自然言語処理 128－7
（1998. 11. 5）

「の」型名詞句における品詞情報と意味情報を併用した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

{nakai,ikehara}@ike.tottori-u.ac.jp

本稿では、構造に多義を持つ日本語名詞句の中でも、基本的でかつ高頻度で現れる「AのBのC」を対象に、品詞情報および意味属性を用いた係り受け規則の自動生成法を提案する。生成される係り受け規則は品詞情報のみを使用した規則、品詞情報と意味属性を混合した規則、意味属性のみを使用した規則の3種類である。ここで、品詞情報のみを使用した規則生成には決定木を用い、品詞情報と意味属性を混合した規則は、同格を抽出することを目的とし、人手で係り受け規則を作成した。また、意味属性のみを使用した規則は汎化を用いた学習によって生成した。これら3種類の係り受け規則を名詞句実験データに組み合わせて適用した結果、係り受け解析正解率が86.8%となり、従来提案した意味属性のみの係り受け規則を適用した場合より、向上することが確認された。

Automatic Generation of Dependency Rules Using Part of Speech and Semantic Attributes for Japanese Noun Phrases with Particles "no"

Shinji NAKAI† Satoru IKEHARA† Satoshi SHIRAI‡

†Tottori University
‡NTT Communication Science Laboratories

{nakai,ikehara}@ike.tottori-u.ac.jp

「A no B no C」 is the most typical noun phrase and frequentcy in Japanese. This paper proposed a method automatically to generate dependency rules. In this method, we obtain 3 types of rules using part of speech,semantic attributes and both part of speech and semantic attributes. In the experiments, these rules were applied to the dependency analysis. As result, we obtain 86.8% accuracy rate. From these result, it was found that the proposed method obtains more accurate than ordinary method.

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、確率的、文法的情報に頼った従来の方法では、これらの問題を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。また必要十分な標本データを収集すること、計算量が膨大となることなどが問題であった。

これに対し、最近大量の標本データから汎化を用いた学習を行いルールを生成する手法が盛んに行われており、中井ら [4] は「A の B の C」の係り受け解析に適用している。この手法は、「A の B の C」を対象に、名詞句の持つ構造に着目して、3つの名詞 A、B、C 間の意味的係り受け規則 (3つの名詞の組の規則、2つの名詞の組の規則、1つのみの名詞の規則) を名詞句標本データから自動生成する。この際、係り受け規則を記述する情報として意味属性を使用しているが、意味属性のみで係り受け規則を記述するのには不適切な場合がある。そこで本稿では、意味属性の他に品詞コードを用いて係り受け規則を記述する。具体的には、1) 品詞コードのみを使用した係り受け規則、2) 品詞コードおよび意味属性を混合した係り受け規則、3) 意味属性のみを使用した係り受け規則の3種類の係り受け規則を生成し、得られた係り受け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

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」型の名詞句を考える。ただし、記号 A、B、C は名詞の出現順序を表す。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。

2.2 名詞句の構造と意味の問題

通常、係り受け解析では前節で述べたように、「A の B の C」を「A の B」と「A の C」の2つに分類する。しかし、この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、必ずしもこの原理が成り立つとは言えず、表現の構造と意味の関係を考えなければならない場合も多い。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 係り受け規則の生成

3.1 解析用データの形式

解析される「の」型名詞句標本データは3つの属性と1つのクラスからなる集合である。3つの属性は名詞Aの属性、名詞Bの属性、名詞Cの属性であり、クラスはB-係り、C-係りの2種類である。

3.2 係り受け規則のタイプ

係り受け規則は以下の3タイプ、計7種類である。

1) 単一の名詞に着目した規則 (1次元規則)
この規則は、構成要素の名詞の1つの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つの名詞の属性とは無関係に、係り受け関係が決定できる規則である。係り受け規則は、次の3種類に分けられる。
(X, *, * : D), (*, Y, * : D), (*, *, Z : D)
* : 任意の属性

