

「の」型名詞句における名詞間の意味的係り受け規則の自動生成

中井 慎司[†] 池原 悟[†] 白井 諭[‡]

[†]鳥取大学大学院 工学研究科

[‡]NTTコミュニケーション科学研究所

1 はじめに

自然言語処理の最大の問題は、表現の構造と意味に関する解釈の曖昧性である。いままで、多くの研究が行われてきたが、文法的情報に頼った従来の方法では、これらの問題と解決するのは困難であった。

日本語名詞句の解析では、従来、コーパスの基づく方法として、単語の共起情報を用いて係り先を決定する方法[1]、意味的クラスの共起情報を用いて係り先が決定される確率を求める方法[2]、また、名詞句解析では、大量の対訳用例の中から意味的に類似した表現を発見し、翻訳結果を得る方法[3]などが提案されてきた。しかし、通常、コーパスから得られる標本はスパースであり、適切な用例がない場合には、結果は保証されない。このように、従来の方法では、必要十分な標本データを収集すること、また、計算量が膨大となることなどが問題であった。

この問題を解決するため、本稿では、名詞句の持つ意味的な構造に着目して、名詞と名詞の間の係り受け規則を名詞句標本データから自動生成する。具体的には、最も基本的な日本語名詞句「AのBのC」を対象に、単語意味属性体系において定義された名詞の意味的用法の上位下位関係を用いて、3つの名詞 A、B、C間の意味的係り受け規則を生成する。この方法は、以下の手順から構成される。まず、日本文中で使用された「AのBのC」型の名詞句を収集し、正しい係り受け関係を付与するとともに、日本語意味辞書[4]を参照して、それぞれの名詞句に含まれる名詞をその名詞の持つ意味属性に置き替える。次に、これを学習用データとし、意味属性間の上下関係に着目して、その中から名詞間の係り受け規則を最も汎用的な規則から順に機械的に生成する。そして得られた解析規則を別の名詞句標本に適用して解析精度を評価する。

2 表現の意味構造

2.1 対象とする名詞句とその意味構造

(1) 対象とする名詞句とその曖昧性

以下では、2つの助詞「の」と3つの名詞 A、B、

Cから構成された「AのBのC」の形の名詞句を考える。ただし、記号 A、B、C は名詞の出現順序をも表すものとする。この名詞句は、係り受け関係に曖昧さのある名詞句の中で最も基本的なものである。以下、この型の名詞句を単に「の型名詞句」という。

日本語では、一般に、表現要素間に後方修飾の原則があることに注意すると、「の型名詞句」では、名詞 B の係り先は名詞 C に特定されるため、先頭の名詞 A について、以下の2通りの係り受け解釈が存在することになる。

1) A→B (&B→C) の場合

例) 「私の母の名前」、
「浴室の脱衣場の壁」

2) A→C (&B→C) の場合

例) 「私の昔の友達」、
「東京の数学の教師」

以下では、簡単のため、1)を「b 係り」、2)を「c 係り」と呼ぶ。

(2) 名詞句の構造と意味の問題

「AのBのC」を「AのB」と「BのC」に分けて考える方法は意味論的にも問題がある。要素合成法の考え方に従えば、表現の意味はそれを構成する部品の意味に還元されるが、言語表現では、この原理が成り立たない場合が多く、表現の構造と意味の関係を考えなければならない場合が多い。例えば、下記の名詞句では、2つの名詞句に分離することは適切でなく、3つの名詞の組とその出現順序に依存して意味が決定される。

例) 「盗人のなれの果て」
「私の気のせい」

このような場合は、表現を分解せず、ひとまとまりのものとして扱うことが必要である。従って、「の型名詞句」の名詞間の係り受け関係を決定する場合も、表現をその構成要素に分解してよい場合と分解できない場合に分けて考えることが重要である。

