確率文脈自由文法に基づく漸進的構文解析

加藤 芳秀　　松原 茂樹　　外山 勝彦　　稲垣 康善　(名古屋大学)

1　はじめに

漸進的構文解析は，自然言語文をその単語の出現順序に従って解析し，入力途中の段階でその構文的関係を捉えるための手法であり，実時間対話処理システムに不可欠な要素技術の一つである．しかし，漸進的構文解析では，入力文全体に対する構文木の部分を構成しない結果的に誤りとなる構文木を生成するといった問題がある．本稿では，入力途中で生成された各構文木に対して，それが正しくなるような単語列がそれ以降に入力される確率を確率文脈自由文法 (Probabilistic Context Free Grammar, PCFG) に基づいて計算する手法を提案する．本手法は，正しい木の選択や不用な木の枝刈りに利用できる．

2　漸進的チャート解析

漸進的チャート解析 [1] は，文脈自由文法に基づく漸進的構文解析の一手法であり，単語が入力されるたびにそれまでの入力に対する部分的な構文木 (部分構文木) を生成する．部分構文木には構造が定まっていない箇所が存在するが，その箇所を新たに入力された単語から生成される構文木で置き換えることにより，新たな部分構文木を生成する．

以下での議論を明確にするため，入力文全体に対して正しい部分構文木の定義を与える．部分構文木 σ に対して置き換え操作を有限回適用することにより部分構文木 σ' が得られるとき，σ は σ' を包含するという．入力文 $w_1 \cdots w_n$ に対するある構文木を σ が包含するとき，σ は $w_1 \cdots w_n$ に対して妥当であるという．

3　PCFG に基づく部分構文木の評価

部分構文木の妥当性は残りの入力に依存して定まるが，本手法では，各部分構文木に対して，残りの入力がそれを妥当にする単語列である確率を PCFG に基づき計算する．その確率が高い部分構文木は，入力文全体に対して妥当である可能性が高い．

PCFG は構文木，及び文の生起確率を計算するためのモデルである．各文法規則に対して確率が与えられており，構文木の生起確率は，その木の構成に用いられた規則の確率の積で計算する．文の生起確率は，その文に対する構文木の生起確率の和で計算する．

まず，j 番目の単語 w_j まで入力されたとき，それまでに生成された部分構文木のそれぞれに関して，それを妥当にするような残りの入力 ω がどのような単語列であるかについて考える．以下では，$T(i)(i \leq j)$ を $w_1 \cdots w_i$ に対する部分構文木からなる集合とする．部分構文木 $\sigma' \in T(j)$ が $w_1 \cdots w_j\omega$ に対して妥当であるのは，σ' の未決定範疇列 (σ' 中の構造が定まっていない箇所の範疇を左から順に並べた系列) から ω が導出されるときである．一方，部分構文木 $\sigma \in T(i)$ が $w_1 \cdots w_j\omega$ に対して妥当であるのは，σ に包含されるある部分構文木 $\sigma' \in T(j)$ の未決定範疇列から ω が導出されるときである．すなわち，σ を妥当にする残りの入力単語列の集合 $\Omega(\sigma, j)$ を，次のように定義できる．

$$\Omega(\sigma, j) = \{\omega \mid \sigma\text{に包含される}\sigma' \in T(j)\text{が存在し，}\sigma'\text{の未決定範疇列から}\omega\text{が導出される}\}$$

$w_1 \cdots w_j$ に続く入力が σ を妥当にする単語列となる，すなわち，$\Omega(\sigma, j)$ の要素となる確率は，次式により定義できる．

$$\sum_{\omega \in \Omega(\sigma,j)} P(\omega \mid w_1 \cdots w_j) = \frac{\sum_{\omega \in \Omega(\sigma,j)} P(w_1 \cdots w_j\omega)}{P(w_1 \cdots w_j)} \tag{1}$$

(1) の右辺の分母は，次のように $w_1 \cdots w_j$ に対する部分構文木から計算できる．

$$P(w_1 \cdots w_j) = \sum_{\sigma' \in T(j)} P(\sigma') \tag{2}$$

一方，(1) の右辺の分子を計算するには，$\Omega(\sigma, j)$ の要素を数えあげる必要があるが，それは一般にはできない．そこで，右辺の分子を次のように近似する．

$$\sum_{\omega \in \Omega(\sigma,j)} P(w_1 \cdots w_j\omega) \approx \sum_{\sigma' \in Tu(\sigma,j)} P(\sigma') \tag{3}$$

ここで，$Tu(\sigma, j)$ は σ に包含される $w_1 \cdots w_j$ に対する部分構文木，及び，それらと未決定範疇列が一致する $w_1 \cdots w_j$ に対する部分構文木からなる集合である．$Tu(\sigma, j)$ 中の部分構文木に包含される構文木はいずれも，$w_1 \cdots w_j\omega(\omega \in \Omega(\sigma, j))$ に対する構文木であり，その生起確率の和 ((3) 式右辺) は，$w_1 \cdots w_j\omega$ に対する構文木の生起確率の和 ((3) 式左辺) より小さい．(1) 〜 (3) より，

$$\sum_{\omega \in \Omega(\sigma,j)} P(\omega \mid w_1 \cdots w_j) \approx \frac{\sum_{\sigma' \in Tu_(\sigma,j)} P(\sigma')}{\sum_{\sigma' \in T(j)} P(\sigma')} \tag{4}$$

(4) は $T(j)$ 中の部分構文木の生起確率から計算できる，すなわち，単語 w_j が入力された時点で計算できる．

4　おわりに

本稿では，漸進的構文解析において，入力途中の段階で，部分構文木が妥当となる可能性を PCFG を用いて評価する手法を提案した．本手法により，解析の正確さの向上が期待できる．漸進的構文解析では，部分構文木を次のモジュールに出力するタイミングが重要であるが，本評価法をもとにそのタイミングを決定する方法について，今後検討したい．

参考文献

[1] Matsubara, S. et al.: Chart-based Parsing and Transfer in Incremental Spoken Language Translation, *Proceedings of NLPRS-97*, pp.521–524(1997).