「の」と「や」で結合された並列名詞句の構造解析法

2N－1

松本　知博　　石川　勉
拓殖大学工学部情報工学科

1．はじめに

日本語処理において並列構造文の係り受け判定は重要な問題である。これについては、これまで動詞を含む複雑な構造については研究されてきているが[1]、名詞だけが「や」などの格助詞で結合された並列構造文についてはあまり研究されてきていない。

本報告では、名詞または複合名詞が「の」と「や」で結ばれた「AのBやCのD」のような並列名詞句の係り受け解析法について提案する。この種の文の解析では、基本的には名詞の意味の違いを理解することが必要となる。本手法では言葉の意味の類似性判定に、これまで開発してきた概念ベース[2]やシソーラス情報[3]を用いる。さらに名詞の種類や共起に関する情報も利用する。

2．係り受け関係の種類

「AのBやCのD」の係り受けのパターンは以下の4種類に別けられる。矢印は係り受けを表す。

・パターン1：A の B や C の D
例　雇用情勢の悪化や消費の低迷

・パターン2：A の B や C の D
例　女性の地位向上や社会進出の問題

・パターン3：A の B や C の D
例　樹木の保護や樹勢回復の業務経験

・パターン4：A の B や C の D
例　日本の経済黒字や内需の弱さ

パターン1は、名詞句“AのB”，“CのD”が“や”によって結合されたものである。パターン2では、“A”が“BやC”に係り、さらにそれが“D”に係る。パターン3は“A”が“B”に係り、“AのB”と“C”が“D”に係るものである。さらに、パターン4では“A”が“B”と“CのD”に係る。

3．識別規則

3.1　識別に用いる情報

以上の4種類の係り受けに対して次の情報を利用して判定規則を作成する。

i）概念ベース：単語間の意味の類似性判別が可能な24万語からなる大規模知識ベース[2]。名詞間の意味の類似性判定に利用。

ii）シソーラス情報：2715のカテゴリからなり12段の木構造を持つシソーラス[3]。名詞間の意味の類似性判定とカテゴリの同一性判定に利用。

iii）品詞情報：形態素解析から得られる、名詞の分類情報（普通名詞、サ変名詞、固有名詞など）。名詞句の分割と名詞の連接の判定に利用。

iv）共起辞書情報：新聞記事をもとに作成した単語間の共起頻度情報。名詞の連接の判定に利用。

3.2　主な識別規則

新聞記事情報（94年度毎日新聞）から対象となる名詞句を抽出し、その特徴について検討し、約30の識別規則を作成した。主な識別規則を以下に示す。

1）パターン1

パターン1の名詞句では、“虎の骨や鹿の角”のように“A”と“C”、“B”と“D”がそれぞれ対応する場合が多く見られた。この場合は、それらがそれぞれ意味的に類似すると考えられる。従って、以下の規則を設定した。

・規則1：$(R[A,C]>\alpha_1) \cap (R[B,D]>\alpha_1)$

ここで、R[X,Y]は“X”と“Y”の類似度、α_1は類似判定のための閾値である。

また、パターン1では並列名詞句全体が“や”を中心に分かれていることから、サ変名詞という意味を打ち切る品詞に注目し、以下の規則を設定した。

・規則2：（B＝D＝“サ変名詞”）∩（C≠“サ変名詞”）

この例としては“イベントの実施やガイドマップの作成”等があり、“実施”や“作成”を動詞と見たとき、それぞれ“イベント”と“ガイドマップ”がそれらの目的格となっている。

2）パターン2

基本的にパターン2は、“投手の肩や肘の保護”のように“B”と“C”が対をなす意味を持つ構造であるため、以下の規則を設定した。

・規則3：　$R[B,C]>\alpha_2$

Coordinate・Structural Analysis of Japanese Noun Phrases Connected by Case Particles
Chihiro Matsumoto, Tsutomu Ishikawa
Department of Computer Science, Takushoku University

3）パターン3

パターン3は、“家畜の豚や人間の血液”のように最後の名詞“D”がそれ以前の名詞句“AのB”と“C”を受けている構造である。従って、単語の共起関係およびカテゴリ（大分類）の同一性を利用し、以下の規則を設定した。

・規則4：（M [A,B]>β_1）∩（M [B,D]>β_1）∩
（M [C,D]>β_1）∩（Ki [B]＝Ki [C]）

ここで、M [X,Y] はX，Yの共起頻度であり、Ki[X] は、Xの上位概念でシソーラスにおけるi段目のカテゴリである。

4）パターン2or3

“D”がサ変名詞で“B”が普通名詞のときは、“世界の河川や湖沼環境の調査”のように、“D”を動詞とみたとき“B”と“C”がその目的格になっていることになっていることが多い。この場合は、少なくてもパターン1、パターン4ではないと考えられる。従って、以下の規則を設定した。

