# 学習型機械翻訳手法 GA-ILMT における帰納的学習を用いた淘汰手法の有効性について

越前谷　博[†]・荒木健治[††]・桃内佳雄[†]・栃内香次[††]

# Method of Selection Using Inductive Learning in GA-ILMT and Its Effectiveness

Hiroshi ECHIZENYA[†], Kenji ARAKI[††], Yoshio MOMOUCHI[†] and Koji TOCHINAI[††]


## 要　旨

我々は，従来の機械翻訳手法の抱える問題点を解決する手法として，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。本論文では，このGA-ILMTを用いて，学習型機械翻訳システム上で獲得された知識に含まれている誤った知識を，帰納的学習により自動的に消去することを試みた。誤った知識を消去する際には，知識獲得に使用された翻訳例を最大限に利用するために，新たな知識を与えるのではなく，翻訳例に基づく帰納的学習により行う。したがって，我々は，GA-ILMTの誤翻訳ルールの消去を目的とする淘汰手法の精度向上を，与えられた翻訳例を用いた帰納的学習の導入により行う手法を提案する。実験の結果，誤翻訳ルールにおける適合率は89.1%から92.8%，再現率は5.9%から65.0%に向上した。さらに，誤翻訳は56.0%，処理時間は52.1%減少した。これらの結果より，学習型機械翻訳システムにおける誤った知識に対する解決策として，帰納的学習の導入が有効となることを確認できた。

## Abstract

We have already proposed a method of Machine Translation using Inductive Learning with Genetic Algorithms(GA-ILMT). In this paper, we aim at removing automatically



† 北海学園大学工学部電子情報工学科
†† 北海道大学大学院工学研究科
† Dept. of Electronics and Information, Hokkai-Gakuen University
†† Division of Electronics and Information, Hokkaido University

erroneous knowledge which is included in gained knowledge by using inductive learning. The system removes erroneous knowledge by inductive learning based on translation examples without giving any new knowledge. In GA-ILMT, we proposed a method of selection using inductive learning which utilizes given translation examples. Our proposed method does not need any analytical knowledge to keep robustness. The results of the evaluation experiments show that precision increased from 89.1% to 92.8% and recall increased from 5.9% to 65.0%. Moreover, erroneous translation results decreased by 56.0%, and CPU time decreased by 52.1%. Therefore, we confirmed that our proposed method is effective to remove erroneous knowledge in machine translation system based on learning capability.

## 1. はじめに

近年，インターネットを介して多くの情報交換が行われている。その際，異言語の情報を正確かつ迅速に処理することが要求される。そうした状況下において，実用的な機械翻訳システムの開発に向け，多くの研究が行われてきた。しかし，翻訳の精度及び品質において，これまでに商用化されている機械翻訳システムはユーザの要求に十分応えられるものとはなっていない。

我々は，こうした背景において，人間の言語及び知識獲得能力の工学的な実現 [1] [2] という観点から，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[3] [4] を提案している。この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では，与えられた翻訳例の字面情報のみから翻訳ルールを帰納的学習により獲得し，その翻訳ルールに基づき翻訳を行う。したがって，良質な翻訳を行うためには，翻訳例を追加していくだけで良いという実例型機械翻訳手法 [5] [8] の利点を持っている。また，帰納的学習を翻訳例の字面情報に対して行うのは，現在の形態素解析や構文解析の精度が十分なものとなっていないためである。このような翻訳例に対する帰納的学習が，解析型機械翻訳手法 [10] [11] の抱えている未知の言語現象に対処することが困難であるという問題点に対して有効になると考えられる。また，遺伝的アルゴリズム [12] [14] の選択交配や突然変異などの遺伝的オペレータを与えられた翻訳例に対して行うことにより，自動的に多くの新翻訳例を生成する。したがって，実例型機械翻訳手法の抱える，十分な翻訳精度及び品質を得るには膨大な量の翻訳例を与えなければならないという問題点に対しても有効となる。このように，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は，予め文法規則を与えることなく，学習的手法により翻訳に必要なルールを自動獲得する能力を有する。そして，遺伝的アルゴリズムの適用により，

さらに，その能力を向上させている。この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を，以後 GA-ILMT (Machine Translation using Inductive Learning with Genetic Algorithms) と呼ぶことにする。

