複合化確率文脈自由文法の提案とその評価

横林 由理枝　冨浦 洋一　日高 達
(九州大学大学院システム情報科学研究科)

1 はじめに

文脈自由文法 (CFG) を用いた構文解析では，一般に入力文に対し意味的に誤った構文木を含む複数の構文木が導出される．そのため解析結果をそのまま意味解析や翻訳等の後処理に渡すと処理の質が悪くなる．そこで，CFG の生成規則にその規則の適用確率を付与することにより，構文木の生起確率を与えることが出来るようにした確率文脈自由文法 (PCFG) を導入することが考えられる．生成規則の適用確率は，意味的に正しい構文木から成る学習データの統計的性質を反映するように推定され，このようにして構成した PCFG を用いて，入力文に対する可能な構文木とその生起確率を求め，この確率値により候補を絞り込むことにより，後処理の質が上がることが期待できる [2].

従来の PCFG では標本列 (構文木列) を単一の発生源から収集されたものとしていた．しかし，実際の言語データが一つの発生源から得られたものか否かわからず，さらに，自然言語が PCFG で完全に表せるとも限らない [1]. そこで本研究では，実際の言語現象に柔軟に対応できるように複数の PCFG をもつ複合化確率文脈自由文法 (以下，複合化文法という) を定義し，その文法数の決定法と生成規則の適用確率の推定法を提案する．また，提案した手法により作成した複合化文法を用いて文法数と標本列の生起確率との関係を調べる実験と認識実験を行なった．

2 複合化文法とパラメタ推定

【 定義 1 】 複合化文法 $G^{(M)}$を，PCFG $G_t = \langle G, p_t \rangle$ と選択確率 c_tとの 2 つ組によって定義する．

$$G^{(M)} = \langle\langle\langle G, p_1\rangle, c_1\rangle, \cdots, \langle\langle G, p_M\rangle, c_M\rangle\rangle$$

ここで，Mは複合化度とする．

また，それぞれの PCFG G_t は $\langle\langle \Sigma, V, P, X_1\rangle, p_t\rangle$ で定義される．ただし，

$$\begin{cases} \Sigma & : \text{終端記号の有限集合} \\ V & : \text{非終端記号の有限集合} \\ P & : \text{生成規則の有限集合} \\ & \quad (P = \{\delta_{ij} \mid \delta_{ij} = X_i \to \alpha_j : i = 1, \cdots, n, \\ & \qquad\qquad j = 1, \cdots, m_i\}) \\ X_1 & : \text{開始記号 } (X_1 \in V) \\ p_t & : P \to (0, 1] \end{cases}$$

である [2].

また，選択確率 $c_1, \cdots, c_M$は次の条件が必要である．

$$\sum_{t=1}^{M} c_t = 1 \qquad 0 < c_t \leq 1 \quad (t = 1, \cdots, M)$$

それぞれの文法 G_tにおける p_t $(t = 1, \cdots, M)$ と G_t の選択確率 c_t $(t = 1, \cdots, M)$ の列を次のようなベクトルで表す．

$$\begin{cases} \boldsymbol{p}^{(M)} & = & p_1^{(M)}, p_2^{(M)}, \cdots, p_M^{(M)} \\ \boldsymbol{c}^{(M)} & = & c_1^{(M)}, c_2^{(M)}, \cdots, c_M^{(M)} \end{cases}$$

□

【 定義 2 】 複合化文法における構文木 T_kの生起確率 $\Pr(T_k \mid \boldsymbol{p}^{(M)})$ は以下のように定義される．

$$\Pr(T_k \mid \boldsymbol{p}^{(M)}) = \sum_{t=1}^{M} c_t^{(M)} \Pr(T_k \mid p_t^{(M)})$$

ただし，$G_t^{(M)}$は複合化文法 $G^{(M)}$の t 番目の PCFG であり，$\Pr(T_k \mid p_t^{(M)})$ $(t = 1, \cdots, M)$ は PCFG $G_t^{(M)}$ における構文木 T_kの生起確率とする．

□

複合化文法における尤度 (構文木列の生起確率) は次式から求められる．

$$L_M(\boldsymbol{p}^{(M)}, \boldsymbol{c}^{(M)}) = \prod_{k=1}^{N} \Pr(T_k \mid \boldsymbol{p}^{(M)})$$

この尤度が最大になるようなパラメタ $\boldsymbol{p}^{(M)}, \boldsymbol{c}^{(M)}$ を Baum 理論 [3] を用いて推定する．Baum 理論とは，変数$\boldsymbol{x}$の値を更新することによって，正係数の多項式 $f(\boldsymbol{x})$

を極大，あるいは極小にするxの値を求める手法である．複合化文法における尤度は一般に複数の極大値が存在するため，Baum 理論を用いた推定ではパラメタの初期値によって尤度がどの値に収束するかが左右される．

3 パラメタの初期値設定法

ここでは，パラメタの初期値設定法の考え方を示す．

ここで，Baum 理論において初期値となるものを $p_{t\ init}^{(M)}(\delta_{ij})$, $c_{t\ init}^{(M)}$，収束計算を行った後の値を $p_{t\ \infty}^{(M)}(\delta_{ij})$, $c_{t\ \infty}^{(M)}$ とする $(\delta_{ij} = X_i \to \alpha_j)$.

