

チャート法に基づく漸進的な英日機械翻訳手法

岩田芳明 松原茂樹 稲垣康善

名古屋大学工学部情報工学科

{iwata, matu, inagaki}@inagaki.nuie.nagoya-u.ac.jp

1 はじめに

計算機による対話翻訳や同時通訳の実現のため、漸進的に原言語を目標言語に変換する機械翻訳手法の開発が必要である [1]。語順の似通っている言語間であれば、左から右への逐次変換により漸進的な翻訳が可能となる。しかしながら、英語日本語間のような語順が異なる言語間の翻訳に単純に左から右へ変換する手法を用いると、意味の通らない翻訳文が生成される。そのため、語順の異なる言語間における漸進的な翻訳処理手法を新たに考案する必要がある。

本稿では、日本語会話文の特徴を用いた漸進的な英日機械翻訳手法を提案する。本手法では、入力される英語文はチャート法に基づいて左から右へ語単位で解析される。その結果構築される句構造木に、繰り返しや省略、語順の柔軟性といった日本語会話文の特徴を積極的に利用して定められた変換規則を逐次適用することによって、漸進的に日本語翻訳文が生成される。

我々は漸進的な英日機械翻訳のプロトタイプシステムを作成した。それを用いて、中学一年生用の英語教科書の英語文に対する翻訳実験を行ない良好な結果を得た。

2 漸進的な翻訳文生成

日本語会話文では、繰り返しや省略、語順の柔軟性を利用して話し手が情報を追加、訂正する現象が頻繁に現れる [2]。本研究では漸進的な日本語翻訳文の生成を可能とするために、これらの特徴を利用する。本節では、英語文 “He threw the ball.” を例に、漸進的な英日翻訳手法における日本語翻訳文生成の基本アイデアを説明する。

漸進的な翻訳文生成の例を図1に示す。まず、“He threw” までを第1文として「彼は投げた」と翻訳する。次に、残りの英語文に対して「投げ

He	threw	the	ball	.
彼は	投げた、		ボールを	投げた。

図1：漸進的な翻訳文生成

た」を繰り返すことにより、第1文に目的語の情報を追加する文として「ボールを投げた」と出力する。第2文では主語の情報は省略されてしまっているが、これは第1文から容易に推測可能である。翻訳文「彼は投げた、ボールを投げた。」は、文単位での処理による翻訳文「彼はボールを投げた。」に比べると、翻訳の質は低下しているものの、その意味内容は正しく維持されている。また、動詞が発話された時点でこれを翻訳するため、話し手の発話の内容を迅速に伝達できる。

本研究では、このように翻訳文を動詞の後で一旦区切り、残りの翻訳文を繰り返しを利用して生成することにより、語順の異なる英日間の翻訳を漸進的に行う。

3 漸進的な翻訳処理

漸進的な英日翻訳手法は解析部と変換部から構成される。図2に本手法の構成を示す。解析部では入力された語を順次にチャート法に基づいて解析し、句構造木を順次更新していく。変換部では解析部によって構築された句構造木に変換規則を適用することにより、目標言語表現を決定する。

本研究では、このように解析部と変換部とが交互に句構造木を参照、更新することにより、漸進的な翻訳処理を実現する。以下では、“He threw the ball.” において、語 “threw” が入力された時の各処理部の動作について説明する。

3.1 漸進的な解析

解析部では、入力文の解析にチャート法を用いる。上昇型のアプローチと下降型のアプローチを組み合わせることによって、左から右へ語単位で

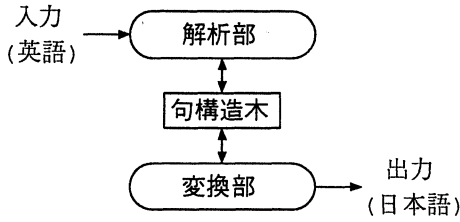


図 2：翻訳処理

解析する。具体的には、語の入力ごとに以下の二つの手続きを実行することにより、漸進的に句構造木を更新する。

- 手続き 1：入力語に対して、上昇型のアプローチを用いて文法を適用し、すべての可能な部分構造木をチャート上に構成する。
- 手続き 2：それまでの入力語によって生成された句構造木と、手続き 1 によって生成された部分構造木から、新しい句構造木を下降的に生成する。

以下、語 “threw” が入力された時の解析例を示すことにより、本手法における解析手順を説明する。“He” に対する翻訳処理が終了した時点での句構造木は

$$[[[!N]_{NP}[?]_{VP}]_S] \quad (1)$$

である。[?] は空所であり、まだ対応する語が入力されていないことを表す。また “!” は “He” が変換部によって置き換えられたもので、“He” に対する変換が終了していることを表す。

解析部は、入力語 “threw” に対して辞書に記述されている項 $[threw]_V$ を初期状態として解析を始める。手続き 1 に従って文法規則 “ $VP \rightarrow V NP$ ” を適用することにより、項 $[[threw]_V[?]_{NP}]_{VP}$ を生成する。これにより表 3 のチャートが構成される。