翻訳例から知識を自動抽出する研究 [5] [6] [7] [8] [9] が近年盛んに行われている。しかし，翻訳例から自動抽出により知識を得る場合，誤った知識の獲得を伴うことが多い。誤った知識に対し，その原因と対処法に関する考察は行われているが，その解決方法の多くが誤った知識の獲得を未然に防ぐことに観点が置かれている。具体的な解決方法としては，翻訳例の変更や追加 [6] [8] [9]，または，解析的な知識を改善すること [5] [6] [7] [8] [9] などが挙げられる。このような解決方法は，確かに特効性のある方法となりえるが，例外的なものに対処することが困難であるといった解析型機械翻訳手法の問題点や，実例型機械翻訳手法における精度向上には豊富な量の翻訳例が必要になるといった問題点を抱え込むことになると考えられる。そこで，我々は，これまでの帰納的学習に基づき知識を獲得するという観点より，この問題の解決を目指した。それはシステム自身が，帰納的学習により誤った知識を消去するということである。誤った知識の消去を帰納的学習により行う際には，知識獲得に用いた与えられた翻訳例を利用する。したがって，新たな知識を与えることなく誤った知識を消去することが可能となる。

我々は，この帰納的学習が学習型機械翻訳システムにおける誤った知識に対して有効となることを確認するために，帰納的学習の導入を GA-ILMT の淘汰処理に対して行った。GA-ILMT では，多くの翻訳ルールが自動的に獲得されるが，その際には，誤翻訳ルールも数多く生成される。そして，そのような誤翻訳ルールは，淘汰処理によって消去されていく。しかし，従来の翻訳精度を用いた淘汰手法では，淘汰処理の精度が低く，誤翻訳ルールに対して十分な淘汰は行われていなかった。また，これまでの GA-ILMT では，翻訳ルールの獲得には帰納的学習が用いられていたが，淘汰処理に対しては用いられていなかった。我々は，淘汰処理に対しても帰納的学習を導入することにより，GA-ILMT を首尾一貫して帰納的学習を用いた，より良い学習型機械翻訳システムに改善できると考えている。そこで，本論文では，GA-ILMT において獲得された翻訳ルール中に存在する誤翻訳ルールを，与えられた翻訳例を利用した帰納的学習により淘汰する手法を提案 [15] [16] [17] する。このように，GA-ILMT の淘汰処理において，従来の翻訳精度を用いた淘汰手法に対し，本論文で提案する帰納的学習を導入した淘汰手法を，帰納的学習を用いた淘汰手法と呼ぶことにする。さらに，GA-ILMT の淘汰処理に帰納的学習を導入することの有効性を確認するために行った評価実験及び考察結果について述べる。

## 2．GA-ILMT

### 2.1　システム構成

GA-ILMT では，図1に示すように，原文とその訳文からなる翻訳例を染色体に，そして，染色体を構成している個々の単語を遺伝子として位置付けている。

また，GA-ILMT は，基本的に原文とその訳文に使用する言語を変更するだけで，多くの言語間での翻訳が可能となる。本論文では，GA-ILMT を用いて英日の翻訳を行う学習型機械翻訳システムを構築し，実験を行った。GA-ILMT に基づくシステムの概要を図2に示す。


図1　染色体と遺伝子
Fig.1　Chromosome and genes.

まず，原文として英文を入力する。すると，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最適な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，本論文で提案する帰納的学習を用いた淘汰手法により誤翻訳ルールを淘汰する。そして，学習部において，翻訳例に対し選択交配と突然変異を行い，多様な翻訳例と翻訳ルールを生成する。選択交配には，一点交叉 [3] と二点交叉 [18] の2つの交