2)2つの名詞に着目した規則 (2次元規則)
この規則は、2つの名詞の属性とそれらの出現位置が与えられれば、残りの名詞の属性とは無関係に係り受け構造が決定できる規則である。係り受け規則は次の3種類に分類される。
(X, Y, * : D), (*, Y, Z : D), (X, *, Z : D)

3)3つの名詞に着目した規則 (3次元規則)
この規則は、3つの名詞すべての属性で係り受け構造が決定できる規則で次の1種類である。
(X, Y, Z : D)

3.3 係り受け規則の記述に使用する情報

係り受け規則の記述に使用できる情報は以下の3つが考えられる。
1) 字面
2) 品詞コード
3) 意味属性 (意味属性体系 [5])

この時、字面は慣用表現を記述する際に使用される。よって、慣用表現以外の係り受け規則の記述に使用される情報は、品詞コードと意味属性である。

次に、品詞コードと意味属性を使用して係り受け規則を生成する場合、以下の3つのパターンが考えられる。
1) 品詞コードのみの規則
2) 品詞コードと意味属性を混合した規則
3) 意味属性のみの規則

次節より上記の3つの係り受け規則の生成法について述べる。

3.4 品詞コードのみの規則

品詞コードのみを使用して係り受け規則を生成する場合、品詞コードはシンボルとみなせるため、決定木を利用することができる。以下では、決定木を利用した係り受け規則の生成法について述べる。

3.4.1 使用する品詞コード

3つの属性に使用する値は、名詞を細分化した26種の品詞コードを用いる。表1に26種の品詞コードを示す。

3.4.2 決定木およびプロダクションルールの生成

決定木の生成には決定木生成プログラムC4.5 [6]を用いる。また人間がモデルをよく理解できるようにするために同じくC4.5を用い、決定木からプロダクションルールを生成する。プロダクションルールの形式は、L->Rである。ここで、条件部のLは属性に基づいたテストの積集合であり、結論部のRはクラスである。クラスの1つはデフォルトクラスに選ばれる。プロダクションルールを使用しての事例の分類は、ルールの条件部に適合すればそのクラスが予測されたクラスであり、どのルールの条件部も満足されなかったならば、その事例はデフォルトクラスに属していると見なされる。

表 1: 使用する品詞コード

一般名詞
一般名詞 (副詞型名詞)
一般名詞 (連体詞型名詞)
用言性名詞 (サ変動詞型 (自))
用言性名詞 (サ変動詞型 (他))
用言性名詞 (サ変動詞型 (自他))
用言性名詞 (形容動詞型 (ダ形 ～な、～に))
用言性名詞 (形容動詞型 (ダ形 ～な))
用言性名詞 (形容動詞型 (タルト形 ～たる、～と))
用言性名詞 (形容動詞型 (タルト形 ～と))
転生名詞 (動詞転生型 (自))
転生名詞 (動詞転生型 (他))
転生名詞 (動詞転生型 (自他))
転生名詞 (形容詞転生型)
時詞
数詞 (数詞のみ)
数詞 (数詞＋助数詞)
代名詞 (人称代名詞)
代名詞 (指示代名詞)
形式名詞
固有名詞
固有名詞 (姓)
固有名詞 (名)
固有名詞 (地名)
固有名詞 (組織名)
その他の固有名詞

3.5 品詞コードと意味属性を混合した規則

品詞コードと意味属性を混合した規則は主に同格抽出用に使用される。この場合、同格を抽出する規則は人手で規定できるため、今回は以下の2つの規則を人手で作成した。

1) 人名を含む同格の規則 (A の B)
A：品詞コード："not 固有名詞 (姓)(名)" かつ
　意味属性：＜人＞
B：品詞コード："固有名詞 (姓)(名)" かつ
　意味属性：＜人＞

2) 地名を含む同格の規則 (A の B)
A：品詞コード："not 固有名詞 (地名)" かつ
　意味属性：＜地域＞
B：品詞コード："固有名詞 (地名)" かつ
　意味属性：＜地域＞

