが存在する。C 係りでは「N_A→ N_B」の係り受けが存在しないことから、「N_B が受け側となる可能性が低ければ C 係りとなる」と仮説を立て、この仮説をもとに判定を行なう (図 3)。これには、２名詞「の」型名詞句に出現する名詞に関して、コーパスをもとに左側出現頻度、右側出現頻度を集計する。左側出現頻度が高いものほど係り側になりやすく、受け側にはなりにくい。

この方法では極端に出現頻度が少ない名詞の場合には困難であると予想できる。

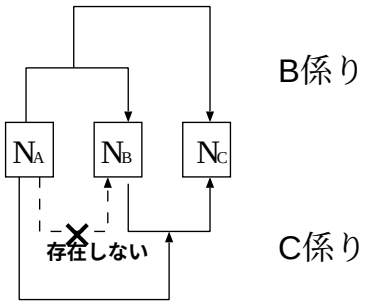

図 3: N_B に着目した C 係り判定の概念図

4.3 N_B に着目した B 係り判定

前節では、C 係りの判定を行なったが、同様に B 係りの判定を行なう手法を提案する。C 係りの場合ほど直観的ではないが、「名詞 N_B が受け側となる可能性が高ければ B 係りとなる」という仮説をもとに判定を行なう。これは、人間が言語をその並び順にしたがっ

て理解しようとしたときに先に「N_A の N_B」の結合が強固ならばそれをそのまま確定しようとするのではないかという考えから来ている。

これは、B 係りの正解率に反映する。

5 評価

5.1 用例とのマッチングによる判定

３名詞からなる「の」型名詞句 1206 例に対して実験を行なった。

２名詞からなる「の」型名詞句約 36 万を用例として用いた。「N_A の N_B」「N_A の N_C」に関して用例とマッチングを行なう。「N_A の N_B」の方が高い類似度でマッチすれば B 係り、さもなくば C 係りとする。判定結果を表 3 に示す。正解率を表 4 に示す。

表 3: 用例とのマッチングによる判定

正解	不正解
1016	190

表 4: 用例マッチングによる判定／正解率

正解率 (%)	B 係り 正解率 (%)	C 係り 正解率 (%)
84.3	89.1	69.3

5.2 N_B による C 係り判定

３名詞からなる「の」型名詞句 1206 例に対して実験を行なった。

２名詞からなる「の」型名詞句約 36 万に含まれる名詞に関して、その左側出現頻度、右側出現頻度を集計する。それより、各名詞の左側出現率を算出する。

ここでは、C 係りのみの判定を行なう。N_B の左側出現率 90%以上であれば、C 係りであるとして判定を行なった。N_B が用例に含まれる名詞と一致しない場合には判定不能とした。判定結果を表 5 に示す。正解率は、判定不能を除いて算出した。

表 5: N_B による C 係り判定

正解	不正解	判定不能	正解率 (%)
156	6	1044	96.3

効果は限定的であるが高い正解率が得られた。

5.3 N_B による B 係り判定

３名詞からなる「の」型名詞句 1206 例に対して実験を行なった。

前節と同様に各名詞の右側出現率を算出する。

N_B の右側出現率 55%以上であれば、B 係りであるとして判定を行なった。N_B が用例に含まれる名詞と一致しない場合には判定不能とした。

判定結果を表 6 に示す。正解率は、判定不能を除いて算出した。

表 6: N_B による B 係り判定

正解	不正解	判定不能	正解率 (%)
359	15	832	96.0

接続強度による判定ではもともと B 係りの正解率は高いため、この正解率では接続強度との組み合わせは困難。

5.4 N_B による判定と用例マッチング判定の融合

３名詞からなる「の」型名詞句 1206 例に対して実験を行なった。

ここでは、N_B による C 係り判定→ N_B による B 係り判定→用例マッチング判定の順に、先に確定したものを優先して判定を行なった。

判定結果を表 7 に示す。

表 7: N_B・C → N_B・B →用例

正解率 (%)	B 係り 正解率 (%)	C 係り 正解率 (%)
87.4	91.8	74.0

B 係り、C 係りともに正解率の向上が見られる。

5.5 N_B による判定と接続強度による判定の融合

３名詞からなる「の」型名詞句 1206 例に対して実験を行なった。

接続強度による判定はもともと B 係り正解率が高いため、N_B による C 係り判定→接続強度の順で先に確定したものを優先して判定する。

判定結果を表 8 に示す。

表 8: N_B・C →接続強度

正解率 (%)	B 係り 正解率 (%)	C 係り 正解率 (%)
91.6	97.4	75.3

また、比較のため接続強度のみの判定結果を表 9 に示す。C 係り正解率の向上が見られる。

表 9: ＜比較＞接続強度のみ

正解率 (%)	B 係り 正解率 (%)	C 係り 正解率 (%)
91.0	97.5	72.8

6 おわりに

「の」型名詞句解析に関して品詞レベルの分類に基づく、接続強度による判定法に加えて用例を用いて語彙依存の傾向を補完するための手法を提案し、その有効性を示した。

主に、「N_A の N_B の N_C」中の N_B に着目して C 係りの正解率を向上させることで全体の正解率を高めた。接続強度による手法、または用例マッチングによる手法ともに C 係り判定のさらなる正解率向上が依然として課題である。

参考文献

[1] 尾嶋基、宮崎正弘:高精度と頑健性を目指した日本語形態素解析とその定量評価
情報処理学会第 56 回全国大会講演論文集 (2)1Q-1(1998).

[2] 江尻秀彰: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(1998).

[3] 金内哲也、宮崎正弘:規則／用例融合型の日本語名詞句構造解析法
言語処理学会第 6 回年次大会発表論文集
pp.403-406(2000).

[4] 池原悟、中井慎司、村上仁一:多義解消のための構造規則の生成方法と日本語名詞句への適用
自然言語処理学会論文誌,Vol.8,No.1
pp.143-174(2001).

[5] 益田裕也、宮崎正弘:名詞間の接続強度を用いた「の」型名詞句構造解析
言語処理学会第 9 回年次大会発表論文集
pp.238-241(2003).