上向パス上の文法記号列からなる縦方向リストを利用した構文解析木の生成確率計算法

椎名 広光: 岡山理科大学 総合情報学部
増山 繁: 豊橋技術科学大学 知識情報工学系

1 はじめに

確率構文解析法は，HMMの拡張として音声認識における音素の予測モデルなどの統計的言語モデルや情報抽出の手法として利用されてきている．従来，文法規則に出現確率を付加したモデルでは，inside-outsideアルゴリズムなどによって生成規則を推定し，尤もらしい構文解析木を求めるモデルが知られている．これに対して，構文解析済みのデータを用いて学習し，学習結果から構文解析の生起確率を求める研究も行なわれてきている．

Briscoe と Carroll は，GLR 構文解析法を応用して状態と先読み文字列の組合せに対して出現確率を付加する確率GLR構文解析法 [2](以降，B&C法と呼ぶ)を提案している．また，Inuiら [4] によって提案されている確率 GLR 構文解析 (以降，Inui法と呼ぶ) では，出現確率の情報を持つ状態と先読み文字列の組合せの方法を工夫することによって，構文解析木の生起確率の計算をより精密なものへと改善している．GLR 構文解析法に対して，Manning [5] らは，GLR 構文解析の拡張法ではなく確率 LC 文法を定義し，文法に対する構文解析木の生起確率を計算している．

しかしながら，学習する構文解析木が複雑な場合は，B&C法，Inui法，Manning法のいずれの方法においても，構文解析木の高さ方向に関する情報を考慮していないため，複雑な構文解析木の生起確率の計算は誤差を多く含むようになってしまう．

そこで，本研究では，その構文解析木の高さ方向に関する情報として本稿で「LC 文法規則リスト」と読んでいる葉から根への文法規則の適用リストの出現頻度を用いる．これによって，構文解析木の生成確率の誤差を改善しようとするものである．また，これまでの確率構文解析法では，終端記号列に対する構文解析木の生起確率を与えていたのに対して，本手法では部分木のように葉の部分が非終端記号であってもその生起確率を求めることができる．

2 準備

文法として，文脈自由文法 $G = (P, N, T, S, \$)$, P: 生成規則の集合, N: 非終端記号の集合, T: 終端記号, S: 開始記号, \$: 入力の終りを示す特殊記号が与えられるものとする．この文法に対して，構文解析木の生起確率を求める確率構文解析法を $M = (G, L, P)$, G: 文法, L: 学習データ, P: 生起確率で表わすことにする．

また，部分木を含めた構文解析木 (T) をその葉に当たる入力文字列 W のもとにおける生起確率を

$$P(T|W) = \frac{\text{部分木 } T \text{ の出現個数}}{W \text{ の出現個数}}$$

で表わすものとする．なお，$W = W_1\, W_2 \ldots W_{|W|}$, $W \in T \cup N, i = 1, \ldots, |W|$ とする．

研究の目的は，$P^*(T|W)$ を生起確率の真の値であるとしたとき，提案する確率構文解析法を M とすると，

$$\sum_{W \in (T \cup N)^+} ||P^*(T|W) - P(T|W)||$$

ができるだけ小さくなるような生起確率 P を求める確率構文解析法 M を求めることである．

3 用語の定義

はじめに述べたように，構文解析木の高さ方向に関する情報を定義するのに，利用する LC 親，及び，LC 文法規則リストの 2 つの用語を定義する．なお，本稿では，図 1 の木を用いて，用語などを例示する．図 1 の木の葉に当たる部分は D,G,H,F の 4 つの文法記号で，木の根には A が対応する．また，\$ と < ROOT > は便宜上付している記号で，\$は入力の終りを示す記号を表わし，木の仮の根 < ROOT > の左側の子供である．

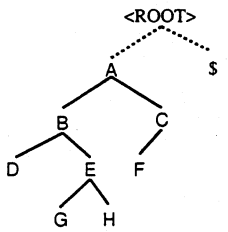


図 1: 木の例

3.1 LC 親

構文解析木やその部分解析木の葉から導出の最左に出現する親を辿っていく時，始めて出合う最左でない親を，「LC 親」とする．

図1に示す木では，葉に対応する文法記号 D,G,H,F に対して，表1に示す LC 親の文法記号が対応する．

表1: LC 親の例

葉	LC 親
D	A
G	E
H	H
F	C

3.2 LC 文法規則リスト

構文解析木を葉から木の根への上の方向に向けて辿るとする．そして，次の葉の LC 親が含まれる文法規則に到着するまでに得られる文法規則を逆順に格納したリストを「LC 文法規則リスト」と呼ぶこととする．