3.6 意味属性のみの規則

3.6.1 意味属性のみの規則の種類

意味属性のみを使用した係り受け規則の生成方法は汎化を用いた学習によって生成する [4]。この方法の特徴は以下の通りである。
「の」型名詞句の係り受け解析では、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できると仮定すると、すべての係り受け規則は3つの名詞の意味属性の組で表現される。ここで「の」型名詞の係り受けの特徴を見てみると、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。
例えば、「私 (A) の本当 (B) の父 (C)」では、「私」は「本当」に係ることができないので c-係りである。つまりこの名詞句は「本当 (B)」のみによって係り先を決定できる。
そこで (A, B, C) の3つの名詞の同時共起で「の」型名詞句の係り受け関係を捉えるだけでなく、(A, B), (B, C), (C, A) のそれぞれ2つの名詞の関係、およびそれら3つの関係、さらに (A), (B), (C) それぞれ1つの意味属性で係り受け関係を捉えるほうが、対象とする名詞句の構造の特徴をより良く捉えられると考えられる。
よって「の」型名詞句の標本データから以下の3つのタイプ、計7種類の係り受け規則を汎化を用いて自動生成する。

3.6.2 意味属性のみの係り受け規則生成法の改良

汎化を用いた係り受け規則の生成において、抽象度の高い名詞 (意味属性体系で上位の意味属性) をあらかじめ除いておくことが必要である。この理由は、抽象度の高い意味属性は下位の意味属性と上位下位関係が成り立たない場合があり、それぞれ分けておかないと必要な係り受け規則が生成されない場合がある。今回は、抽象度の高い意味属性をシソーラス上で深さ0、1、2にある意味属性、計9種と規定して、これらの意味属性を含む用例はあらかじめ除き、汎化を用いた係り受け規則の生成には使用しない。

以下に9種の意味属性をあげる。(＜x：n＞：x:

意味属性の名称、n:シソーラス上の深さ)
＜名詞：0＞、＜具体：1＞、＜抽象：1＞、
＜主体：2＞、＜場：2＞、＜具体物：2＞、
＜抽象物：2＞、＜事：2＞、＜抽象的関係：2＞

表 2: 品詞コードのみを使用した規則と意味属性のみを使用した規則の正解率

	品詞コードのみを使用した規則	意味属性のみを使用した規則
カバー率	91.8%	96.0%
適合率	84.8%	88.4%
正解率	77.8%	85.1%

4 実験

4.1 実験対象と実験方法

実験に使用する名詞句データ(「A の B の C」)は、新潮文庫小説 100 冊 (約 900 万単語) より抽出した 10,021 個のデータを用いた。ここで品詞コードは形態素解析プログラム ALT-JAWS の結果を使用した。意味属性は一般に 1 つの名詞に複数付与されているが、今回は人手で付与した。また、係り先についても人手で付与した。

なお、実験は 10 分割 Cross Validation で行った。

実験は以下の 3 種類を行った。
1) 品詞コードのみを使用した係り受け規則を実験データに適用する
2) 意味属性のみを使用した係り受け規則を実験データに適用する
3) 品詞コードのみを使用した係り受け規則、品詞コードと意味属性を混合した規則、意味属性のみを使用した係り受け規則を組み合わせて実験データに適用する
3) の 3 つの係り受け規則を実験データに適用する順序は、まず品詞コードと意味属性を混合した規則によって同格を含む用例を抽出する。残った学習データに対し、決定木を構築し予測正解率 α %以上の係り受け規則を実験データに適用する。ここで予測正解率とは、各プロダクションルール毎の未知の用例に対する予想される正解率である。最後に残った学習データに対し、汎化を用いた学習を行い係り受け規則を生成し、実験データに適用する。