次に手続き 2 に従って、現在の句構造木 (1) と手続き 1 により構成された表 3 のチャートから新しい句構造木を構築する。(1) の最初の空所 $[?]_{VP}$ は、統語範疇が VP である。したがって、表 3 のチャートから統語範疇が VP である No.2 の項 $[[threw]_V[?]_{NP}]_{VP}$ を選択し、(1) の空所に代入する。

解析部の以上の動作により、

$$[[[!N]_{NP}[[threw]_V[?]_{NP}]_{VP}]_S] \quad (2)$$

No.	term
1	$[threw]_V$
2	$[[threw]_V[?]_{NP}]_{VP}$

表 3：上昇型チャート

t-rule	VP													
source	(11) = V	(12) = NP												
target	(21) = V	(22) = NP												
	phon : (21) (22) “を”(21)													
subseq	<table border="1"> <tbody> <tr> <td>st-rule</td> <td>V</td> </tr> <tr> <td>source</td> <td>(11)</td> </tr> <tr> <td>target</td> <td>(21)</td> </tr> </tbody> </table>	st-rule	V	source	(11)	target	(21)	<table border="1"> <tbody> <tr> <td>st-rule</td> <td>NP</td> </tr> <tr> <td>source</td> <td>(12)</td> </tr> <tr> <td>target</td> <td>(22)</td> </tr> </tbody> </table>	st-rule	NP	source	(12)	target	(22)
st-rule	V													
source	(11)													
target	(21)													
st-rule	NP													
source	(12)													
target	(22)													

図 4：変換規則の例

という “threw” 入力時までの原言語の句構造木が構築される。解析部では、このようにして語の入力ごとに句構造木を更新する。

3.2 漸進的な変換

変換部では、解析部により句構造木が更新されるたびに、以下の二つの手続きを実行する。

- 手続き 1：句構造木に変換規則を適用することにより、目標言語表現を決定する。
- 手続き 2：句構造木中の目標言語表現が決定した部分を、“!” で置き換える。

本手法で用いる変換規則は漸進的な独英変換で Amtrup が使用した規則 [3] に基づく。変換規則は文法規則に対応して用意される。例として文法規則 “ $VP \rightarrow V NP$ ” に対応する変換規則を図 4 に示す。これは文法規則 “ $VP \rightarrow V NP$ ” によって生成された句構造木中の VP の部分に対して適用される変換規則である。また、VP 中の V の部分に V に対する変換規則を、NP の部分に NP に対する変換規則をそれぞれ適用した結果を、phon に記述された順に出力することを表す。

ここでは、入力語 “threw” を解析して得られた句構造木 (2) の中の VP に対する部分構造木

$$[[threw]_V[?]_{NP}]_{VP}$$

に対して図 4 の変換規則を適用した場合について考える。まず、 $[threw]_V$ の部分に V に対する変換規則が適用される。その結果 “threw” は「投げた」に変換される。この「投げた」は翻訳文の最後に再度用いられるため、変換部内で保持される。

入力	解析部による更新	出力	変換部による更新
	[?]S (初期状態)		
He	[[[He]N]NP[?]VP]S	彼は	[[[!]N]NP[?]VP]S
threw	[[[!]N]NP[[threw]V[?]NP]VP]S	投げた	[[[!]N]NP[[!]V[?]NP]VP]S
the	[[[!]N]NP[[!]V[[the]DET[?]N]NP]VP]S		[[[!]N]NP[[!]V[[[!]DET[?]N]NP]VP]S
ball	[[[!]N]NP[[!]V[[[!]DET[ball]N]NP]VP]S	ボールを 投げた	[[[!]N]NP[[!]V[[[!]DET[!]N]NP]VP]S
.			

表5：漸進的な翻訳処理

次に [?]NP に対して NP に対する変換規則が適用される。しかし、NP の項は空所であるので、この段階では NP に対する変換は行なわれない。結果的に、図4の変換規則の適用によって「投げた」という目標言語表現が生成される。

次に手続き2に従い、変換の終了した“threw”を“!”に置き換える。VP に対する部分構造木は

$$[[!]V[?]NP]VP$$

に更新される。“!”は、その語に対する変換が終了していることを表す。したがって、次の語“the”の入力によって、この部分構造木に再び変換規則が適用されたとき、“threw”に対する訳は出力されない。変換部の以上の動作により、漸進的に翻訳文が生成される。

表5に、英語文“He threw the ball .”に対する一連の翻訳処理の流れを示す。原言語の入力ごとの解析部と変換部の処理により、目標言語表現を漸進的に決定できることがわかる。

4 翻訳実験

本稿で提案する漸進的な英日翻訳手法のプロトタイプシステムを、Common Lisp を用いてワークステーション上に作成した。本手法を評価するために、中学一年生用の英語教科書の本文中のすべての会話文を対象に、本システムによる翻訳実験を行なった。辞書は主に教科書の巻末辞書を利用して作成した。

4.1 プロトタイプシステムの諸性能

プロトタイプシステムは解析部、変換部、文法、変換規則、辞書から構成されている。文法は60規則からなり、各文法規則に対して変換規則が用意されている。辞書には単語とその統語範疇が記述されている。辞書の規模は350語程度である。

