

名詞間の接続強度を用いた「の」型名詞句構造解析

3M-03

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…」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 3名詞からなる「の」型名詞句とその構造

3名詞からなる「の」型名詞句を「 N_A の N_B の N_C 」とする。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、その後後に

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。

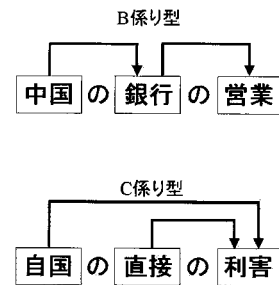


図 1: 係り受けの例

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[1]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、その名詞自体の出現頻度等)、試行錯誤を加えて設定した(表1参照)。

表 1: 接続強度

品詞	右側接続強度	左側接続強度
普通名詞	5	11
サ変名詞	4	12
動作名詞	10	10
状態名詞	10	10
形容詞転生名詞	6	4
形容動詞転生名詞	6	4
連体詞性名詞	6	2
数詞	6	8
時詞	1	6
副詞型名詞	6	4
固有名詞	4	2
形式名詞	10	10
代名詞	4	12

4 係り受け構造の判定

まず、名詞句「 N_A の N_B の N_C 」に対して、図2のようにそれぞれ評価点を算出する。続いて、表1の接続強度を用いた「 N_X の N_Y 」名詞句の評価点 P_{XY} を、式(1)で計算する。

$$P_{XY} = R_X * L_Y * W_{XY} \quad (1)$$

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である。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。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。

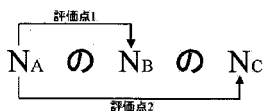


図 2: 評価点の算出

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2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

$$P_{AB} = 6 * 11 * 1.00 = 66$$

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、

$$P_{BC} = 6 * 12 * 0.85 = 61.2$$

となる。この結果、

$$P_{AB} > P_{BC}$$

が成立するため、B係り型であることがわかる。

5 評価

文献[1]の接続強度を用いた解析法は、式(2)のように評価点を算出しており、その解析正解率は約84%（固有名詞を含むもの限定では約73%）であった。

$$P_{XY} = (R_X + L_Y) * W_{XY} \quad (2)$$

「 N_A の N_B の N_C 」型名詞句1214例を用いて評価実験を行った結果、本手法での解析正解率は約88%であった。また、固有名詞を含む名詞句の正解率は約87%と、大きく向上している。この結果から、本手法の有効性が確認された。

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。

参考文献

- [1] 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）
- [2] 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）