文脈自由文法の変換に基づく漸進的な話し言葉翻訳手法

松原 茂樹† 小川 浩司‡ 外山 勝彦‡§ 稲垣 康善‡

†名古屋大学言語文化部 ‡名古屋大学大学院工学研究科

§名古屋大学統合音響情報研究拠点 (CIAIR)

matubara@lang.nagoya-u.ac.jp

1 はじめに

構文トランスファー方式に基づく機械翻訳ではまず，原言語の解析により構文構造を作成するが，この段階で十分に曖昧性を解消できなければ，不適切な構造に対して変換・生成処理が実行され，システムは誤った翻訳結果を作り上げることになる．特に，原言語入力に対してできる限り同時進行的に目標言語を生成する漸進的翻訳では，通常の曖昧性に加え，それ以後の入力の予測に関する曖昧性が生じるため，問題は一層深刻となる．実際，著者らが提案している漸進的英日話し言葉翻訳システム Sync/Trans[3] は，文脈自由文法 (以下， CFG) を用いた漸進的解析をベースとしているが，入力途中であまりに多くの解析木が作成されることがある．解析処理時間が大きくなるとともに，不適切な解析木が作成され，翻訳に失敗する可能性が高くなるという問題がある．

一方， CFG を用いた漸進的解析では，入力予測に関する曖昧性を減らすために， M 標準形に等価変換された CFG の利用が効果的であることが実験的に示されている [2]．解析木の中の予測に関する部分を抽象化し，いくつかの解析木を同一化できるようにあらかじめ CFG を変形することにより，作成される解析木の数を抑制することができる．

そこで本稿では， Sync/Trans の解析フェーズにおいて，M 標準形に準拠した CFG を用いることにより，翻訳処理時間を削減，並びに翻訳精度を向上させる手法を提案する．また本手法の有効性を評価するために，英語対話文を用いた翻訳実験を実施したので，その結果についても報告する．

2 漸進的英日話し言葉翻訳システム

漸進的な英日話し言葉翻訳システム Sync/Trans は，英語音声の入力に対して同時的に対応する日本語を作成する [3]．システムは，語が入力されるたびにそれまでの入力に対する解析木を作成する**漸進的解析**と解析木から対応する日本語を作り上げる**漸進的変換**から構成される [4]．漸進的解析は， CFG に対する構文解析手法の一つであるチャート法をベースとしており，その時点で可能なすべての解析木を作成する．また，漸進的変換では，作成された解析木の中から一つを選択し，それに対して，文法規則と一対一に対応して作成された変換規則を適用することにより，漸進的に日本語翻訳結果を作り上げる．それ以後の入力によっては，選択した解析木が不適切であったと判明することがあるが，その場合には，再度解析木を選択し，誤った翻訳結果については言い直し表現を用いて修正を試みる．対話翻訳実験を通して，システムの有用性をすでに確認しているが，一方で，あまりに多くの言い直しを生成すると，翻訳結果の了解度が下がり，翻訳精度が低下することが分かっている．

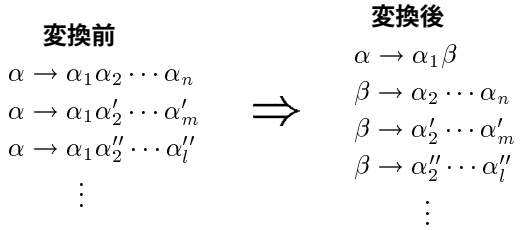

図 1: CFG の M 標準形への変換

3 文脈自由文法の M 標準形

本節では， CFG の M 標準形への変換，及び，それを用いた漸進的解析について述べる．

漸進的解析は，通常のチャート法と同様，文法規則の適用操作，及び，項の置き換え操作を駆使して解析木を作成する [4]．文法規則の適用はボトムアップであり，範疇が α_1 である解析木 σ，及び文法規則 $\alpha \to \alpha_1\alpha_2\cdots\alpha_n$ に対して，解析木 $[\sigma[?]_{\alpha_2}\cdots[?]_{\alpha_n}]_\alpha$ を作成する．ここで $[?]_{\alpha_i}$ は未決定項とよばれ， $[\sigma[?]_{\alpha_2}\cdots[?]_{\alpha_n}]_\alpha$ は、範疇列 $\alpha_2\cdots\alpha_n$ を構成する単語列の入力を予測できることを示している．

漸進的解析では，文の入力途中において予測されうるすべての入力を，それぞれ解析木として表現するために，結果として作成する解析木が多くなる．すなわち，上述した文法規則の適用において，文法規則 $\alpha \to \alpha_1\alpha'_2\cdots\alpha'_m$ が存在すれば，解析木 $[\sigma[?]_{\alpha'_2}\cdots[?]_{\alpha'_n}]_\alpha$ も作成するというように，左辺の範疇が α でかつ右辺の最左範疇が α_1 であるすべての文法規則に対して同様の操作を行うことになる．それらの解析木はすべて未決定項，すなわち，予測部分が異なるだけであり，それらを別々に作成するのは非効率である．

この問題に対して，新たな文法範疇を導入することにより，別々に作成されていた未決定項列を一つの未決定項で表現することが考えられる．すなわち，図 1に示すように，範疇 β と文法規則 $\beta \to \alpha_2\cdots\alpha_n, \beta \to \alpha'_2\cdots\alpha'_m, \cdots$ を導入し，左辺が α でかつ右辺の最左範疇が α_1 である文法規則

Simultaneous Machine Interpretation based on A Normal-Form Conversion of CFG: Shigeki Matsubara, Hiroshi Ogawa, Katsuhiko Toyama and Yasuyoshi Inagaki (Nagoya University)