	文数	割合
(1)	180文	83.3%
(2)	13文	6.0%
(3)	10文	4.6%
(4)	13文	6.0%

表6：実験結果

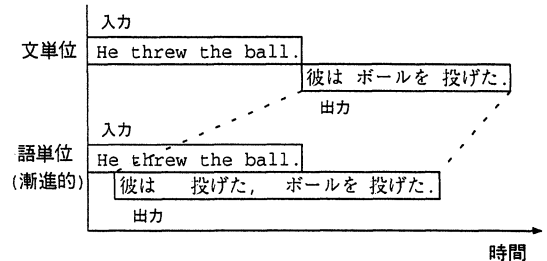


図7：翻訳文を出力するタイミングの比較

なお、英語には“look for”や“get up”など直後の語と呼応してある意味をなす単語もあり、語単位での翻訳では正しく処理できない。本システムではこれらを熟語として扱い、次の単語が入力された後、その翻訳処理を駆動する。

4.2 実験結果と評価

本実験で用いた原言語データは27対話216文からなる。評価は、プロトタイプシステムによって得られた翻訳文をその質によって、

- (1) 正しく意味が通じ、かつ違和感もない日本語文
- (2) 意味は通じるが違和感のある日本語文
- (3) 意味が誤っている日本語文

に分類し、(1)と(2)を成功、(3)を失敗とした。表6に実験の結果を示す。(4)は構文解析に失敗し、翻訳文が得られなかった原言語文である。

漸進的な翻訳処理によって、文例データ 216 文のうち約 89.4% に相当する 193 文についての確かな意味内容を持つ翻訳文を得ることができた。翻訳の対象とした原言語文の多くは、単純な構造を持つ。しかしながら、これらの結果は漸進的な翻訳手法においても一定水準以上の翻訳の質を維持できることを示唆する。

翻訳処理の時間に関しては、語の入力に対して実時間で出力されることが確認できた。図 7 に本システムによる漸進的な翻訳文の出力と、従来の文単位での処理による翻訳文の出力とのタイミングの比較を示す。漸進的に翻訳処理を進めることにより、翻訳文の出力を開始するまでの時間と終了するまでの時間を大幅に短縮できる。これは、漸進的な処理によりユーザーの待ち時間を大幅に短縮できることを意味している。

4.3 失敗例の分析

翻訳に失敗した 23 文について分析する。まず、23 文中の 11 文は、入力文の曖昧性など機械翻訳一般に生じる問題であった。しかし、それ以外の 12 文は漸進的な翻訳手法に特有と思われる失敗である。これらは、その原因に基づき以下の 3 つに分類される。

- 並列句の解析の失敗 (6 文)

“and” や “or” などによる並列句を含んだ原言語文を解析できない。これは、接続詞が入力された時点では並列句の構造が明らかでないことによる。

(例) “I change clothes and eat dinner .”

- 統語範疇を複数持つ語の解析の誤り (3 文)

入力語が複数の統語範疇を持つ場合、解析に失敗することがある。これは生成しうる句構造木の候補が、複数存在することによる。

(例 1) “Are these books yours ?”

“Are these” の時点で “these” を名詞として解析し、その結果「これらは」を訳出した。

(例 2) “I do my homework .”

“do” を助動詞として解析した。

- 慣用表現に関する失敗 (3 文)

慣用表現に対して適切な翻訳文を生成する

ことができない。これらの表現も、熟語として扱えば翻訳可能である。ただし、

- (1) 訳出可能となるまで多くの入力語を必要とするため、漸進性が大きく損なわれる
- (2) 慣用表現とそれ以外の表現との区別が困難である

などといった問題が生じる。

(例) “Here you are .”
“I beg your pardon .”

これらはいずれも漸進的な翻訳処理に特有の問題であり、高精度の漸進的な機械翻訳システムを実現するためには、これらの問題を解決する必要がある。

5 おわりに

本稿では、語単位で左から右へ翻訳処理を進める英日機械翻訳手法を提案した。本手法のプロトタイプシステムによる翻訳実験により、語単位の漸進的な翻訳処理でも意味内容の通じる程度の翻訳の質を維持できることを確認した。また、入力と同時進行的に翻訳処理を進めることによって、ユーザーの待ち時間が大幅に短縮されることを示した。

実験によって判明した漸進的な翻訳処理に特有な失敗例の解決法を考案すること、曖昧な入力文の処理方法を検討すること、意味解析により翻訳の精度をあげることなどは今後の課題である。

参考文献

- [1] Inagaki, Y. and Matsubara, S.: Models for Incremental Interpretation of Natural Language, *Proc. of the 2nd Symposium on Natural Language Processing*, pp. 51-60 (1995).
- [2] 伝 康晴: 話し言葉における非文法的現象とその機械的処理, 人工知能学会言語・音声理解と対話処理研究会, SIG-SLUD-9503-3, pp. 9-16 (1996).
- [3] Amtrup, J. W.: Chart-based Incremental Transfer in Machine Translation, *Proc. of 6th International Conference of Theoretical and Methodological Issues in Machine Translation*, pp. 188-195 (1995).