図1に示す木の葉 H に対しては，次の葉 F の LC 親が C であるので LC 文法規則リストは "$E \rightarrow GH$", "$B \rightarrow DE$", "$A \rightarrow BC$" の順で現れる文法規則リストの逆順となる．なお，H を含めた図1の木の葉に対応する LC 文法規則リストを表2に示す．

表2: LC 文法規則リスト

葉	LC 文法規則リスト
D	"$B \rightarrow DE$"
G	"$E \rightarrow GH$"
H	"$A \rightarrow BC$" $\rightarrow$ "$B \rightarrow DE$" $\rightarrow$ "$E \rightarrow GH$"
F	"< ROOT >$\rightarrow A\$$" $\rightarrow$ "$A \rightarrow BC$" $\rightarrow$ "$C \rightarrow F$"

4 構文解析木の生起確率の計算

入力記号列 W に対する特定の構文解析木 T の生起確率を計算する方法について説明する．

4.1 文法規則による積の場合

生起確率を求めたい構文解析木が与えられた場合，構文解析木に適用されている文法規則の生起確率の積を構文解析木の生起確率とするのが単純な方法である (本稿では，この方式で得られる生起確率を P_1 で表わし，確率構文解析法 $M_1 = (G, L, P_1)$ とする)．

$$P_1(T|W) = \prod_{p_i} P(p_i)$$

なお，$p_i = A \rightarrow \beta$ の生起確率を，

$$\therefore\ P(p_i) = p(A \rightarrow \beta) = \frac{N(A \rightarrow \beta)}{\sum_{\alpha \in N} N(\alpha \rightarrow \beta)}$$

で表わし，$N(A \rightarrow \beta)$ を $A \rightarrow \beta$ の出現個数とする．

図1の構文解析木の生起確率は，次のように表わされる．

$$P_1(T|DGHF) = P(A \rightarrow BC) \cdot P(B \rightarrow DE) \cdot P(C \rightarrow F) \cdot P(E \rightarrow GH)$$

4.2 LC 文法規則リストによる場合

前節の方法は，構文解析木の高さ方向に関する情報がないため，文法が曖昧でない文法であれば，良い結果が得られるが，そうでなければ木の高さが高くなるほど誤差が大きくなる．そこで，葉の LC 親の文法記号と，1つ前の葉から得られる LC 文法規則リストの組の出現頻度を数え，そこから計算される葉ごとの生起確率の積を構文解析の生起確率とする．

ここで，構文解析木の葉の左から i 番目を W_i とし，その W_i の LC 親を $LCP(W_i)$ とする．また，$i-1$ 番目に現れる LC 文法規則リストを L_{i-1} と表わす (図2)．

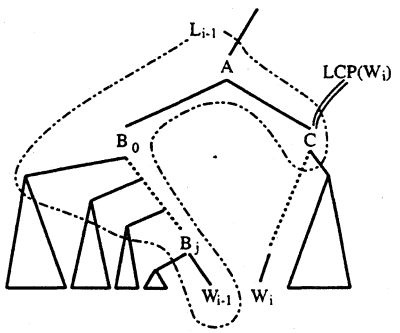


図2: LC 親と LC 文法規則リスト

この時，W_i に対する L_{i-1} の生起確率は，次のように定義される．

$$P(L_{i-1}|LCP(W_i)) = \frac{N(L_{i-1}, LCP(W_i))}{N(LCP(W_i))}$$

- $N(L_{i-1}, LCP(W_i))$: $LCP(W_i)$ と L_{i-1} が現れる出現個数
- $N(LCP(W_i))$: W_i の LC 親 $LCP(W_i)$ の出現個数

しかし，LC 文法規則リスト L_{i-1} は，木の根から葉の方向への文法規則のリストであるので，木の高さに依存し，学習データに依存する．そのため，LC 文法規則リスト L_{i-1} の種類は，木の高さに関

して指数的に増大する．そこで LC 文法規則リスト L_{i-1} を連続する 2 つの文法規則に分割する．そして，それぞれの 2 つの連続する文法規則と LC 親の組み合わせの生起確率の積を先の L_{i-1} の生起確率として近似する．

$$P(L_{i-1}|LCP(W_i)) \approx \prod_{j=1}^{|L_{i-1}|} P(p(L_{i-1},j)|LCP(W_i)) \cdot P(p(L_{i-1},j),p(L_{i-1},j+1),LCP(W_i))$$