・規則5：（B＝“普通名詞”）∩（D＝“サ変名詞”）

5）パターン4

パターン4は、“周囲のゴミや犬のふん”のように基本的には“A”が“B”と“D”に係るもので、“A”と“C”は全くカテゴリが異なる場合が多い。従って、以下の規則を設定した。

・規則6：（Ki[A]≠Ki[C]）∩（R [B,D]>α_3）
∩（M [C,D] >β_2）

4．実験結果

前述した新聞記事から抽出した並列名詞句を対象に実験を行った。具体的には、各パターンごとに20個の並列名詞句を評価データとして用意したが、パターン3，4は抽出されたデータが少なかったため人手で作成した並列名詞句を加えている。

4．1 各規則の正解率と適用率

図1、図2にそれぞれ規則1、規則3に対するの実験結果を示す。ここで適用率とは、この規則が全80の評価データに対して適用された割合である。当然のことながら、基準値αを大きくすると正解率は向上するが、適用率は下がることが分かる。

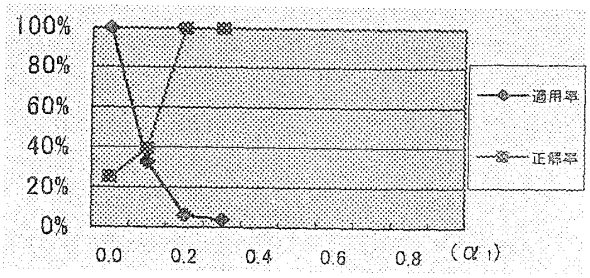

図1　規則1（パターン1）の実験結果

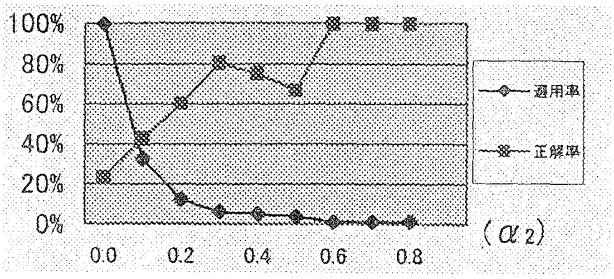

図2　規則3（パターン2）の実験結果

また、規則2および規則4～6に対する実験結果を表1に示す。ここでは、各パラメータは規則4と規則6におけるiは3、β_1、β_2は30、また規則6におけるα_3は0.3としている。

表1　規則2および規則4～6に対する実験結果

	正解	不正解	未判定	適用率	正解率	備考
規則2	6	3	71	11%	67%	パターン2
規則4	4	11	65	19%	27%	パターン3
規則5	18	4	58	28%	82%	パターン2or3
規則6	4	4	62	10%	50%	パターン4

規則4は正解率が極めて低いが、誤りはパターン4をパターン3と判定し生じていた。

4．2　総合的な正解率

上記の6個の規則の他に9個の規則を加え、全80の評価データを4パターンに分類する実験を行った。なお、規則1のα_1は0.2、規則3のα_2は0.3としている。

各規則は正解率が高い順に適用することとした。また、規則5を初期の段階で適用しパターン1、4とパターン2、3を分類した後、他の規則を適用した。その結果、正解：20、不正解：17、未判定：43が得られた。未判定の43データのうち11データは、パターン2または3に絞られている。従って、これらのデータの半分を正解とし、他の未判定データの4分の1が正解と扱うと、全体としての正解率は約42%になると期待される。

5．まとめ

「AのBやCのD」のような並列名詞句の係り受け解析法について提案した。この名詞句は4つのパターンに分類されるが、“AのB”，“CのD”が“や”によって結合されるパターン以外については、確実性の高い規則が少いため、全体の正解率としては42%とあまり高くなかった。これらのパターンに対する有効な規則について今後検討する必要がある。

参考文献

[1] 江尻　：名詞の統語・意味的制約と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54回全国大会p61
[2] 帆苅　譲、笠原　要、石川　勉：言葉の意味に関する階層型大規模概念ベースの構築、信学技報AI98－65（1999－01）P25
[3] 池原　悟，宮崎　正弘，白井　論、横尾　昭男、その他　日本語語彙体系　全5巻　岩波書店