GA-ILMTにおける翻訳例の局所的対応関係に着目した帰納的学習手法の有効性について

越前谷 博[†]· 荒木 健治[††] 栃内 香次[†]

[†]北海道大学大学院工学研究科 [††]北海学園大学工学部

1. はじめに

近年、コンピュータネットワークの普及に伴い、迅速かつ正確に母国語と異なる言語で表現されている情報を得ることが重要視されている。このような状況下において、これまでに多くの機械翻訳手法が提案されてきた[1, 2, 3]。そして、それに伴い、現在、いくつかの機械翻訳システムが商用化されている。しかし、商用化されている機械翻訳システムは、その翻訳精度及び品質の面において、ユーザの要求を十分に満たしているとはいえない。商用機械翻訳システムに取り入れられている手法の主流となっているのは、解析型機械翻訳手法[1]である。この手法では、有限個の文法規則を利用するため、多様な言語現象に十分対処できないことが問題点となっている。こうした問題点を解決するために、実例型機械翻訳手法[2][3]が提案されているが、実用化に向けては、大量の翻訳例が必要になるという問題点を抱えている。

我々は、より翻訳精度及び品質の高い学習型機械翻訳システムの実現に向け、従来より、遺伝的アルゴリズムを適用した帰納的学習による機械翻訳手法(GA-ILMT)を提案している[4]。GA-ILMTは、与えられた翻訳例に対して帰納的学習を行うことにより、翻訳ルールを自動的に獲得する。そして、その翻訳ルールを用いて翻訳を行う。また、遺伝的アルゴリズムを適用することにより、新たな翻訳例の生成と効率的な翻訳結果の生成を行い、翻訳ルールの学習能力を向上させている。

ところで、GA-ILMTの学習処理における翻訳例からの翻訳ルールの獲得は、翻訳例全体のみに着目することにより行われている。したがって、翻訳例の原文とその訳文の対応関係において、局所的には対応関係が得られる場合でも、文全体において対応関係が得られなければ、翻訳ルールを獲得することはできない。この問題を解決するために、本稿では、翻訳例の原文とその訳文の間の局所的対応関係に着目することのできる帰納的学習手法を提案する。そして、翻訳例の局所的対応関係に着目した帰納的学習手法の導入により、GA-ILMTの学習能力の向上を図った結果について述べる。

2.GA-ILMTの概要

GA-ILMTに基づき構築した英日機械翻訳システムの構成を図1に示す。

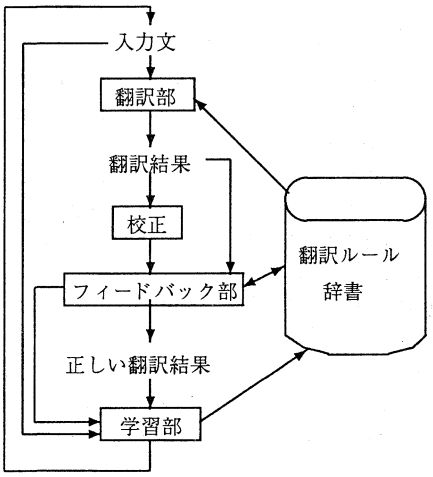

図1: GA-ILMTのシステム構成図

まず、原文として英文が入力されると、翻訳部において、それまでに獲得された辞書中の翻訳ルールを用いて、翻訳結果を生成する。その翻訳結果に誤りが含まれている場合には、人手による校正を行い、正しい翻訳結果を得る。次いで、フィードバック部において、正しい翻訳結果に基づき、翻訳に使用された翻訳ルールに対する適応度の決定と淘汰を行う。そして、学習部において、与えられた翻訳例に対し、選択交配と突然変異といった遺伝的アルゴリズムの基本操作を行い、多くの翻訳例を自動的に生成する。さらに、翻訳例に対し、帰納的学習を

