規則／用例融合型の日本語名詞句構造解析法

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、多種多様な名詞句が出現し、その統語構造や意味もまた多岐にわたる。それらの名詞句の中でも代表的な形式の一つとして、いくつかの名詞を助詞「の」や「と」で結合したものがあげられるが、その形が単純であるがゆえに、正しい統語構造や意味を判別することが困難となっている。また、これらの名詞句を取り扱った研究の多くは構成する名詞の数や「の」の使われ方を限定している場合が多く、その限定に収まらない名詞句の解析は不可能であった。

そこで、本稿では、名詞が助詞「の」で結合されたものに焦点をあて、構成する名詞の数によらない解析法を提案する。その解析法としては、構成要素である名詞の統語的制約を用いたもの、および、大量の文の解析結果から獲得した「NのN」型名詞句の用例を用いたものを考え、その2つを効果的に統合する方法について提案し、その有効性について論じる。

2 「の」型名詞句の構造解析

2.1 「の」型名詞句とその構造

多種多様な種類が存在する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、複数の名詞が助詞「の」によって連結された名詞句がある。このような名詞句は一般に「の」型名詞句と呼ばれている。「の」型名詞句は、その単純さゆえにさまざまな意味が考えられるうえに、3名詞以上から構成される場合は係り受け構造にも曖昧性がある。

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。

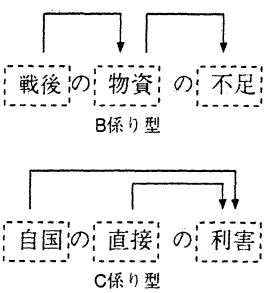

図1: B係り型とC係り型の例

図1において、名詞句「戦後の物資の不足」は、「戦後」が「物資」に係り、そして「物資」が「不足」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である「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。

「N_AのN_BのN_C」型名詞句のうち約7割はB係り型となっており、その中で、いかにC係り型を判定するかが重要となっている。そして、同様の曖昧さは4個以上の名詞からなる名詞句の場合にも当然存在する。

本稿では、特に出現頻度の高い3名詞からなる名詞句に重点を置きつつ、4名詞以上からなる、より一般的な名詞句にも応用できる構造解析法を提案する。

Japanese Noun Phrase Structure Analysis using Syntax Constraint and Noun Phrase Corpus
Tetsuya Kaneuchi, Masahiro Miyazaki
Niigata University

2.2 統語的制約を利用した解析

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、それぞれに右側接続強度と左側接続強度を設定した(表1)。この分類と接続強度の設定にあたっては、経験的に得られたものである[1]および、日本経済新聞1994年の記事データから獲得した「NのN」型名詞句データにおいて、各品詞が「の」の左右に出現する回数の統計結果を利用した。[1]の接続強度を統計結果と照らし合わせ、不整合がある部分では統計結果を反映した数値を設定してある。

表1: 接続強度

品詞	右側	左側
具体名詞	2	10
抽象名詞	6	12
関係名詞	8	12
サ変名詞	4	12
動作名詞	10	10
状態名詞1	10	10
状態名詞2	10	12
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

例として「N_Aの N_Bの N_Cの N_Dの....」という名詞句を解析する場合、まず、N_1の係り先を判定する。この際、N_xとN_yとの間の評価点をC_{xy}とした場合の計算式を次のように設定する。

$$C_{xy} = (R_x + L_y) * W_c^{d-1} \qquad (1)$$

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、dは隣り合った名詞間を1とする距離、W_cは距離による重み定数とする。本稿では、試行錯誤の結果W_cの最適値として

$$W_c = 0.8$$

としている。N_Aと、続く各名詞との評価値を計算し、もっとも高い評価値を持つ組合わせが係り受け関係にあると判定する。

解析例 名詞句:「戦後(副詞型名詞)の物資(具体名詞)の不足(サ変名詞)」の解析

先頭の名詞(この例の場合は「戦後」)と後に続く各名詞との間の評価点を計算し、どの名詞に係るのかを判定する。したがって、最初に評価するのは「戦後の物資」という構造である。

