コーパスに基づく日本語文法の自動獲得法

1H－1

横田 和章　阿部 賢司　藤崎 博也
東京理科大学　基礎工学部

1. はじめに

近年、計算機のハードウェアの進歩に伴い、ワープロのかな漢字変換や構文チェック機能など、自然言語の解析処理が実用化されつつある。しかし、これらの機能は構文+浅い意味情報に基づいており、人間による修正無しでは満足な結果が得られない。一方、より良い結果を得るため、文の深い意味を調べる方法も報告されているが、この方法では、解析に使う知識をあらかじめ明示的に組み込んでおかなければならない。このため、処理対象となる話題を限定しないと、組み込むべき知識が膨大になって統一がとれなくなり、実現困難となる欠点を持つ。

そこで、コーパスから言語的知識を獲得する方法が最近注目されている [1, 2, 3]。この方法では知識をあらかじめ書き込んでおかなくても、獲得により性能を向上できる特徴を持つ。また、新しい話題に対しても、獲得により対応できる。本稿では、この様な見地から、コーパスに基づいて日本語文法を自動獲得する方法について報告する。

2. 文法モデル

一般の構文解析法では、図 1(b) の様な文法規則を構成し、その規則に基づいて、テキストから同図 (a) の様に構文木を作成する。この構文木の各節点には、それぞれ１つのラベルが対応している。

従って、逆にこの構文木から文法を獲得するには、各節点にラベルをつけて、節点の親子関係から規則を生成すれば良い。ところが、ラベル数や規則数を無限大に増やさずに各節点にラベルをつけるのは難しい。

これに対し、ここで提案する方法では、図 2(a) の様に１つの節点に対し２つのラベルをつける。このうち左のラベルはその句の左に共起する句の性質を、

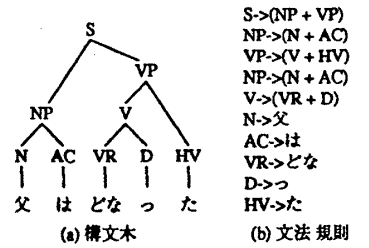

図 1. 一般の構文解析で用いる文法規則の例

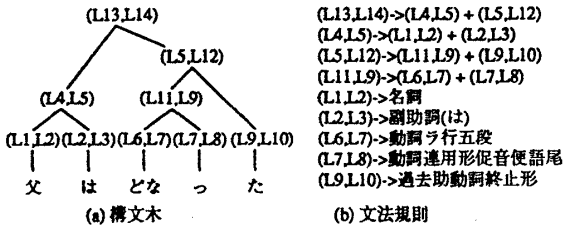

図 2. 本研究で用いる文法規則の例

右のラベルはその句の右に共起する句の性質を示すものとする。そして、ある句が別の句と共起する時、先行する句の右ラベルと後接する句の左ラベルに同じラベルを与えることによりラベル数を制限する。

この木においては葉の節点は形態素である。このため、各形態素がどの様な左ラベルと右ラベルを持ちうるかという規則が別に必要になる。本研究で用いる EDR 電子化辞書では、各形態素は接続属性という情報によって分類されている [4]。従って本研究では、この接続属性を左右ラベルに変換する規則も作成する。すると、文法は、図 2(b) の様に表される。

3. 文法の自動獲得法

日本語の任意の文を処理するのに必要な文法を、EDR 電子化辞書研究所の EDR コーパスから自動的に獲得することを試みる。このコーパスには、新聞記事などの例文が約 20 万文収録されており、例文を構成している形態素の情報や、構文木を作るための情報が付与されている。従って本研究では、この構文木から文法を作成して以後の解析に用いる。

まず、これらの情報を利用して構文木を生成し、次

A Method for Automatic Acquisition of the Japanese Grammar
Kazuaki Yokota, Kenji Abe, Hiroya Fujisaki
Science University of Tokyo
2641 Yamazaki, Noda, 278, Japan

