日本語表現に対する意味的構造規則の自動獲得法の研究

中井 慎司

1999年2月19日

概要

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。従来、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方法では、これらの問題を解決するのは困難であった。

本論文では、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、名詞意味属性を用いた構造規則の自動生成法を提案した。この手法は名詞句の持つ意味的な構造に着目して、1つの名詞句学習データから、3つの名詞A、B、C間の意味的構造規則（3タイプ、計7種類）を自動生成する。次に、得られた構造規則を別の名詞句標本データ(10,021個)に汎用規則（1，2，3次元規則の順）から適用した。その結果、判定率96.0%、適合率88.4%の精度で係り先が判定でき、最終的な正解率は85.1%であった。これより、提案した方法は名詞句に対し、精度の高い構造規則が得られることが分かった。

目 次

第1章　はじめに　1

1.1　研究背景と目的　1

1.2　本論文の構成　2

第2章　意味属性を使用した構造規則の記述　3

2.1　構造規則の一般形　3

2.2　「の」名詞句の構造の意味表現　3

2.2.1　「の」型名詞句の規則構造の形式　3

2.2.2　「の」型名詞句の構造　3

2.2.3　名詞句の構造と意味の問題　4

第3章　汎化を用いた構造規則の自動生成　5

3.1　汎化を用いた構造規則の一般化　5

3.1.1　背景知識を用いた事例からの学習　5

3.1.2　汎化を用いた構造規則の生成アルゴリズム　6

3.2　特徴次元の抽出　6

3.2.1　複数の学習データの作成　6

3.2.2　複数の構造規則の適用　8

第4章　実験　10

4.1　実験対象とその構造　10

4.1.1　実験対象　10

4.1.2　実験の手順　10

4.1.3　実験対象データの性質　10

4.2　結果　11

4.2.1　個別の規則ごとの適用結果　11

4.2.2 組み合わせた場合の適用結果 13
4.2.3 単一次元と本手法との比較 13

第5章 考察 15
5.1 品詞情報を用いた規則生成 15
5.1.1 品詞情報のみを用いた規則生成法 15
5.1.2 品詞コードと意味属性を混合した構造規則 17
5.1.3 品詞コードのみを使用した構造規則、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構造規則を組み合わせた場合の正解率 19
5.2 包含関係を持つ規則生成 20
5.2.1 名詞句の性質 20
5.2.2 包含関係を持つ規則 20
5.2.3 包含関係を持つ規則の生成法 21
5.3 「の」型名詞句の構造の細分類 22

第6章 結論 24

図 目 次

3.1 汎化の過程の例 7

4.1 実験の手順 11

5.1 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 19
5.2 包含関係を持つ構造規則 21
5.3 包含関係を持つ構造規則の関係 22

表 目 次

4.1 名詞と意味属性の関係 12
4.2 自動生成されたここの構造規則の数と適用回数および正解率 12
4.3 汎用的規則から適用した場合のカバー率および適合率 13
4.4 単一次元と本手法との比較 14

5.1 使用する品詞コード 16
5.2 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 . 17
5.3 予測正解率上位の構造規則 18
5.4 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 . . . 19

第1章　はじめに

1.1　研究背景と目的

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[4]、意味的クラスの共起情報を用いて係り先を決定する方法[5][6]、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[2]などが提案されてきた。しかし、共起情報を用いる手法は、意味的な関係を考慮に入れないため、精度に限界がある。また、用例翻訳は、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、動詞の意味の解析用に意味属性体系[1]が開発された。この辞書の知識は、格要素として使用された名詞の意味を動詞との共起によって解析する場合も使用できる。さらに、名詞の意味を詳細に分類（約2,700種類）しているため、そこで規定された単語意味属性の情報は、名詞句内で使用された名詞の意味解析にも有効と期待される。

そこで、名詞句の持つ意味的な構造に着目して、名詞と名詞の間の構造規則を意味属性を用いて記述する。このとき、構造規則を標本から得られたデータから自動生成する。

具体的には、日本語名詞句の中でも最も基本的でかつ、高頻度で出現する「の」型名詞句(「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概念を用い、3つの名詞A、B、C間の意味的構造規則を生成する。このとき、3つの名詞すべてが決まらなくても係り受け関係が決定できることを考慮にいれ、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の3タイプの規則を名詞句標本データから自動生成する手法を提案する。そして、得られた構造規則を別の名詞句標本に適用して解析精度を評価する。