行うことにより新たな翻訳ルールを獲得し、以後の翻訳に活用する。

3. 局所的対応関係に着目した帰納的学習手法

3.1 概要

GA-ILMTは、翻訳例に対し、帰納的学習を行うことにより翻訳ルールを獲得する。まず、与えられた翻訳例中から、英文とその日本語訳文のそれぞれにおいて、共通部分を持つ2つの翻訳例を選択する。図2に示す2つの翻訳例間では、英文においては「He likes」、日本語訳文においては「彼/は」と「が/好き/です」が共通部分となる。次いで、翻訳例の英文とその日本語訳文の差異部分に着目する。その結果、英文とその日本語訳文において、差異部分が1対1に対応付けられる場合にのみ差異部分と共通部分を抽出し、翻訳ルールを獲得する。差異部分が1対1に対応付けられるということは、文を構成している全ての単語について対応関係が一意に決定できることを意味している。本稿では、英文とその日本語訳文において全ての単語の対応関係が得られる場合を文全体の対応関係が成立していると位置付ける。しかし、図3に示す2つの翻訳例からは、差異部分が英文とその日本語訳文において複数存在するため、文全体の対応関係が得られず、翻訳ルールを獲得することができない。したがって、GA-ILMTの帰納的学習では、翻訳ルールを獲得する際、英文とその日本語訳文の対応関係が文全体に対して得られなければ、全く翻訳ルールの獲得が行えないことになる。これは、GA-ILMTの帰納的学習が、常に翻訳例全体のみに着目していることを意味する。しかし、人間は、英文とその日本語訳文の対応関係を決定する場合、常に文全体のみに着目しているのではなく、局所的に見ることにより、部分的に対応関係を決定することができると考えられる。そこで、我々は、翻訳例の局所的対応関係に着目することのできる帰納的学習を提案する。局所的対応関係に着目する際には、過去に与えられた情報に基づき対応関係の曖昧さを解消する。

3.2 処理過程

局所的対応関係に着目した帰納的学習の処理過程を述べる。

(1) 英文とその日本語訳文のそれぞれにおいて、共通部分を持つ2つの翻訳例を選択する。図3の2つの翻訳例では、英文においては「likes」、日本語訳文においては「は」と「が/好き」が共通部分となる。

(He likes tennis. ;彼/ は/テニス/が/好き/です.)
(He likes tea. ;彼/ は/お茶/が/好き/です.)

↓

差異部分
(tennis ; テニス), (tea ; お茶)
共通部分
(He likes @0. ; 彼/は/@0/が/好き/です.)

図 2: 差異部分と共通部分の抽出

(I do not like baseball.
; 私/は/野球/が/好き/ではありません.)
(You like volleyball.
; あなた/は/バレーボール/が/好き/です.)

↓

共通部分
(likes ; は), (likes ; が/ 好き)

図 3: 差異部分が複数存在する場合の翻訳例

(2) 共通部分を用いることにより、英文とその日本語訳文のそれぞれに対して部分的な抽出を行い、それらを組み合わせる。図4に（I do not like baseball. ; 私/は/野球/が/好き/ではありません。）に対して行われた局所的対応関係の抽出結果を示す。図4中の正解率については処理過程の（3）でその詳細を述べる。

(3) 正しい局所的対応関係を決定する。処理の詳細を以下に示す。

- 抽出された各々の局所的対応関係において、共通部分を除く英単語と日本語単語の全ての単語の組合せを取り出す。
- 単語の組合せに対し、与えられた翻訳例を参照することにより、個々の単語の組合せがどの程度正しいのかを以下の式により求める。

$$\text{単語の正解率 (\%)} = \frac{\text{単語の組合せの出現回数}}{\text{英単語の出現回数}} \times 100.0$$

図4の中の局所的対応関係（I do not like; は/野球/が/好き/ではありません）に対する結果を表1に示す。この場合、共通部分は（like ; は）であり、共通部分を除いた（I do not ; 野球/が/好き/ではありません）を用いて、個々の単語の正解率を求める。

翻訳例:(I do not like baseball.
; 私/は/野球/が/好き/ではありません.)

