名詞間の接続強度を用いた「の」型名詞句構造解析

益田裕也　　宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

日本語文に数多く出現する名詞句の中でも特に出現頻度の高いものに、複数の名詞を助詞「の」や「と」で結合した名詞句がある。名詞が助詞「の」によって結合された「NのNのN…」のような型の名詞句は「の」型名詞句と呼ばれ、非常に単純かつ基本的なものでありながら、「の」によって作られる名詞同士の関係が多様であるために、名詞句の構造すなわち係り受け構造に曖昧性が生じる。この「の」型名詞句の係り受け構造は、その意味を理解したり、他の言語へ翻訳する際などに不可欠な情報であるが、現状では、この係り受けを高い精度で解析することは困難な課題となっている。

本稿では、3名詞からなる「の」型名詞句を対象に、名詞間の統語的・意味的制約を反映した、接続強度を用いた構造解析法を提案し、その有効性を示す。

2 「の」型名詞句とその構造

日本語名詞句の中でも、最も基本的で数多く出現するものの一つとして、名詞を助詞「の」によって結合したものがある。このような名詞句は一般的に「の」型名詞句と呼ばれている。「の」型名詞句はその単純さゆえに様々な意味が考えられるうえに、3つ以上の名詞が「の」で結合された場合は、係り受け構造にも曖昧性が生じる。

係り受け構造に曖昧性がある「の」型名詞句のうち代表的なものは3名詞からなる「N_AのN_BのN_C」という形式のものである。この型の名詞句の構造としては、N_AがN_Bに係る場合(B係り型)とN_Bを飛び越えてN_Cに係る場合(C係り型)が考えられる。それぞれの例を図1に示す。

図1において、名詞句「中国の銀行の営業」は、「中国」が「銀行」に係り、「銀行」が「営業」に係るという直線的な係り受け構造がある。これに対し、名詞句「自国の直接の利害」の「自国」の場合は、その直後にある名詞「直接」に係ると考えるよりは、「利害」に係って「自国の利害」という構造があり、同時に「直接の利害」という構造も存在すると考えた方が自然である。

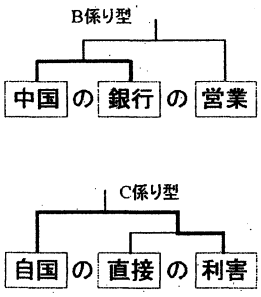

図 1: 係り受けの例

3 接続強度を用いた構造解析法

名詞は、その働きの違いにより、普通名詞やサ変名詞などといった多くの種類に分けることができる。文献[1]の形態素辞書にある情報をもとに、名詞を表1のような13種類に分類した。

Japanese Noun Phrase Structural Analysis
using Connection Cost between Nouns
Yuuya Masuda , Masahiro Miyazaki
Niigata University

表 1: 名詞の分類

品詞	品詞コード	例
普通名詞	1100	ビル、窓枠
サ変名詞	1210	不足、発掘
動作名詞	1220,1230	表れ、鋼製
状態名詞	1320,1330	危険、不均衡
形容詞転生名詞	1510	早さ、速さ
形容動詞転生名詞	1520	重要さ、大切さ
連体詞性名詞	1600	本当、不慮
数詞	17**	一連、一人
時詞	1810	戦後、古代
副詞型名詞	1820,1830	全て、一部
固有名詞	19**	日本、太郎
形式名詞	1a**	こと、もの
代名詞	1b**	私、彼女

また、名詞にはその種類により、助詞「の」の左側に出現しやすい（係り側になりやすい）ものと、右側に出現しやすい（受け側になりやすい）ものがある。例えば「架空」のような連体詞性名詞は、「架空の～」のように、他の語を修飾することが多く、被修飾語となることは少ない。従って連体詞性名詞は係り側になりやすく、受け側になりにくいと考えられる。本稿の接続強度とは、これらを数値化したものであり、右側接続強度が大きいものは「の」を挟んで右側に名詞を取りやすいということであり、左側接続強度が大きいものは「の」を挟んで左側に名詞を取りやすいということである。

接続強度は文献[2,3]で提案されているものを基本とし、そこにコーパスから得た名詞句データの統計情報による検討（左右の出現頻度等）、試行錯誤を加えて設定した。表1の品詞分類に従い、各品詞が助詞「の」の左右どちら側に現れたかを集計した結果を表2に示す。

表2において、右側出現頻度よりも左側出現頻度が高い品詞は「の」による係り側になることが多いことになるため、その差に応じて右側接続強度を高く設定した方が良いと考えられる。逆に、左側出現頻度よりも右側出現頻度が高い品詞は「の」による受け側になることが多いことになり、左側接続強度を高く設定した方が良いと考えられる。文献[2,3]において設定された接続強度に、前述したような統計データを元にした修正を加え、新たに設定した接続強度を表3に示す。

表 2: 各品詞が「の」の左右に現れる頻度

品詞	左側出現頻度	右側出現頻度
普通名詞	289820	296367
サ変名詞	64695	115761
動作名詞	11600	23520
状態名詞	12513	7493
形容詞転生名詞	448	2726
形容動詞転生名詞	56	498
連体詞性名詞	31290	8508
数詞	9488	5409
時詞	37867	18590
副詞型名詞	15402	7723
固有名詞	20913	4860
形式名詞	12150	19441
代名詞	7749	3095

表 3: 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

4 係り受け構造の判定

4.1 ルールの適用

基本的に係り受けの判定は、後述する評価点の大小比較によって行われるが、名詞の特徴によって係り受けが高い確率で確定できる場合がある。本論文では、固有名詞の特徴に着目して、以下のルールを設定した。