は $\alpha \rightarrow \alpha_1\beta$ だけであるとすれば，解析木 $[\sigma[?]_\beta]_\alpha$ のみが作成されることになる．これは，各解析木の未決定範疇列 $[?]_{\alpha_2}\cdots[?]_{\alpha_n}$,$[?]_{\alpha'_2}\cdots[?]_{\alpha'_n}$,$\cdots$ を一つの未決定項 $[?]_\beta$ でまとめたことにほかならない．

以下では，左辺の範疇及び右辺の最左範疇がともに一致する文法規則が複数存在しない CFG は M **標準形**であるという．M 標準形の CFG を用いることにより，文法規則の適用回数が減少するため，漸進的チャート解析の効率化が可能になる [2]．なお，任意の CFG を M 標準形に変換することができる．

4 M 標準形に準拠した変換規則

前節で述べたように，漸進的解析で M 標準形の CFG を用いれば作成される解析木の数が抑制されるため，翻訳処理で解析木の選択に誤る可能性が低下する．これにより，翻訳結果における言い直しの生成頻度が減り，結果として翻訳精度が向上することが期待できる．しかし，そのような文法のもとで作成された解析木は，変換前の文法を用いて生成されたときとは異なるため，変換規則もそれに併せてあらかじめ変更する必要がある．

Sync/Trans における変換規則は，文法規則と一対一に対応し，それぞれ目標言語における生成順序や生成する助詞など，生成に関する情報が付け加えられている．このため，M 標準形への変換手続きに従って変換規則を変形すると，生成に関する情報が破壊されることになる．これを避けるために，本研究では，M 標準形の条件を生成の情報を含めたものに強め，そのような M 標準形の変換規則に変換し，それを用いて翻訳結果を作り上げることを試みる．すなわち，左辺の範疇と右辺の最左範疇が等しいだけでなく，その生成に関する情報もまた一致する変換規則に対して，まとめ上げを行う．なお，文法規則と変換規則の 1 対 1 対応を維持するために，漸進的解析では，変換後の変換規則から生成の情報を取り除いた CFG を用いる．

5 実験と評価

本手法の有効性を評価するために， GNU Common Lisp 2.2 を用いて実験システムを作成し，対話翻訳実験を行った．実験では， ATR 対話データベース [1] に収録された旅行申し込みをタスクとする対話の中から，英語話者による発話 278 文 (平均語数 7.2 語) を使用した．その処理のために，辞書及び変換規則をそれぞれ 570 語， 185 規則の規模で作成した．変換規則を M 標準形に変換した結果，新たに 77 個の範疇が導入され， 262 規則に増加した．変換前と変換後とで，翻訳正解率，及び翻訳時間に関する実験を行った．実験結果を表 1に示す．

【(1) 翻訳正解率に関する評価】

実験に用いた英語発話文をその翻訳結果の了解性に従って分類した．表 2に示す評価基準のもと， A) 及び B) に分類された英語文を翻訳正解とした．変換後の規則を使用することにより，変換前の規則を使用した場合と比べて， 30 文 (10.5 %) が新たに翻訳正解と判定された．不自然な翻訳結果であると判断されていた対話文が，言い直しの生成回数が減少したことにより，新たに意味が通じるようになったことがその主な原因である．以上より， Sync/Trans の精度や品質における， CFG の M 標準形への変換の有効性を確認した．

表 1: 英語発話 278 文に対する実験結果の比較

評価項目	変換前	変換後
翻訳正解率 (%)	59.7	70.5
一文あたりの平均言い直し数 (個)	3.25	2.25
一語あたりの平均翻訳時間 (秒)	2.65	0.046
一語あたりの平均解析木数 (個)	522.7	43.6

表 2: 英語発話 278 文の翻訳正解率の比較

	変換前	変換後
評価基準	文数 (割合)	文数 (割合)
A) 翻訳正解 (言い直し無)	0(0.0 %)	35(12.6 %)
B) 翻訳正解 (言い直し有)	166(59.7 %)	161(57.9 %)
C) 不自然な翻訳結果	76(27.3 %)	38(13.6 %)
D) 翻訳誤り	35(12.6 %)	43(15.5 %)
E) 解析失敗	1(0.4 %)	1(0.4 %)
翻訳正解 A)+B)	166(59.7 %)	196(70.5 %)

【(2) 翻訳処理時間に関する評価】

Sun UltraSparcII ワークステーション (メモリ： 512MB, CPU： 248MHz) 上で，翻訳処理時間を計測した．一単語あたりの平均処理時間は， 2.65 秒から 0.046 秒に短縮した．変換は選択された一つの解析木に対して処理を実行しさえすればよいの対して，解析は作成されたすべての解析木に対して処理を行う．すなわち，翻訳処理時間のほとんどは解析に費やされた時間であることから，処理効率においても本手法が有効であることがわかった．

6 おわりに

M 標準形の CFG を漸進的解析に用いることにより，予測部分に関する曖昧性が小さくなり，作成される解析木の数を削減することができる．本稿では，文法規則及び変換規則の M 標準形への変換に基づく漸進的な話し言葉翻訳手法について述べた．また、対話翻訳実験による評価の結果，本手法の有効性を確認した．

参考文献

[1] 江原 他: ATR 対話データベースの内容, Technical Report TR-I-0186, ATR 自動翻訳研究所 (1990).

[2] 松原，村瀬，加藤，外山，稲垣: 文脈自由文法の変換に基づく漸進的構文解析の効率化, 平成 11 年度電気関係学会東海支部連合大会講演論文集 (1999).

[3] Matsubara, S and Inagaki, Y.: Incremental Transfer in English-Japanese Machine Translation, *IEICE Trans. on Infomation and Systems*, Vol.E80-D, No.11, pp.1121–1129 (1997).

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