上記の近似式中で使われている生起確率や記号を次に定義する．

- $p(L_i,j)$: L_i の先頭から j 番目に対応する文法規則
- $P(p(L_{i-1},j)|\alpha_i) = LC$ 親が α_i であるとの条件のもとで文法規則が "$p(L_{i-1},j)$" である確率
- $\dfrac{P(p(L_{i-1},j+1),p(L_{i-1},j+1),\alpha_i) = (p(L_{i-1},j),p(L_{i-1},j+1),\alpha_i) \text{ の出現個数}}{\sum_{\beta\in(TUN)}(p(L_{i-1},j),p(L_{i-1},j+1),\beta) \text{ の出現個数}}$

以上の LC 文法規則リストと LC 文法規則リストを分割した場合の入力文字列 W に対する構文解析木の生起確率の計算式を次式で定義する (本稿では，この方式で得られる生起確率を P_2 で表わし，確率構文解析法 $M_2 = (G,L,P_2)$ とする)．

$$P_2(T|W) = \prod_{i}^{|W|} P(L_i|LCP(W_i)) = \prod_{i}^{|W|}\prod_{j=0}^{|L_{i-1}|} P(p(L_{i-1},j)|LCP(W_i)) \cdot P(p(L_{i-1},j),p(L_{i-1},j+1),LCP(W_i))$$

図 1 の葉に当たる文字列を $W = DGHF$ とすると，その時に構文解析木 T が得られる計算式は，次の通りになる．

$$P_2(T|DGHF) = P(B \to DE|E) \cdot P(E \to GH|H) \cdot P(E \to GH|C) \cdot P(B \to DE|C) \cdot P(E \to GH, B \to DE, C) \cdot P(A \to BC|C) \cdot P(B \to DE, A \to BC, C) \cdot P(F \to C|\$)$$

4.3 LC リストの親から位置を用いる方法

前節の方法では，構文解析木の高さ方向に関する情報として，LC 文法規則リストの連続して適用される文法規則の確率を用いていた．本節では，それに加えて適用される文法規則の LC 文法規則リストにおける出現順を追加することによって，構文解析木の生起確率を求める (本稿では，この方式で得られる生起確率を P_3 で表わし，確率構文解析法 $M_3 = (G,L,P_3)$ とする)．

$$P_3(T|W) = \prod_{i}^{|W|} P(L_i|LCP(W_i)) = \prod_{i}^{|W|}\prod_{j=1}^{|L_{i-1}|} P(p(L_{i-1},j),p(L_{i-1},j+1),j,LCP(W_i))$$

$$P(p(L_{i-1},j),p(L_{i-1},j+1),j,\alpha) = \frac{N(p(L_{i-1},j),p(L_{i-1},j+1),j,\alpha)}{N(p(L_{i-1},j),j,\alpha)}$$

- $N(p(L_{i-1},j),j,\alpha)$: (LC 親が α，LC 文法規則リスト L_{i-1} の上から j 番目が文法規則 $p(L_{i-1},j)$，j+1 番目が文法規則 $p(L_{i-1},j+1)$) である組の出現個数
- $N(p(L_{i-1},j),j,\alpha)$: (LC 親が α，LC 文法規則リスト L_{i-1} の上から j 番目が文法規則 $p(L_{i-1},j)$) である組の出現個数

前節と同様に，葉に当たる文字列 $W = DGHF$ とすると，その時に構文解析木 T が得られる計算式は，次の通りになる．

$$P_3(T|DGHF) = P(B \to DE, E \to GH, 1, C) \cdot P(E \to GH, *, 2, C) \cdot P(A \to BC, C \to F, 1, *) \cdot P(F \to C, *, 2, *)$$

なお，$*$ には文法記号の 1 文字が対応する．

5 実験結果

日本電子化辞書研究所 (EDR) の日本語コーパスのうち，構文解析済みデータ 9856 文を学習データとして，実験を行なった．本稿では，図 3 に示す **<複合名詞句>** が 3 個並んだ曖昧性を含む構文解析木の部分木 T_a と T_b の結果を示す．なお，文法規則 "< 複合名詞 >→< 複合名詞 > < 複合名詞 >"，部分木 T_a，部分木 T_b の出現回数は，それぞれ 114 回，11 回，1 回であった．また，文法規則 "< 複合名詞 >→ < 複合名詞 > < 複合名詞 >" を p_i に，文法記号 **<複合名詞句>** を **<複>** で省略する．

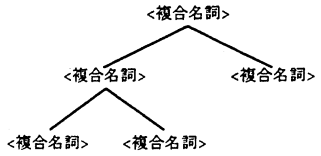
学習データから得られる T_a と T_b の生起確率は，

$$P^*(T_a| < 複 >< 複 >< 複 >) = \frac{11}{12} = 0.9167$$
$$P^*(T_b| < 複 >< 複 >< 複 >) = \frac{1}{12} = 0.083$$