抽出された局所的対応関係	正解率
(I do not like ; は/ 野球/が/好き/ではありません)	68
(I do not like ; 私/は)	59
(I do not like ; が/ 好き/ではありません)	51
(I do not like ; 私/は/野球/が/好き)	95
(I do not like ; は/野球)	43
(I do not like ; 野球/が/好き)	43
(like baseball ; は/野球/が/好き/ではありません)	100
(like baseball ; 私/は)	60
(like baseball ; が/好き/ではありません)	20
(like baseball ; 私/は/野球/が/好き)	100
(like baseball ; は/野球)	100
(like baseball ; 野球/が/好き)	100

図 4: 局所的対応関係の抽出結果

表 1: 単語の正解率

	野球	が	好き	ではありません
I	17%	14%	14%	17%
do	100%	50%	50%	50%
not	12%	6%	6%	87%

- 単語の正解率に基づいて、各々の抽出された局所的対応関係がどれだけ正しいのかを以下の式により求める。

$$\text{局所的対応関係の正解率 (\%)} = \frac{\text{単語の正解率の最大値の合計}}{\text{構成英単語数}}$$

表 1 の局所的対応関係（I do not like ; は/野球/が/好き/ ではありません）の正解率は (17+100+87)/3=68.0%となる。

(4) 抽出された局所的対応関係の中から、正解率が最大のものを正しい局所的対応関係であると決定し、それを翻訳ルールとして獲得する。したがって、図 4 の例では、破線で囲まれた 4 つの翻訳ルールが得られる。また、正しい翻訳ルールとしては、(like baseball ; 野球/が/好き）が得られることになる。そして、翻訳例に対して、抽出箇所を変数に置き換えることにより、翻訳ルール（I do not @0. ; 私/は/@0/ ではありません。）が得られる。

このように、帰納的学習において、翻訳例の局所的な対応関係に着目することにより、翻訳例からの学習能力が向上し、より多くの翻訳ルールを獲得することが可能となる。

4. 評価実験

4.1 実験方法

実験は、最初に、本稿で提案する局所的対応関係に着目した帰納的学習手法を、図 1 に示す GA-ILMT の学習部に導入し、実験システムを作成した。次いで、辞書の初期状態を空として、中学 1 年生用教科書ガイド・ワンワールド [5] に掲載されている英文とその日本語訳文の１５０組（Lesson1～Lesson4）を学習データとして用い、翻訳と学習を 1 文ずつ繰り返し行った。そして、翻訳評価データとして、同じく中学 1 年生用教科書ガイド・ワンワールド [5] に掲載されている英文１５４文（Lesson5～Lesson6）を用いて翻訳を行った。翻訳結果が生成された後には、1 文ずつ正しい日本語訳文を与え、学習を行った。

4.2 評価方法

GA-ILMT より生成される翻訳結果に対する評価方法について述べる。有効な翻訳は以下の 2 つである。

1) 未登録語を含まない正翻訳

2) 未登録語を含む正翻訳

未登録語を含む正翻訳は、未登録語に名詞句や形容詞などの単語の翻訳ルールを与えることにより、容易に未登録語を含まない正翻訳が得られる翻訳結果である。

1) 未登録語を含まない誤翻訳

2) 未登録語を含む誤翻訳

3) 翻訳不能

未登録語を含む誤翻訳は、名詞句や形容詞以外の単語の翻訳ルールを必要とする翻訳結果である。翻訳不能は、入力文に対し、適用可能な文の翻訳ルールが全く存在せずに、翻訳が行えなかった翻訳結果である。

4.3 実験結果

表2に、局所的対応関係に着目した帰納的学習手法を導入しなかった場合の実験結果を示す。表3には、局所的対応関係に着目した新たな帰納的学習手法を導入した場合の実験結果を示す。