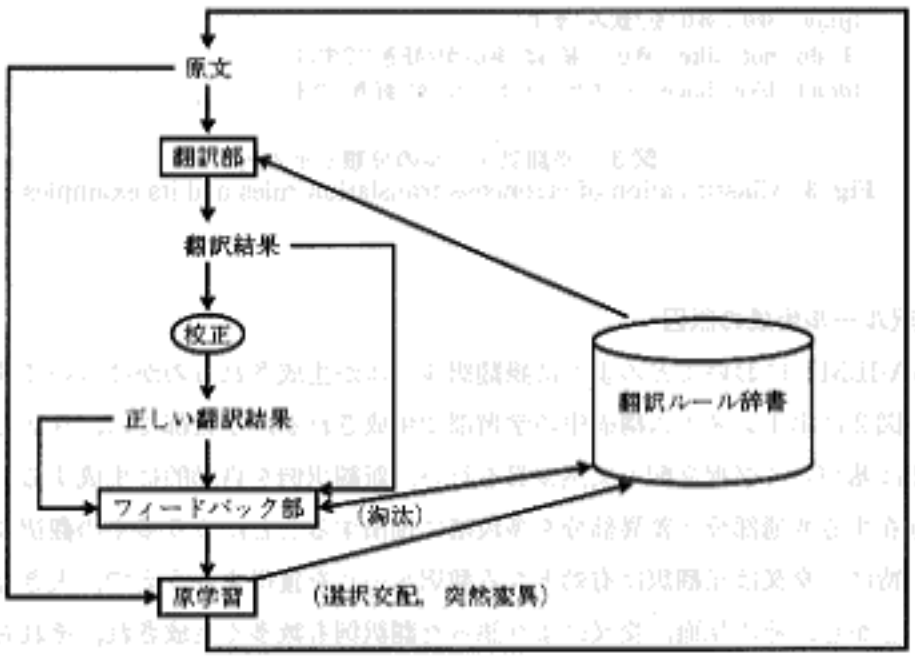


図2　GA-ILMT のシステム構成
Fig. 2　System configuration of GA-ILMT.

叉手法を取り入れている。このような処理を繰り返すことにより，システム自身がより良い学習型機械翻訳システムへと進化していく。

### 2.2　誤翻訳ルールの定義

GA-ILMTにおいて生成される誤翻訳ルールは，以下の2つに分類することができる。

① 意味的な誤りを含む誤翻訳ルール

② 英文とその日本語訳文の対応関係が成立しない誤翻訳ルール

図3にそれぞれの具体例を示す。

図3の①の意味的な誤りを含む誤翻訳ルールは，意味的には誤っていると考えられるが，英文とその日本語訳文の対応関係は成立している。それに対し，図3の②の英文とその日本語訳文の対応関係が成立しない誤翻訳ルールは，対応関係が成立せず，明確に誤った翻訳ルールであると判断できる。このような対応関係が成立しない誤翻訳ルールは，GA-ILMTにおいて誤翻訳の生成の原因となる。そこで，本論文では，英文とその日本語訳文の対応関係が成立せず，人間が見た場合，明らかに誤りとわかる誤翻訳ルールを淘汰の対象とする。

①意味的な誤りを含む誤翻訳ルール
(She is my father.；彼女/は/私の/父/です。)
(I am not my classmate.；僕/は/僕の/クラスメイト/ではありません。)
②英文とその日本語訳文の対応関係が成立しない誤翻訳ルール
(Akiko is not my teammate.；あき子/は/僕の/チーム仲間/です。)
(not my teammate；僕の/チーム仲間/です)
(play @0；@0/を/飲み/ます)
(I do not like @0.；私/は/@0/が/好き/です。)
(don't like juice；バスケットボール/が/好き/です)

図3　誤翻訳ルールの分類とその例
**Fig. 3**　Classification of erroneous translation rules and its examples.

### 2.3　誤翻訳ルール生成の原因

次に，GA-ILMTにおいてどのように誤翻訳ルールが生成されるのかについて述べる。翻訳ルールは，図2に示すシステム構成中の学習部で生成される。学習部では，与えられた翻訳例の字面情報に基づいて選択文配と突然変異を行い，新翻訳例を自動的に生成する。そして，翻訳例間に存在する共通部分と差異部分を多段階に抽出することにより多くの翻訳ルールを生成する[19]。特に，交叉は正翻訳に有効となる翻訳ルールを獲得するうえで，大きな役割を果たしている。しかし，その反面，交叉により誤った翻訳例も数多く生成され，それらが基となり多くの誤翻訳ルールが生成される。図4に一点交叉により生成される誤翻訳例の具体例を示す。