3 意味的係り受け規則の自動生成

3. 1 係り受け規則の形式とタイプ

(1) 係り受け規則の形式

名詞句「AのBのC」の意味構造を(X,Y,Z)で表す。ただし、X,Y,Zは、それぞれ、名詞A,B,Cの意味属性番号とする。次に、この意味構造の名詞句に対する係り受け規則を図1のように(X,Y,Z:D)で表す。ただし、Dは係り受けのタイプで、D=bは前方係り受け(b-dependency)、D=cは後方係り受け(c-dependency)を表すものとする。

(X,Y,Z:D): 係り受け規則の一般形

X: 名詞Aの意味属性番号

Y: 名詞Bの意味属性番号

Z: 名詞Cの意味属性番号

D: 係り受けの種類 (bまたはc)

図1 係り受け規則の記述形式

(2) 係り受け規則のタイプ

この名詞句の解析では、「3つの名詞の意味属性の組が決まれば係り受け関係が一意に決定される」と仮定したが、必ずしも3つ名詞の意味属性のすべてが決まらなくても、係り受け関係が決まる場合がある。そこで、この点に着目して、係り受け規則を以下の通り4種類に分けて考える。

1) 単一の名詞に着目した規則 (1次元規則)

構成要素の名詞の一つの意味属性とその名詞が何番目の名詞として使用されたかが分かれば、残りの2つの名詞の意味属性とは無関係に、係り受け関係が決定できるもの。対象とする名詞句では3つの名詞が使用されているため、その位置に対応して、解析規則は、3グループに分けられる。

2) 2つの名詞に着目した規則 (2次元規則)

2つの名詞の意味属性とそれらの出現位置が与えられれば、残りの名詞の意味属性とは無関係に係り受け構造が決定できるもの。意味属性が無関係となる名詞の位置に応じて、解析規則は3種類に分類される。

3) 3つの名詞すべてに着目した規則 (3次元規則)

3つの名詞すべての意味属性とその出現位置の関係に係り受け構造が決定できるもの。

4) それ以外の規則 (例外規則)

単語の意味属性と出現位置では係り受け構造が決

定できないもの。単語の字面を用いて規則を記述する必要があるため、ここでは例外的ルールと考える。

以上の規則の抽出では、規則が抽出されるごとに、その規則作成に使用した名詞句の標本は使用済みとし、残された名詞句標本から次の係り受け規則を抽出するものとする。従って、実際の名詞句解析では、1次元規則から順に適用し、適用できる規則があった段階で係り受け解析は終了する。このように係り受け規則を構成することによって、解析精度を落とさないうえ、規則数全体の増大を防ぐことができると期待される。

3. 2 係り受け規則の抽出と汎化

(1) 1次元規則の抽出と汎化

名詞A,B,Cのそれぞれについて単語意味属性体系の木を用意し、図2に示すように、各木の各ノードにリスト(m_i, n_i)を対応させる。ただし、 m_i は、i番目の意味属性を持つ名詞が使用された名詞句の内、b-dependencyの数、 n_i はi番目の意味属性を持つ名詞が使用された名詞句の内、b-dependencyの数を表すもので、いずれも、3.1節で作成した標本データを集計することによって得られる値である。

ここで、1次元規則は、3つの名詞のうちのどれか一つの名詞の意味属性のみに依存して係り受け関係が決定できるものであるから、十分大量の標本データから m_i, n_i の値が求められているとすれば、 m_i または n_i のどちらか一方が0となるとときに1次元規則が存在する。逆に、両者の値がいずれもゼロでないときは、その意味属性iを用いた1次元規則は存在しない。

すなわち、名詞Aの意味属性体系で、 $m_i \neq 0$ かつ $n_i = 0$ なる意味属性iからは、係り受け規則($i, 0, 0: b$)が得られ(図3参照)、 $m_j = 0$ かつ $n_j \neq 0$ なる意味属性jからは、係り受け規則($j, 0, 0: c$)が得られる。 $m_k \neq 0$ かつ $n_k \neq 0$ なる意味属性kでは、1次元規則は存在しないから、次項以下の方法で、2次元規則、3次元規則の有無を調べる。なお、 $m_i = 0$ かつ $n_i = 0$ なる1では、改めて規則を作成する必要はない。