1.2 本論文の構成

本論文では、2章で意味属性をしようした意味構造表現について述べる。3章では、日本語表現からの意味的構造規則の自動生成について説明する。4章では、3章で説明した手法を日本語名詞句に適用する。5章では、4章の実験結果についての考察および、規則生成の拡張について述べる。そして、5章で本研究をまとめ、今後の課題を述べる。

第2章　意味属性を使用した構造規則の記述

2.1　構造規則の一般形

構造規則の一般形を以下に示す。

$(A_1,A_2,...,A_n$:C)

A_i:属性 C:クラス

本論文では、属性が意味属性である場合を考える。ここで、クラスは自然言語処理分野では、係り受け構造、意味関係、対訳などが考えられる。

2.2　「の」名詞句の構造の意味表現

2.2.1　「の」型名詞句の規則構造の形式

「の」型名詞句(「AのBのC」: A,B,Cは単語)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、すべての構造規則は3つの名詞の意味属性の組で表現される。この場合の構造規則は以下のように表される。

(X,Y,Z:D)

属性：X,Y,Z(X,Y,Zは単語A,B,Cの意味属性)

クラス：D(係り先：B係りとC係りの2種類)

2.2.2　「の」型名詞句の構造

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。

2.2.3　名詞句の構造と意味の問題

従来の共起頻度を用いた係り受け解析では、「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。

第3章　汎化を用いた構造規則の自動生成

3.1　汎化を用いた構造規則の一般化

3.1.1　背景知識を用いた事例からの学習

事例からの学習

事例からの学習技術とは、事例がもつクラスに従って、事例の表現空間をあらかじめ分割しておき、未知の事例に対し、その分割に従って事例のクラスを推定する技術である。しかし、収集できる学習データの数が少ない場合に学習結果の精度が不十分なことがある。そこで、背景知識を利用してこの問題を解決する。

背景知識の利用

学習データの少なさによる正解率低下を防ぐための手法に、背景知識の利用がある。ここでは対象世界における概念の上下階層関係を背景知識として利用する。自然言語処理分野では、意味に関する知識を表現するためにis-a関係を使用し、木構造に表現する。名詞を例に挙げると、名詞間の関係は、is-a関係およびhas-a関係を用いシソーラスの形で表現される[1]。背景知識を利用すれば、少数の事例からそれらをカバーする知識を生成できる。

汎化による構造規則の一般化

例として名詞句「AのB」の意味関係の付与について考える。標本から得られた例として「(彼)の(自転車):所有関係」、「(先生)の(机):所有関係」の2つの場合を考える。背景知識を用い、この2つの用例の「彼」「先生」を<人>、「自転車」「机」を<人工物>に汎化することで、より一般的な知識「<人>の<人工物>:所有関係」を導くことができる。

3.1.2　汎化を用いた構造規則の生成アルゴリズム

汎化を用いた構造規則の生成の手順を以下に示す。

1. すべての学習データを最終的に得られる構造規則とする。
2. それらの構造規則を少しづつ汎化して一般的な構造規則を生成していく。
   ここで汎化とは、階層構造を持つ属性のあるノードをひとつ上の親ノードに変更することにより、より抽象度の高いノードにしていく操作である。
3. 汎化の過程で、汎化の対象としている属性が以下の2つの条件のどちらかを満たした時、その属性に対する汎化を終了させる。
   (a) あるノードを汎化した時、他のクラスの領域と重複した場合（図3.1参照）
   (b) あるノードを汎化した時、その汎化によって新しい用例が包含されない場合
4. 2.と3.を繰り返し、どの属性も汎化できなくなった時、最終的に処理を終了させる。

条件3aについては、実際にはノイズを考慮に入れ、次のように条件を緩くするのが良いと思われる。

> あるノードを汎化した時、他のクラスの領域を包含する場合でも、不正解率がある一定の閾値内であれば、他のクラスの用例をノイズとみなし、汎化していく。

これは決定木の枝刈り（pruning）に相当する。この時の最適な閾値は、学習データに依存するので実験的に得る必要がある。

条件3bは、構造規則の記述は、より広い範囲をより詳細な規則で記述するという一般的な規則の記述の方針に従って、それ以上汎化可能であっても、汎化をストップする。これにより、必要な範囲だけをカバーする構造規則が得られ、人間がみて理解しやすい構造規則が得られる。なお、テストデータに構造規則を適用する時、どの構造規則にも当てはまらない場合もあるので、デフォルトルールを作成しておく必要がある。

3.2　特徴次元の抽出

3.2.1　複数の学習データの作成

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