交叉では，まず，英文とその日本語訳文において，共通部分が存在する2つの翻訳例を選択

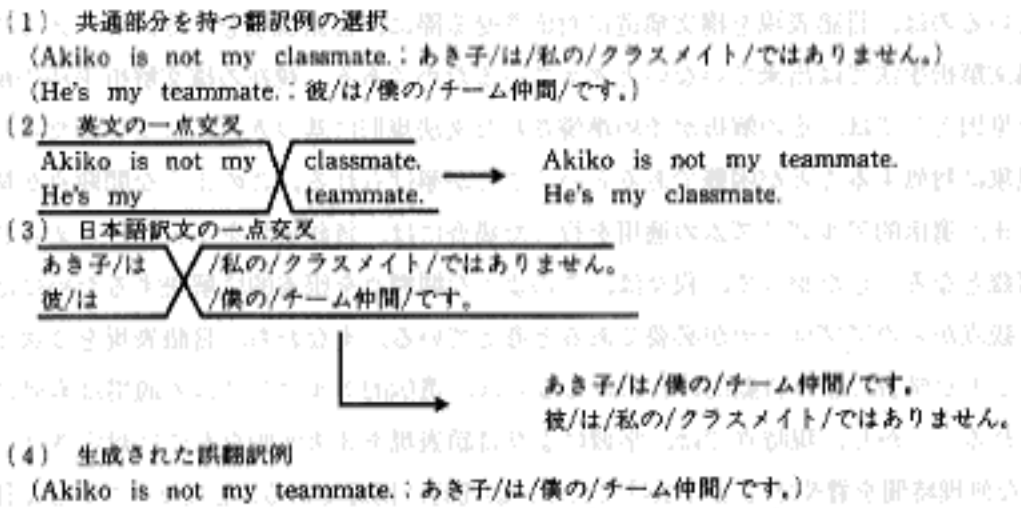


図4　一点交叉による誤翻訳例の生成例

**Fig. 4**　Examples of erroneous translation examples generated by one-point crossover.

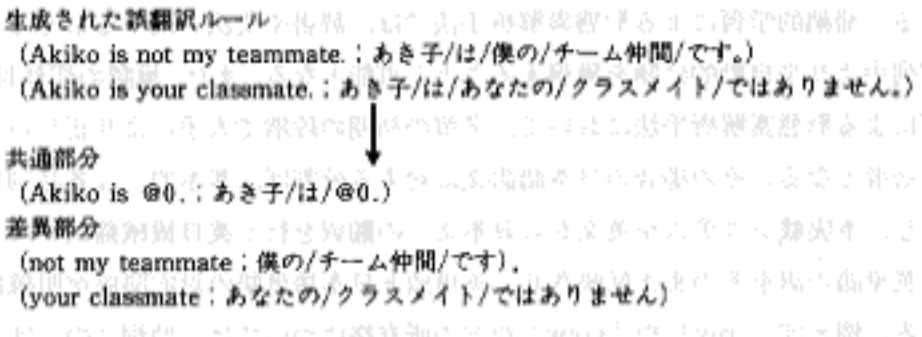


図5　誤翻訳ルールの生成例

**Fig. 5**　Examples of the generation of erroneous translation rules.

する。そして，その共通部分を交叉位置として英文とその日本語訳文のそれぞれに対して交叉を行う。図4に示す例では，英文においては「my」，日本語訳文においては「は」が交叉位置となり，図4の（4）に示す誤翻訳例が生成される。このような誤翻訳例が辞書中に蓄積されると，それらを用いた共通部分と差異部分の多段階の抽出が行われ，誤翻訳ルールが急激に増加することになる。図5に図4の（4）で示した誤翻訳例と他の誤翻訳例から生成される誤翻訳ルールの具体例を示す。

生成された翻訳ルールは，翻訳ルールに含まれる変数の数ごとに分けられ，逐次，辞書に登録される。そして，与えられた翻訳例や新たに生成された翻訳例も学習後，辞書に翻訳ルールとして登録される。また，図4に示す誤翻訳例は，交叉の際に，英文とその日本語訳文において，言語間の否定表現の位置の違いを認識できずに生成される。

我々が，遺伝的アルゴリズムを適用するにあたって，このように言語表現として単語列のみ

を扱っているのは，言語表現を構文構造に対応させる際に，言語現象を十分にとらえることが現在の構文解析手法では出来ていないと考えているためである。現在の構文解析手法の精度が不十分な原因としては，その解析が予め準備された文法規則に基づき行われているため，未知の言語現象に対処することが困難であるということが挙げられる。このような問題点が解決されないまま，遺伝的アルゴリズムの適用を行った場合には，言語現象を十分にとらえることはやはり困難となる。したがって，我々は，このような問題点を根本的に解決するためには，学習という観点からのアプローチが必要であると考えている。すなわち，言語表現を 3 次元的にとらえることが学習により自動的に行われるならば，遺伝的アルゴリズムの適用は有効になると考えられる。しかし，現時点では，学習により言語表現を 3 次元的なものに対応させることは，膨大な処理時間を費やすことは避けられず，物理的に困難である。したがって，GA-ILMT では，現在の構文解析の問題点を抱え込まないということを重視し，1 次元的に言語表現をとらえたまま，遺伝的アルゴリズムの適用を行っている。