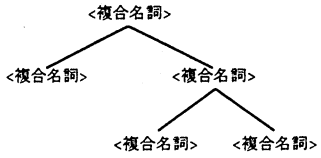
となる．これに対して，**第 4.1 節**の方式では，

$$P_1(T_a| < 複 >< 複 >< 複 >) = P(p_i) \cdot P(p_i) \cdot P(p_i) = \frac{121}{121} \cdot \frac{121}{121} \cdot \frac{121}{121} = 1 = P_1(T_b| < 複 >< 複 >< 複 >)$$

と計算される．



(a)最左導出T_a



(b)最右導出T_b

図 3: 曖昧な構文解析木の部分木の例

第 4.2 節の方式では,

$$P_2(T_a|<\text{複}><\text{複}><\text{複}>) = P(p_i|<\text{複}>)\cdot P(p_i|<\text{複}>)\cdot P(p_i,p_i|<\text{複}>)\cdot P(p_i|*) = 1\cdot 1\cdot \frac{11}{12}\cdot 1 = 0.917$$

$$P_2(T_b|<\text{複}><\text{複}><\text{複}>) = P(p_i|<\text{複}>)\cdot P(p_i|<\text{複}>)\cdot P(p_i|*)\cdot P(p_i,p_i|*) = 1\cdot 1\cdot 1\cdot \frac{12}{12} = 1$$

と計算される. T_b の最後の葉に関して計算に誤差が生じてしまう問題点があるため, 誤差が生じている.

第 4.3 節の方式では,

$$P_3(T_a|<\text{複}><\text{複}><\text{複}>) = P(p_i,*,1,<\text{複}>)\cdot P(p_i,*,1,*) = \frac{11}{12}\cdot\frac{12}{12}$$

$$P_3(T_b|<\text{複}><\text{複}><\text{複}>) = P(p_i,p_i,1,*)\cdot P(p_i,*,2,*) = \frac{1}{12}\cdot 1 = 0.833$$

と計算される.

また, 図 4 は**<読点>**が並び構文解析木の部分木である. この部分木はコーパスに現れないので, 生起確率は 0 となる. 以下に 3 つの方式の生起確率を示す. なお, 文法規則 "< 名詞句 >→< 名詞句 >< 読点 >" を p_j で省略する.

$$P_1(T_c|<\text{名詞句}><\text{読点}><\text{読点}>) = P(p_j)\cdot P(p_j) = \frac{68}{68}\cdot\frac{68}{68} = 1$$

$$P_2(T_c|<\text{名詞句}><\text{読点}><\text{読点}>) = P(p_j|<\text{読点}>)\cdot P(p_j,p_j|<\text{読点}>)\cdot P(p_j|*) = 1\cdot 0\cdot 1 = 0$$

$$P_3(T_c|<\text{名詞句}><\text{読点}><\text{読点}>) = P(p_j,*,1,<\text{読点}>)\cdot P(p_j,*,1,*)$$

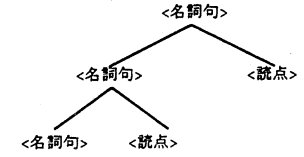


図 4: コーパスに現れない部分木

$$= 0\cdot 1 = 0$$

6 おわりに

本研究では, 構文解析木の高さ方向に関する情報として LC 文法規則リストを用いることで, 構文解析木の生起確率を計算する方法を提案した. 例えば, 従来の方法では, あり得ない構文解析木の生起確率を文法規則の適用順序を計算することによって正しく 0 と求めている.

今後は, EDR のコーパス以外にも学習データを増やすとともに, 特定の部分木の生起確率だけではなく, 可能な構文解析木の生起確率についても評価を行なう予定である. また, 学習結果の記憶容量や削減や確率構文解析法のモデルの評価方法についても改善をしたい.

なお, 本研究の一部は, 科学研究費補助金 奨励研究 (A) 課題番号 12780248 の援助を得て行なった.

参考文献

[1] M. Tomita, "An Efficient Augmented Context-Free Parsing Algorithm", Computational Linguistics, 13,1-2, pp31-46, 1987.

[2] T. Briscoe and J. Carroll, "Generalized Probabilistc LR Parsing of Natural Language(Corpora) with Unification-Based Grammars", Computational Linguistics, Vol.19, No.1, pp.25-59, 1993.

[3] E. Charniak, "Statistical langugae learning", MIT press, 1993.

[4] K. Inui, V. Sornlertlamvanich, H. Tanaka and T. Tokunaga, "Probabilistic GLR Parsing: A New Formalization and Its Impact of Parsing Performance", Journal of Natural Language Processing, Vol. 5, No. 3, pp.33-52, 1998.

[5] H. Bunt and A. Nijholt, "Advances in probabilistic and other parsing technologies", Kluwer Academic Publishers, 2000.