に下位の節点から以下の様な操作を施す。

1. その節点が葉の時

節点が葉の場合は、その節点は形態素に対応する。その形態素の接続属性を C とする。今まで獲得した文法の中に $(L_x, L_y) \rightarrow (C)$ という形の規則があれば、その節点に (L_x, L_y) というラベルを付与する。ないときは、新たなラベル L_1、L_2 を生成してその節点に付与し、次の様な規則を文法に加える (獲得する)。

$$(L_1, L_2) \rightarrow (C) \qquad (1)$$

2. その節点が葉でない時

下位の節点の左右ラベルをそれぞれ (L_1, L_2)、(L_3, L_4) とする。今まで獲得した文法の中に $(L_x, L_y) \rightarrow (L_1, L_2) + (L_3, L_4)$ という形の規則があれば、その節点にラベル (L_x, L_y) を付与する。ないときは、新しいラベル L_5、L_6 を生成し、次の様な規則を獲得する。

$$(L_5, L_6) \rightarrow (L_1, L_2) + (L_2, L_4) \qquad (2)$$

そして、今まで獲得した文法に現れる L_3 を全て L_2 に置き換える (統一化)。

この様な操作を各節点について行ない、規則を書える。重複した規則は消去する。

もし、自然言語の文法を構成する規則の数が有限であるとするならば、数多くの文に対してこの操作を行なうことにより、全ての規則を獲得しうるものと期待される。

4. 獲得した文法の評価

上記の方法によって文法を獲得するシステムを作り、獲得した文法に基づいて実際に構文解析を行なうことにより評価を行った。その結果を表 1に示す。

表中、学習文数とは文法を獲得するために用いた文の数である。獲得した文法を用いて、別に用意した 100 文を構文解析し、構文木を得た。この構文木が、コーパスに付与されていたものと一致した場合を正解とした。

表 1 文法自動獲得結果

学習文数	中間ラベル数	規則数数	ラベル数	正解率 [%]
100	49	456	527	38
200	42	669	765	54
300	47	844	872	60
400	43	871	973	52
500	43	951	1061	46
1000	9	1405	1624	11

解析は最長一致法によった。表中の中間ラベルとは、式 (2) における L_2 の様な、二つの節点をつなぐラベルのことを示す。

正解率は学習文数 300 文では 60%に達するが、この点をピークに減少している。この原因は中間ラベル数の減少にあると考えられる。中間ラベル数が減ると、構文解析処理でのあいまい性が増す。この減少は、今回用いた EDR コーパスに不適切な係り受けが存在し、前節の手順 2 により、本来異なるはずの 2 つのラベルが統一化されることに起因する。

従って、本方法による結果はコーパスの質に大きく依存する。

5. まとめ

本稿では、教師入力として与えた構文木から文法を獲得する方法を示した。この方法による解析の正解率は、コーパスに含まれる誤りによって、大きく影響されることが明らかとなった。現在、この点を改善するため、統計的な情報を用いて、若干の誤りを含むコーパスからでも正しい文法を獲得できる方式を検討中である。

参考文献

[1] 宇津呂武仁, 松本裕治: コーパスを用いた言語知識の獲得, 人工知能学会誌, vol.10, No.2, pp.197-204(1995).

[2] 工藤育男, 井上直己: コーパスに基づく共起知識の獲得とその応用, 人工知能学会誌, vol.10, No.2, pp.205-212(1995).

[3] 横田和章, 藤崎博也: コーパスに基づく構文木の自動生成法, 平成 7 年春期電子情報通信学会全国大会論文集, vol D, p.120(1995).

[4] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (1993).

[5] 藤崎博也, 亀田弘之: 人間の言語処理過程のモデルに基づく自然言語理解システムの構築, 言語情報処理の高度化研究報告 8, 言語情報処理の高度化の諸問題, pp.657-684(1989).