4.2 実験結果

4.2.1 品詞コードのみを使用した係り受け規則を適用した場合の正解率

表 2 に決定木を用い品詞コードのみを使用した係り受け規則を適用した場合のカバー率、適合率、正解率を示す。ここで、デフォルトルールしか適用されない用例に対しては、係り先は決定しないとした。なおカバー率および適合率の定義は以下の通りである。
カバー率＝計算機が係り先を決定した個数／全実験データの個数
適合率＝係り先の正解の個数／計算機が係り先を決定した個数
正解率＝カバー率×適合率

表 2 より、品詞コードを用いた係り受け規則は、意味属性のみを使用した係り受け規則を適用した場合より、正解率が低い。これについて以下の理由が考えられる。
1) 品詞コードの分類が少ない。
2) 品詞コードの大半が (一般名詞) である。
3) 品詞コードの中でも名詞句の係り先に影響を与える品詞は一部に限られる。

表 3 に、生成された品詞コードのみを使用した係り受け規則のうち、予測正解率の上位 (予想正解率 85%以上) の係り受け規則とその規則に当てはまる用例をあげる。

表 3: 予測正解率上位の係り受け規則

1	(*) + (形式名詞) + (*):B-係り 昔のままの姿　　大人のための物語
2	(*) + (*) + (形式名詞):B-係り 玄関の石段のところ　　声楽の勉強のため
3	(*) + (指示代名詞) + (*):B-係り 湿原の向うの林　　海の彼方の国々
4	(*) + (*) + (副詞型名詞):B-係り 彼の小説の数々　　新宿の二丁目の近く
5	(*) + (*) + (指示代名詞):B-係り 窓のガラスの向う　　逢坂の関の彼方
6	(*) + (形容詞転生型) + (*):B-係り 事件の残酷さの意味　　もとの静けさのなか
7	(*) + (動詞転生型 (自)) + (*):B-係り 砂のくぼみの中　　岬のつづきの丘
8	(人称代名詞) + (サ変動詞型 (他)) + (*):B-係り 彼の指揮の下　　私たちの受験の頃
9	(指示代名詞) + (*) + (*):B-係り ここの学校の子　　こちらの膝の上
10	(*) + (*) + (時詞):B-係り 父の死の直前　　山本の訓示のあと
11	(*) + (サ変動詞型 (自他)) + (*):B-係り ピアノの稽古のため　　司祭の不在の間
12	(*) + (*) + (形容詞転生型):B-係り 漁夫の生活の厳しさ　　父の声の暗さ

表 4: 品詞コードのみと意味属性を混合した規則を適用した場合の正解率

	人名を含む同格	地名を含む同格
カバー率	74.5%(76/102)	100.0%(13/13)
適合率	100.0%(76/76)	100.0%(13/13)
正解率	74.5%(76/102)	100.0%(13/13)

4.2.2 品詞コードと意味属性を混合した規則を適用した場合の正解率

表 4 に品詞コードと意味属性を混合した規則を適用した場合の正解率を示す。
この規則は同格を抽出するために生成された規則であるため、正解率の分母は実際の同格を含む用例の数である。
表 4 より、同格を含む用例は全実験データ (10,021 個) の約 1%程度にすぎない。しかし、意味属性のみを使用した係り受け規則を生成する場合、品詞情報を持たないため、同格かそうでないかの判断ができず、正しい係り受け規則が生成されない。よって、意味属性のみを使用した係り受け規則を生成する前に、同格を含む用例の抽出を行うことが必要である。