ある適当な $p : P \to (0, 1]$ に対して

$$\begin{cases} \boldsymbol{p}_{init}^{(M+1)} &= \boldsymbol{p}^{(M)} \ , \ p \\ \boldsymbol{c}_{init}^{(M+1)} &= \boldsymbol{c}^{(M)} \ , \ 0 \end{cases}$$

とすると，

$$\begin{cases} p_{t\ \infty}^{(M+1)} &= p_t^{(M)} \quad (1 \le t \le M) \\ c_{t\ \infty}^{(M+1)} &= c_t^{(M)} \quad (1 \le t \le M) \end{cases}$$

となり，この場合，複合化度 Mと $M+1$の文法による尤度は等しくなる．そこで，上記の $\boldsymbol{p}_{init}^{(M+1)}, \boldsymbol{c}_{init}^{(M+1)}$ の値を少し変えたもの (pの与え方としては，$\boldsymbol{p}^{(M)}$ の要素の中でどれとも傾向がちがうものに設定し，$c_{M+1\ init}^{(M+1)}$ も 0でない値にする) を初期値として与えてやると，複合化度 Mと $M+1$の文法による尤度は少なくとも異なることが期待できる．そこで，複合化度 $M+1$の文法による尤度が複合化度 Mの文法による尤度より大きくなるように，いろいろな変え方で試す．

具体的には複合化の手順を次のように考える．

まず，$c_1 = 1, p_1^1(\delta_{ij})$ を最尤推定法により求めた値に設定する．

次に，複合化文法 $G^{(M)}$から $G^{(M+1)}$を新たにつくる手法を次のように考える．

ここで，$G_{t_0}^{(M)}$において，左辺 X_iが同じ生成規則の適用確率 $p_{t_0}^{(M)}(\delta_{i1}), \cdots, p_{t_0}^{(M)}(\delta_{i\ m_i})$ の中の最大値，最小値をそれぞれ

$$\begin{cases} Y_{t_0}^{(M)}(X_i) &= Max(p_{t_0}^{(M)}(\delta_{i1}), \cdots, p_{t_0}^{(M)}(\delta_{i\ m_i})) \\ Z_{t_0}^{(M)}(X_i) &= Min(p_{t_0}^{(M)}(\delta_{i1}), \cdots, p_{t_0}^{(M)}(\delta_{i\ m_i})) \end{cases}$$

と表すことにする．このとき，複合化度 $M+1$の文法のパラメタの初期値を次のように与える．

$$\begin{aligned} p_{t\ init}^{(M+1)}(\delta_{ij}) &= p_t^{(M)}(\delta_{ij}) \quad (1 \le t \le M) \\ p_{M+1\ init}^{(M+1)}(\delta_{ij}) &= \frac{\{Y_{t_0}^{(M)}(X_i) + Z_{t_0}^{(M)}(X_i)\} - p_{t_0}^{(M)}(\delta_{ij})}{m_i \cdot \{Y_{t_0}^{(M)}(X_i) + Z_{t_0}^{(M)}(X_i)\} - 1} \\ c_{M+1\ init}^{(M+1)} &= \frac{1}{2} \cdot c_{t_0}^{(M)} \\ c_{t_0\ init}^{(M+1)} &= \frac{1}{2} \cdot c_{t_0}^{(M)} \\ c_{t\ init}^{(M+1)} &= c_t^{(M)} \quad (1 \le t \le M, t \ne t_0) \end{aligned}$$

上式で $t_0 = 1, \cdots, M$ においてすべての t_0に対して $\boldsymbol{p}_{init}^{(M+1)}, \boldsymbol{c}_{init}^{(M+1)}$ を計算し，それをパラメタの初期値として Baum 理論を用いて収束計算を行ない，それを用いて M通りの尤度 $L_{M+1}(\boldsymbol{p}_{\infty}^{(M+1)}, \boldsymbol{c}_{\infty}^{(M+1)})$ を求める．その中で尤度の値が最も大きくなるような $\boldsymbol{p}_{\infty}^{(M+1)}, \boldsymbol{c}_{\infty}^{(M+1)}$ をパラメタ $\boldsymbol{p}^{(M+1)}, \boldsymbol{c}^{(M+1)}$ として選ぶ．

以上の繰り返し計算をϵが十分小さな正数とすると，

$$L_{M+1}(\boldsymbol{p}^{(M+1)}, \boldsymbol{c}^{(M+1)}) - L_M(\boldsymbol{p}^{(M)}, \boldsymbol{c}^{(M)}) \le \epsilon$$

となるまで行う (このとき，複合化度は Mとなる)．

4 実験

実験には曖昧な CFG から計算機によってランダムに発生させた構文木を用いた．それを学習データとして本手法で複合化文法を作成し，複合化度と尤度との関係を調べる実験と，学習データに対する認識実験を行なった．実験結果から，複合化文法による尤度は複合化度が高くなるにつれて単調増加することが確認できた．また，複合化度を上げたときの尤度の上昇率が高いとき認識率も高くなることが多くの場合に見られた．

5 最後に

本研究では，複数の確率文脈自由文法をもつ複合化確率文脈自由文法を定義し，その複合化度 (文法数) の決定法とパラメタ推定における初期値設定法を提案し，その有効性を確認した．

今後の課題として，コーパスなどの実際のデータに対して実験を行ない，構文構造の曖昧性がどのくらい解消されるかを確かめる．

参考文献

[1] 林田 憲昭:確率文脈自由文法の複合化, 九州大学大学院修士論文 (1997)

[2] 日高 達:確率文法, 情報処理学会誌 Vol.36 No.2(1995)

[3] Baum,L.E.:An Inequality and Associated Maximization Technique in Satistical Estimation of Probabilistic Functions of a Markov Process, Inequalities,3(1972)