漸進的構文解析のための統計情報を用いた有限状態変換器作成手法

湊 恵一　　加藤 芳秀　　松原 茂樹　　稲垣 康善　　(名古屋大学)

1 はじめに

同時通訳システムなど, 入力発話と同時的な処理が要求される場面では, 解析処理の漸進性, 実時間性が重要となる. これまで, 文脈自由文法 (CFG) に基づく漸進的構文解析手法が提案されている [1] が, この手法は, 実時間性については十分ではない. 一方, 著者らは有限状態変換器 (FST) に基づく漸進的構文解析手法を提案している [2]. この手法では, CFG を近似変換した FST を用いて構文解析を実行するため, 高速な解析処理が可能となる. しかし, 近似変換の結果, もとの CFG で解析できる文が, FST では解析できない場合がある. 本稿では, 統計情報を利用した CFG から FST への変換について提案する. FST を構成する弧は構文木の節点と対応しているが, 出現頻度の高い節点に対応した弧を優先的に作成することにより, より多くの文を解析できる FST を作成できる.

2 FST を用いた漸進的構文解析

FST は弧に出力ラベルを持つ有限状態オートマトンであり, 入力ラベルに終端記号を, 出力ラベルに構文木を割り当てることで, 構文解析を実現できる. 文献 [2] の手法では, CFG から漸進的構文解析を実現する FST を作成する. まず, 各範疇に対応した状態遷移ネットワークを作成する. 次に, 開始記号を入力ラベルとする弧を持つ FST を初期 FST とし, 弧を, その入力ラベルの範疇に対応したネットワークで置き換える操作を繰り返す. 図 1の左の図は, NP を入力ラベルに持つ弧を, 左辺が NP である文法規則を表現するネットワークで置き換える操作を示している.

FST は, 入力に従って状態を遷移し, その出力を連結するという単純な処理で構文解析を実現するため, 高速な処理が可能である. しかし, 一般的にネットワークの置き換え操作は無限に適用することになるが, 実際には記憶領域の制限があり, 実行可能な置き換え操作は有限回である. そのため, 文解析に十分な回数の置き換えがされない場合, もとの CFG で解析できる文であっても, FST では解析できない.

3 統計情報に基づく CFG から FST への変換

FST の弧は構文木の節点と対応づけられる. 例えば, 図 1は弧と節点の対応を示しており, 同じ番号の弧と節点が対応している. 従って, 出現頻度の高い節点に対応する弧に優先的に置き換え操作を適用して作成した FST は, より多くの文を解析できることが期待できる. ところで, 節点 X_n に対応した弧を遷移することは, 構文木の根から節点 X_n へのパスに現れる文法規則 $r_1, r_2, \ldots, r_n$ を用いた解析が行われることを意味する (図 2参照). そこで, これらの文法規則が適用されて X_n が導出される確率 $P(r_1, r_2, \ldots, r_n)$ を, 弧の優先度とし, 優先度の高い弧から順に置き換え操作を適用する. $P(r_1, r_2, \ldots, r_n)$ を次のように定義する.

$$P(r_1, r_2, \ldots, r_n) = \prod_{i=1}^{n} P(r_i \mid c_i) \qquad (1)$$

ここで, c_i は, 文法規則を適用する条件であり, 文脈情報を用いる. 文脈情報として, 文法規則が適用される節点の周囲に存在する節点の情報を利用する.

4 解析実験

統計情報を用いた FST 作成の効果を検討するために, 解析実験を行った. その

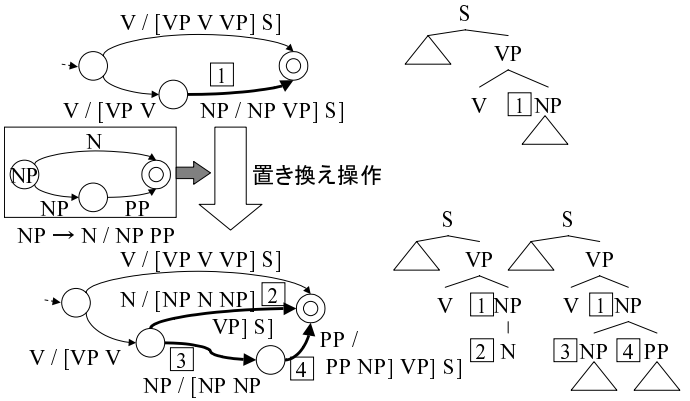

図 1: 置き換え操作例・弧と節点の対応

表 1: 実験結果

文脈情報	なし	C_0	C_1	C_2
正解率	36%	57%	64%	68%

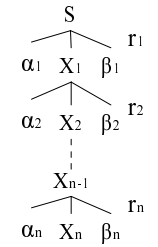

図 2: X_n の導出

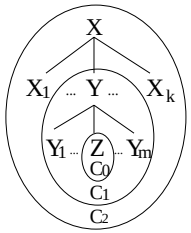

図 3: 文脈の範囲

際, 条件付き確率の条件として, 3 通りの文脈情報 C_0, C_1, C_2 を用いた. 節点 Z に文法規則を適用する場合に, それぞれの文脈情報を用いるときに参照する節点の範囲を図 3に円で示す. 実験では, 構文木付き ATR 音声言語データベースを用いた. 学習データとして 9081 文を用い, 文法規則 (698 種類) とその統計情報を獲得した. 優先度を用いない場合と, 各文脈情報に基づいて計算した優先度を用いた場合について, それぞれ FST を作成した. ただし, 状態数が 1,000,000 に達するまで, 置き換え操作を適用した. 学習データに含まれない 1874 文を用いて解析実験を行い, 正解率を求めた. 解析は幅優先探索で行い, 正解率は, 文全体に対して得られた解析結果の中に, 正解の構文木が存在した文の割合とした. 実験結果を表 1に示す. 文脈情報なしは, 優先度を用いなかった場合の結果である. 実験結果より, 優先度に基づく FST の作成の有効性を確認した. また, 利用する文脈情報が広範囲であるほど, より正解率の高い FST が作成できることがわかった.

5 おわりに

本稿では, 構文解析に用いる FST 作成時の, 統計情報の利用に関する有効性について述べた. 今後の課題には, 条件付き確率の条件に関する検討があげられる.

参考文献

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

[2] 湊 他, "有限状態変換器を用いた漸進的構文解析", 平成 13 年度電気関係学会東海支部連合大会論文集,p.279 (2001).