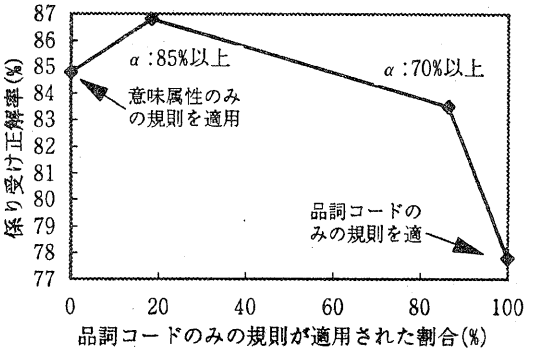

図 1: 品詞コードのみの規則と意味属性のみの規則の適用の割合を変化させたときの係り受け解析正解率

4.2.3 品詞コードのみを使用した係り受け規則、品詞コードと意味属性を混合した規則、意味属性のみを使用した係り受け規則を組み合わせた場合の正解率

図 1 に品詞コードのみの規則と意味属性のみの規則の適用の割合を変化させたときの係り受け解析正解率の変化を示す。ここで、品詞コードと意味属性を混合した係り受け規則は同格を含む用例を抽出する目的で、一番初めに適用される。

本実験では、予測正解率 85%以上および 70%以上の品詞コードのみを使用した係り受け規則を適用し、適用されなかった実験データに対し、意味属性のみを使用した係り受け規則を適用し、最終的な正解率を出した。

結果は、図 1 より予測正解率 85%以上の品詞コードのみを使用した係り受け規則を適用し、その後意味属性のみを使用した係り受け規則を適用した場合、最も係り受け解析正解率が高くなった。これより、品詞コードを使用した係り受け規則の有効性が示された。

5 考察

「A の B の C」型名詞句の係り受け解析において、正解率の向上を妨げる原因の一つに人間が係り先を付与する際、係り先に揺れが生じるといった問題がある。この原因として、従来は「A の B の C」

型名詞句の係り受け構造をB-係り、C-係りの2分類にして考えてきたが、2分類ではどちらの構造に当てはまるか迷う用例が数多く出現し、係り先を統一できないためと考えられる。

そこで、従来のB-係り、C-係りの2分類をそれぞれさらに2分類し、さらに同格の2種類を追加し、計6つの構造に細分化する。人間が係り先を付与する場合、この6つの構造に分類することにより、より正確に構造を分類できると考える。

また、係り受け規則の生成についても解析データのクラスを2分類から6分類にすることにより、それぞれの係り受け構造をよく特徴づける係り受け規則が生成できると考える。

以下に6分類した構造を示す。

○ B-係り A → B(&B → C)

(B-1)A → B(&B → C)[A → C(共起する),A not → C(意味的には成り立たない)]

ex. 私の母の名前

(B-2)A → B(&B → C)[A not → C(共起する)]

ex. 赤色のブランコのそば

○ C-係り A → C(&B → C)

(C-1)A → C(&B → C)[A → B(共起する)]

ex. 私の自転車のカギ

(C-2)A → C(&B → C)[A not → B(共起しない)]

ex. 私の本当の父

○同格

(D-1)A=B(&B → C)

ex. 同僚の田中さんの机

(D-2)A → B(&B=C)

ex. 会社の同僚の田中さん

6 おわりに

本稿では、構造に多義を持つ日本語名詞句の中でも、基本的でかつ高頻度で現れる「A の B の C」を対象に、品詞情報および意味属性を用いた係り受け規則の自動生成法を提案した。生成された係り受け規則は品詞情報のみを使用した規則、品詞情報と意味属性を混合した規則、意味属性のみを使用した規則の3種類である。これら3種類の係り受け規則を「A の B の C」型の名詞句の係り受け解析に組み合わせて適用した結果、係り受け解析正解率が86.8%となり意味属性のみの係り受け規則を適用した場合より向上した。これより、「A の B の C」型の名詞句の係り受け解析には品詞コードと意味属性の両方を使用することが有効であることが分かった。なお、本方式は他の名詞句の係り受け解析にも適用可能と思われる。

参考文献

[1] 佐々木, 坂本：文書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.*, pp. 585–594 (1992).

[4] 中井, 池原, 白井：「の」型名詞句における名詞間の係り受け規則の自 動生成法, 信学技報, Vol. 98, No. 53, pp. 15–22 (1998).

[5] 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).

[6] Quinlan, J.: AI によるデータ解析, トッパン (1995).