2R–5　構文構造付きコーパスからの確率文脈自由文法の自動抽出

白井清昭　　徳永健伸　　田中穂積

東京工業大学理工学研究科

1　はじめに

コーパスなどの言語データから自動的に獲得された文脈自由文法は適用範囲が広く、また文法獲得に要する人的負担が軽いなどの利点がある。文法の自動獲得に関する研究としては、Inside-Outside アルゴリズム [2][3] によるものや、統語解析に失敗した事例から文法規則を獲得する研究 [1] などがあるが、多くの計算量を必要とするなどの問題点も多い。本論文では、括弧付けによる構文構造が付加されたコーパスから、効率良くかつ実用的な確率文脈自由文法を自動的に抽出する方法を提案する [4]。

2　確率文脈自由文法の抽出

図 1 は、本論文で使用する EDR コーパスのテキスト、およびそれに付加された括弧付けによる構文構造の例である。

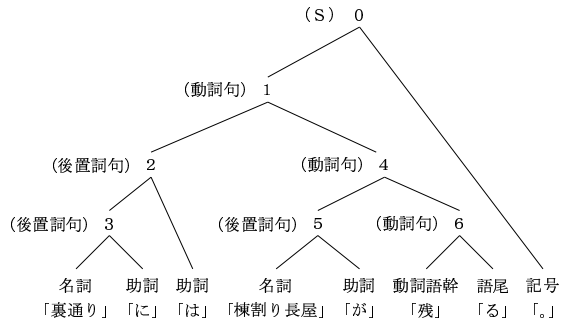


図 1: EDR コーパスの例文とその構文構造

この構文構造から文法を抽出するには、数字で表されている内部ノードに適切な非終端記号を割り当てればよい。ここでは、前の要素が後ろの要素を修飾する、すなわち句の主辞はその句における一番最後の要素であるという日本語の特徴に着目し、句の一番最後の要素に “句” をつけた非終端記号 (e.g. “名詞句”) を、その句の親ノードに与えることにした。また、主辞が非終端記号 (e.g. “名詞句”) である場合には、右再帰を用いてそれをそのまま親ノードに与えることにした。ただし、次のような場合は例外的な非終端記号を与える。

- 句の一番最後の要素が “語尾”, “記号” のとき
  これらの品詞は主辞にはならないので、その左の要素を主辞とみなして親ノードに非終端記号を与える。
  例.　6 → 動詞語幹 語尾 (6 を “動詞句” に)
- 主辞が “助詞” のとき
  左辺に “後置詞句” という非終端記号を与える。
- 主辞が “接尾語” のとき
  EDR コーパスにおいては、“接尾語” は「日」「メートル」などの単位が多く、句全体で名詞句を構成していると考えられるので、その親ノードには “名詞句” という非終端記号を与える。
- 構文構造の根ノードには、開始記号 “S” を与える。

図 1 の各ノードの左にある (　) 内の記号が、上記の方法によりノードに与えられた非終端記号である。

以上の操作をコーパスの全ての例文に対して行い、非終端記号のついた構文木を文法規則に変換することにより、文法を抽出することができる。また、抽出した規則の確率は、その規則のコーパスにおける出現回数をもとに、式 (1) のように推定することにした。

$$P(A \to \zeta) = \frac{A \to \zeta \text{の出現回数}}{\sum_i A \to \zeta_i \text{の出現回数}} \qquad (1)$$

3　文法の改良

この章では、抽出した文法を改良するいくつかの方法について述べる。

3.1　誤りと思われる規則の削除

“語尾” が右辺の先頭に現れる規則など、誤りと思われる規則を半自動的に検出した。63 個の規則を誤りと判断し、それを削除した。

Automatic Extraction of Probabilistic Context-Free Grammar from Bracketed Corpus
Kiyoaki Shirai, Takenobu Tokunaga, Hozumi Tanaka
Department of Computer Science
Tokyo Institute of Technology
2–12–1 Ookayama Meguro–ku, Tokyo, 152 JAPAN