- 固有名詞同士の係り受けを確定

「の」型名詞句において、固有名詞同士の結び付きは非常に強いので、この係り受けを確定する。このルールは、名詞句の中に2つの固有名詞が存在したときのみ適用され、1つ、または3つの固有名詞が含まれる場合は適用されない。
このルールの適用例を以下に示す。

例 名詞句
「日本（固有名詞）の東京（固有名詞）の気温（普通名詞）」

この場合、「日本」と「東京」の係り受けを確定し、図2のような木構造が構成され、B係り型と判定する。

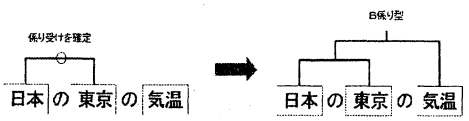

図 2: 係り受け確定の例

4.2 評価点の大小比較

2.2節の図1を参照すればわかるとおり、名詞 N_A が、名詞 N_B に係ればB係り型、名詞 N_C に係ればC係り型となる。そこで、図3のように評価点を与え、その大小比較によって係り受けを判定する。

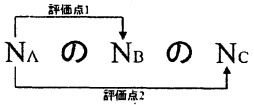

図 3: 評価点の算出

「N_Xの N_Y」という名詞句の接続強度を用いた評価点 P_{XY} を、次の式で計算する。

- $P_{XY} = (R_X + L_Y) * W_{XY}$

ここで、R_X は N_X の右側接続強度、L_Y は N_Y の左側接続強度、W_{XY} は名詞間の距離による重み定数を表している。名詞同士は距離が遠くなるほど係り受けしにくくなることを考慮し、
評価点1の算出:$W_{XY} = 1.00$
評価点2の算出:$W_{XY} = 0.85$
と設定した。評価点2の重み定数については、定量評価を行い最も正解率が高くなった0.85を採用した（図4参照）。

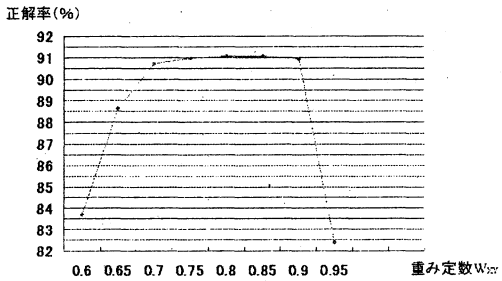

図 4: 重み定数のグラフ

名詞句「N_Aの N_Bの N_C」におけるB係り、C係り評価点の算出式はそれぞれ以下のようになる。

- B係り評価点:$P_{AB} = (R_A + L_B) * 1.00 \cdots (1)$
- C係り評価点:$P_{AC} = (R_A + L_C) * 0.85 \cdots (2)$

この2つの評価点のうち、高いものを N_A の係り先とし、名詞句の構成要素である名詞間の係り受け構造を判定する。すなわち、$P_{AB} \geqq P_{AC}$ の場合、その名詞句は「B係り型」、$P_{AB} < P_{AC}$ の場合、その名詞句は「C係り型」とする。

4.3 解析例

実際の解析例を以下に示す。

- 「異種（連体詞性名詞）の金属（普通名詞）の接合（サ変名詞）」

図3より、まず最初に評価するのは「異種の金属」という名詞句である。表1により求まる接続強度を式

(1) に代入すると、評価点は

$P_{AB} = (6 + 11) * 1.00 = 17$

となる。次に評価する、「異種の接合」という名詞句についても同様に式 (2) に代入して、

$P_{BC} = (6 + 12) * 0.85 = 15.3$

となる。この結果、

$P_{AB} > P_{BC}$

が成立し、この名詞句はB係り型であることがわかる (図5参照)。

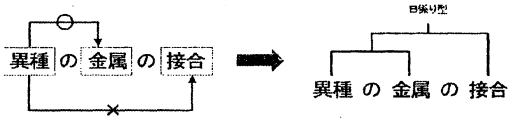

図 5: 名詞句「異種の金属の接合」の解析

5 評価

5.1 評価実験の結果

「N_A の N_B の N_C」型名詞句1206例を用いて評価実験を行なった。実験の結果、解析正解率は約91%であった (表4参照)。

表 4: 評価結果

	総合	ルール無し	B係り	C係り
評価例数	1206	1206	845	319
正解数	1098	1096	823	233
正解率	91.04%	90.88%	97.40%	73.04%

「の」型名詞句の構造解析正解率は、文献 [3,4] 共に約85%であり、本手法が有効であることが示された。また、文献 [3] において、B係りの正解率が約95%、C係りの成功率が約67%であったことを考えると、名詞句解析の難題である、C係り型の判定にも大きな効果があったといえる。なお、B係りとC係りの評価例数を足しても全体の1206例にはならないが、これはB係りともC係りともとれる名詞句が42例存在したため、それを差し引いた数だからである。

5.2 不正解データ

評価実験における不正解例を以下に示す。

B係り型をC係り型にしたもの
- 電柱の真下の路面
- 事故の直前の記憶

C係り型をB係り型にしたもの
- 学校の歴史の教科書
- 当時の片道の運賃
- 買収後の会社の資産
- 本当の恐竜の専門家

6 おわりに

本稿では、接続強度を用いた「の」型名詞句構造解析法を提案し、その有効性を示した。今後、接続強度の見直しや、用例解析との組み合わせも検討し、解析正解率の向上をはかる必要がある。また、B係り型の正解率に比べ、C係り型の正解率が著しく低いことから、名詞句構造解析の精度向上には、C係り型の判定を正確にすることが重要である。

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)