表1より、「戦後」の品詞である副詞型名詞の右側接続強度は6、「物資」の品詞である具体名詞の左側接続強度は10、さらに、2つの名詞の間には他の名詞を挟まないため、距離による重みは考慮しなくて良い。したがって、「戦後の物資」の評価点C_{AB}は式1より、

$$\begin{aligned} C_{AB} &= (R_A + L_B) * W_c^0 \\ &= (10 + 6) * 1 \\ &= 16 \end{aligned}$$

となる。

続いて、「戦後の不足」という名詞句を評価する。「不足」はサ変名詞であり、右側接続強度は4である。さらに、「戦後の不足」は「戦後」が途中にある名詞「物資」を1つ越えてより遠くへ係る構造であるため、距離の離れた分の重みとして、W_c^1すなわち0.8をかける。したがって、「戦後の不足」の評価点C_{AC}は

$$\begin{aligned} C_{AC} &= (R_A + L_C) * W_c^1 \\ &= (6 + 4) * 0.8 \\ &= 8 \end{aligned}$$

となる。この結果、

$$C_{AB} \geq C_{AC}$$

が成立するため、「戦後→物資」という係りが確定する。

次の名詞「物資」に関しては、係り先が「不足」以外に考えられないため、「物資→不足」という係りが確定し、

戦後→物資
物資→不足

という係り受け構造が確定される(図2)。

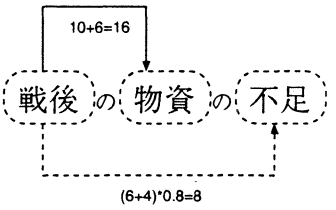

図 2: 「戦後の物資の不足」の解析

2.3 用例を利用した解析

名詞句の構造の評価用に、「N_Aの N_B」型の名詞句約75万個34万種類からなる用例データベースを用意した。これらは、接続強度の設定でも用いた、日本経済新聞の記事データから獲得した名詞句データを元に、各名詞の主名詞を取り出し、抽象度を上げたものから構成されている。つまり、例えば名詞句データ「民主党の役員集会」という名詞句は「党の集会」という用例としてデータベースに登録されている。データベースには、表記の他に品詞コード、意味カテゴリコード (第1候補のみ) も登録した。

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。ここでの一致条件は

X 表記; 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種類とし、それぞれに表2にある評価点を与える。

表 2: 一致条件とその類似度

左側 \ 右側	X	Y	Z
X	10	3	0.1
Y	5	×	×
Z	0.2	×	×

統語的制約を用いた解析の場合と同様、この評価点と名詞間距離による重みをもとに構造を解析する。

評価に用例を用いることを除いては、接続強度を用いた解析と大きな違いはない。「N_Aの N_Bの N_Cの N_Dの....」という名詞句を解析する場合の N_x と N_y の評価点 E_{xy} の計算式を、次のように設定する。

$$E_{xy} = M(N_x, N_y) * W_e^{d-1} \qquad (2)$$

ここで、$M(N_x, N_y)$は、「N_xの N_y」という名詞句と用例の類似度により表2から与えられる評価点である。dは隣り合った名詞間を1とした距離であり、W_eは名詞間の距離による重み定数とする。本稿では、試行錯誤の結果、接続強度と同様に

$$W_e = 0.8$$

と設定した。

解析処理では、先頭の名詞から順に、後に続く名詞との間の評価点を式2より計算し、係りの交叉が起こらない範囲で、最も高い評価点が得られた名詞に係る。評価点が等しい場合は、距離が近い方の名詞に係る。