3.2 冗長な規則の削除

文法の生成能力を低下させることなく文法サイズを縮小するために、文法規則の中から次に定義する冗長な規則を自動的に検出し、それを削除した。

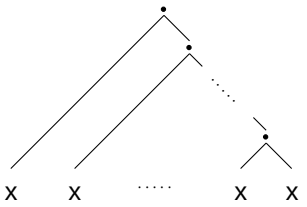
> ある規則 $A \to \zeta$ に対して、文法中のそれ以外の規則を用いて非終端記号 A を記号列 ζ に展開できる場合、その規則は冗長である。

また、削除した冗長な規則の出現回数は、その規則の右辺の記号列を生成するのに用いた規則の出現回数に加え、残った規則の確率の推定に反映させた。

3.3 統語的曖昧性の抑制

1. 同一品詞列の取り扱い

   “名詞 名詞 名詞 ...” などのような同じ品詞によって構成される句の構造の解析は意味情報を必要とする。そこで、同一品詞列の構造の解析は意味解析の段階で行うことにし、統語解析の段階では下図のような右下がりの構造のみを生成し、統語的曖昧性を抑制するように文法を修正した。



2. 品詞の細分化

   EDR コーパスで用いられている品詞は 15 種類しかない。そこで、“記号” と “助詞” の品詞を次のように細分化した。

   - 「、」には “読点” という品詞を与えた。
   - 「。」と「？」については “文末記号” という品詞を与えた。
   - 助詞毎にユニークな品詞を与えた。例えば、「が」には “助詞が” という品詞を与えた。

   このように品詞を細分化してから文法を抽出することにより、統語的曖昧性を約 10 %に減少させることができた。

4 実験

EDR コーパスの 75,000 文のうち、90 %を訓練データとして確率文脈自由文法を抽出したところ、1781 個の規則からなる文法が得られた。次に、残りの 10 %の文をテストデータとし、抽出した文法を用いてこれらの文の統語解析を行った。結果は、以下の通りである。

解析成功率	93.99 %
平均解析木数	9.33×10^9
括弧付けの正解率	74.78 %
文の正解率	24.97 %

括弧付けの正解率と文の正解率は、生成確率の上位 15 位の解析木のうち、矛盾する括弧付けの最も少ないものについて調べたものである。前者はコーパスの括弧付けと矛盾しない解析木の括弧付けの割合を示し、後者は全ての括弧付けがコーパスの括弧付けと矛盾しない解析木が得られた文の割合を示している。94 % 近い解析成功率が得られていることから、適用範囲の広い文法が抽出できたことがわかる。

5 まとめ

構文構造付きコーパスの内部ノードに非終端記号を与えて確率文脈自由文法を抽出する方法と、それを改善する方法をいくつか提案した。実験により適用範囲の広い文法が得られたことを確認したが、依然としてかなりの統語的曖昧性が生じることもわかった。これを抑制するような解決策を考えることが、今後の課題としてあげられる。

謝辞 本研究で使用した EDR コーパスの利用を許可くださいました日本電子化辞書研究所の横井所長、コーパスに関する技術的な支援をしてくださいました日本電子化辞書研究所の仲尾氏に感謝いたします。

参考文献

[1] M. Kiyono and J. Tsujii. Hypothesis selection in grammar acquisition. In *COLING '94*, Vol. 2, pp. 837–841, 1994.

[2] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech and languages*, Vol. 4, pp. 35–56, 1990.

[3] Fernando Pereira and Yves Schabes. Inside-outside reestimation from partially bracketed corpora. In *ACL '92*, pp. 128–135, 7 1992.

[4] 白井清昭, 徳永健伸, 田中穂積. コーパスからの文法の自動抽出. 情報処理学会自然言語処理研究会, Vol. 101, pp. 81–88, 1994.