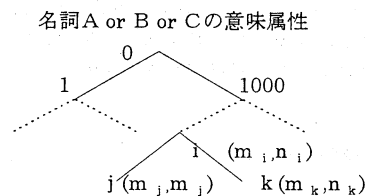


図2 1次元規則の抽出

表2 自動生成された係り受け規則の数と適用回数および正解率

1次元ルール				2次元ルール				3次元ルール			
タイプ	ルール数	適用頻度	正解率	タイプ	ルール数	適用頻度	正解率		ルール数	適用頻度	正解率
A	89.6	780	91.8	A&B	1103.7	5437	91.3	計	1417.9	1053	69.7
B	89.1	699	89.3	B&C	1065.0	6600	92.1	例外ルール			
C	31.1	606	94.4	C&A	1029.7	5859	91.8		ルール数	適用頻度	正解率
計	209.8	1932	91.6	計	3198.4	6615	91.0	計	400	0	0

2) 2次元規則について

約3200の係り受け規則が得られ、全体の66%をカバーしている。

3) 係り受け規則数は約1400と多いが、カバーしている率は1次元規則と同じ程度である。

4) 例外規則は全体の4%と少なく、3次元規則まででほぼカバーできることが分かる。

4. 3 係り受け解析

生成された規則を10,000件の実験データに適用したところ、全体の96.0%に適用され、適合率は88.8%であった。適用されなかった名詞句はわずか4.0%であった。表2に適用された頻度、それぞれの規則における正解率を示す。この表より以下のことがいえる。

- 1) 1次元規則、2次元規則ともに正解率が高い。またこの2つの規則で全体の85.5%をカバーすることができる。
- 2) 一方、3次元規則の正解率は悪い。
- 3) 3つの規則を合わせると全体の96.0%をカバーすることができる。

4. 4 従来の方法との比較、検討

表3に従来の方法である3次元規則のみを使った解析と本手法との解析精度の比較を示す。この表からわかるように、3次元規則のみの解析の場合は約20%が解析できなかったのに対し、本手法で解析できないものはわずか4%であった。その結果、最終的な正解率も10%強向上した。このことより汎用性のある規則(1次元規則、2次元規則、3次元規則の順)から順に適用することの有効性が示された。

表3 従来の手法との正解率の比較

	3次元規則のみ	本手法
判定率	80.5	96.0
適合率	92.0	88.8
正解率	74.2	85.2

* 正解率 = 判定率 × 適合率

5 おわりに

本稿では、係り受け解析に曖昧性のあるもっとも典型的な名詞句である「AのBのC」を取り上げ、名詞の意味属性を使用した意味的係り受け規則の自動生成法を提案した。この方法によって大量の解析済みの名詞句標本から名詞間の意味的構造と係り受け解析の関係を自動的に学習し、汎用性のもっとも高い規則から順に3種類の規則を生成した。

本方式を日本語名詞句10,000件に適用した実験では、3種類の規則として、それぞれ210件、3198件、1418件が得られた。また、これらの規則を別の名詞句標本10,000件の解析に適用した結果によれば、係り受け解析の正解率は88.8%であった。これより、提案した方法は精度の高い解析規則が得られることが分かった。

また本方式は別の名詞句の係り受け解析「AのBとC」、「形容詞+AのB」などにも適用可能であると思われる。

参考文献

- [1] 佐々木、坂本：文章一括処理による係り受け関係の解析、言語処理学会第1回年次大会、pp.101-104、(1995)
- [2] 小林、徳永、田中：名詞間の意味的共起情報を用いた複合名詞の解析、自然言語処理、Vol.3、No.1、pp.29-43(1996)
- [3] Sumita, E. and Iida, H: Example-based Transfer of Japanese Adonominal Particles into English、IEICE Trans. Inf. & Syst., pp585-594(1992)
- [4] 池原、宮崎、白井、横尾、中岩、小倉、大山、林：日本語語彙体系、岩波書店、(1997)