解析例 名詞句「今回 (副詞型名詞) の米国 (固有名詞) の要求 (サ変名詞)」の解析

「今回の米国」に関して用例データベースを検索すると

類似度	回数
完全マッチ	0
左側カテゴリ＋右側表記	0
左側表記＋右側カテゴリ	9
左側品詞＋右側表記	0
左側表記＋右側品詞	5

となる。表2と式2より評価点 E_{AB} を計算すると、

$$\begin{aligned} E_{AB} &= M(N_A, N_B) * W_e^{d-1} \\ &= 3 * 9 + 0.1 * 5 \\ &= 27.5 \end{aligned}$$

となる。

つづいて、「今回の要求」について用例データベースを検索すると、

類似度	回数
完全マッチ	3
左側カテゴリ＋右側表記	0
左側表記＋右側カテゴリ	0
左側品詞＋右側表記	30
左側表記＋右側品詞	2664

となる。表2と式2より評価点E_{AC}を計算すると、

$$\begin{aligned} E_{AC} &= M(N_A, N_C) * W_e^{d-1} \\ &= (10*3+0.2*30+0.1*2664)*0.8^1 \\ &= 302.4 \end{aligned}$$

となり、

$$E_{AB} \leq E_{AC}$$

が成立する。したがって、

今回→要求
米国→要求

という構造であると判定される (図3)。

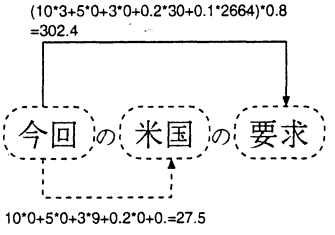

図 3: 「今回の米国の要求」の解析

3 統語的制約と用例を用いた解析

「N_Aの N_Bの N_C」型の名詞句解析において、統語的制約による解析と、用例による解析を比較したところ、前者はB係りの正解率が非常に高くC係りの正解率が低くなっており、後者はB係り、C係りともに同程度の正解率となっている。「N_Aの N_Bの N_C」型の名詞句では約7割がB係りであることから、両方の解析結果がC係りであると判定した場合にのみC係りであると判定し、その他はB係りであると判定することにより、B係りの正解率をさらに高め、C係りの正解率の低下もさほど起こらないと考えられる。

現在、大規模データによる定量的評価は完了していないが、簡単な評価実験では約83%の正解率が得られている。

4 助詞「と」が入った名詞句の解析

「の」型名詞句の構造解析法を応用し、「と」を挟んだ「の」による係り受け関係を判定することにより、助詞「と」が入った「の/と」型名詞句の並列構造も抽出可能であると考えられる。この際は、「の」によって「と」を越えた係り受けを認定する条件、具体的には評価点の閾値を適切に設定する必要がある。解析例を図4に示す。

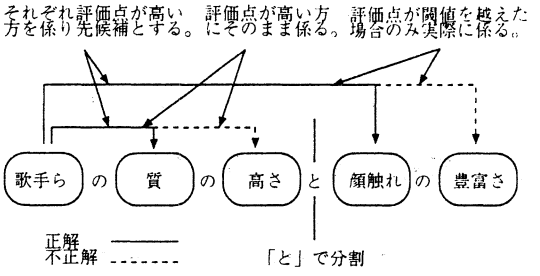

図 4: 「歌手ら」の係り先判定処理

5 おわりに

本稿では、統語的制約と用例を用いた「の型」名詞句の構造解析法を提案した。例としては「N_Aの N_Bの N_C」型の名詞句のみを取り上げたが、解析方法には構成する名詞数による部分はなく、4名詞以上の場合でもそのまま適用できる。今後は「と」が入った名詞の場合も含め、定量的評価を行ない、2つの解析法をより高度に統合する方法を検討する必要がある。

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、日本経済新聞記事データ (1994年版) を提供くださった日本経済新聞社出版局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(1998.3)

[2] 尾嶋, 宮崎：日本語形態素解析システムにおける部分的再試行機構の導入とその効果, 情報処理学会第58回全国大会講演論文集 (2),1E-4(1999.3)