表 2: 従来の帰納的学習手法を用いた場合の実験結果

		翻訳率	合計
有効な翻訳	正翻訳	33.1%	47.4%
	未登録	14.3%	
無効な翻訳	誤翻訳	14.3%	52.6%
	未登録	21.4%	
	翻訳不能	16.9%	

表 3: 新たな帰納的学習手法を用いた場合の実験結果

		翻訳率	合計
有効な翻訳	正翻訳	41.6%	58.5%
	未登録	16.9%	
無効な翻訳	誤翻訳	15.6%	41.5%
	未登録	16.2%	
	翻訳不能	9.7%	

5. 考察

表2と表3に示すように、有効な翻訳率は47.4%から58.5%に増加した。これは翻訳評価データ１５４文において、無効な翻訳から有効な翻訳へと改善された翻訳結果が２１文、また、有効な翻訳から無効な翻訳へと移行した翻訳結果が４文、結果として、有効な翻訳が１７文増加したことに相当する。表4に、無効な翻訳から有効な翻訳へと改善された翻訳結果２１文の内訳を示す。

表4より、未登録語を含まない誤翻訳から有効な翻訳になったものが3文、未登録語を含む誤翻訳から有効な翻訳になったものが１１文、そして、翻訳不能から有効な翻訳になったものが7文であった。したがって、本稿で提案する帰納的学習手法を導入していない場合の実験結果において、未登録語を含まない誤翻訳の13.6%が、また、未登録語を含む誤翻訳の33.3%が、そして、翻訳不能の26.9%が改善されたことになる。したがって、本稿で提案する帰納的学習手法が、無効な翻訳において最も占有率の高い未登録語を含む誤翻訳と翻訳不能に対して有効であったことを示している。これは、翻訳例の局所的対応関係に着目することにより、翻訳ルールの学習能力が向上し、より翻訳に有効な翻訳ルールが増加したためであると考えられる。

表 4: 有効な翻訳に改善された翻訳結果の内訳

	文数
誤翻訳（未登録語無）→正翻訳（未登録語無）	3
誤翻訳（未登録語有）→正翻訳（未登録語無）	6
誤翻訳（未登録語無）→正翻訳（未登録語有）	5
翻訳不能→正翻訳（未登録語無）	2
翻訳不能→正翻訳（未登録語有）	5
合計	21

6. おわりに

本稿では、GA-ILMTの帰納的学習の能力向上のために、局所的な対応関係に着目することにより、翻訳ルールを獲得する手法を提案した。実験の結果、翻訳率は47.4%から58.5%に増加し、本手法の導入が翻訳ルールの学習能力を向上させる際に有効であったことが確認できた。また、本手法は、翻訳例の局所的な対応関係に着目しているため、翻訳例全体からは容易に対応関係を決定することが困難なより長い文から構成されている翻訳例に対しても有効であると考えられる。今後は、より長い文から構成されている翻訳例を対象とした性能評価実験を行い、より翻訳精度及び品質の高い学習型機械翻訳システムの実現に向けての研究を行う予定である。

参考文献

[1] 野村浩郷（編）:言語処理と機械翻訳, 講談社 (1991).

[2] 古瀬蔵, 隅田英一郎, 飯田仁：経験的知識を活用する変換主導型機械翻訳, 情報処理学会論文誌, Vol. 35, No. 3, pp. 414–425 (1994).

[3] 北村美穂子, 松本裕治：対訳コーパスを利用した翻訳規則の自動獲得, 情報処理学会論文誌, Vol. 37, No. 6, pp. 1030–1040 (1996).

[4] 越前谷博, 荒木健治, 桃内佳雄, 栃内香次：実例に基づく帰納的学習による機械翻訳手法における遺伝的アルゴリズムの適用とその有効性, 情報処理学会論文誌,Vol.37,No.8,pp.1565-1579,(1996).

[5] 教科書ガイド教育出版ワンワールド, 日本教材, 東京,(1991).