また，翻訳例中の日本語訳文に対する形態素解析は，帰納的学習による形態素解析手法 [1] を用いている。帰納的学習による形態素解析手法では，辞書や文法を用いることなく，学習によって文字列中より半自動的に語を獲得することが可能となる。また，単語の切れ目の判断は，帰納的学習による形態素解析手法において，学習の初期の段階で人手により正しい分割結果を与える際に必要となる。その場合の日本語訳文に対する分割は，基本的には各品詞毎に行っている。しかし，本実験システムが英文から日本文への翻訳を行う英日機械翻訳システムであることから，英単語の訳をそのまま反映させ，英単語と日本語単語の対応関係が明確になるように行っている。例えば，「my」や「your」などの所有格については，助詞「の」は一定であるため，訳語である「私の」や「あなたの」を分割せずにそのまま用いている。また，否定表現「not」や助動詞「can」などの動詞に付与した形で使用されることが多い英単語については，英単語の訳が反映されるように，「ではありません」や「ことができます」という訳を分割せずに使用している。このような日本語訳文の分割に，帰納的学習による形態素解析手法を用いているのは，本形態素解析手法が文法に依存した分割を行うものではなく，ユーザの要求に応じた分割を可能とする適応的な形態素解析手法となっているからである。例えば，文法的には，「私の」は名詞「私」と助詞「の」に分割するのが一般的といえるが，帰納的学習を用いた形態素解析手法では，システムが学習により，自動的にユーザの意図を反映させ，効率の良い分割を行うことが可能となる。したがって，本論文の日本語訳文の分割は，英日機械翻訳における独自な分割であるため，英単語の訳語を反映させるような分割基準を設け，帰納的学習を用いた形態素解析手法により自動的に行った。

## 2.4 翻訳精度を用いた淘汰手法とその問題点

本項では，これまでに，誤翻訳ルールをどのように判定し，辞書中から消去していたのかについて述べる。初めに，従来の淘汰手法が翻訳過程及び翻訳結果を利用していることから，GA-ILMT の翻訳処理について説明する。図6に翻訳部で行われる翻訳処理の具体例を示す。

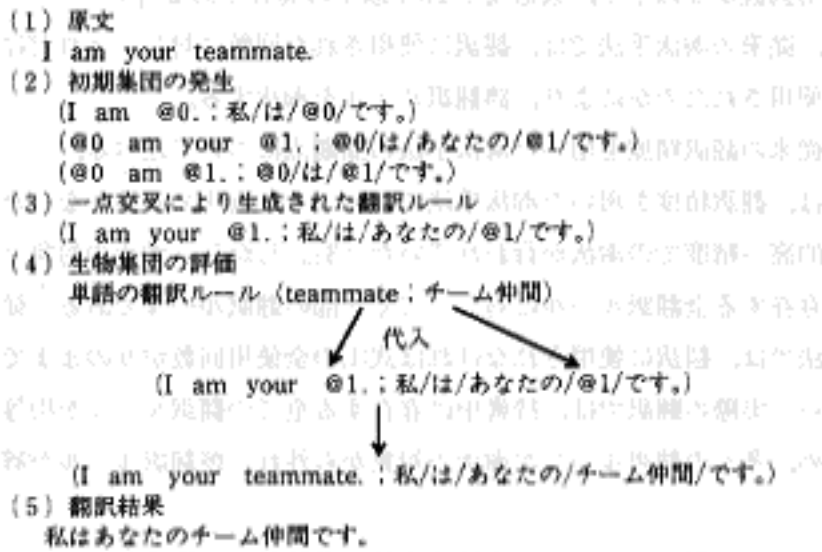


図6　翻訳結果の生成例

**Fig. 6**　An example of how the translation result is produced.

翻訳部では，まず，それまでに生成された辞書中の翻訳ルールから，入力文に対し基本的な構造を表現している文の翻訳ルールを選択し，その集団を初期集団とする。入力文に対し基本的な構造を表現している文の翻訳ルールとは，翻訳ルールの英文と入力文を比較した際に，翻訳ルールの英文から変数を除いた単語列が，入力文の一部もしくは全体と完全に対応している翻訳ルールを指している。次いで，集められた翻訳ルールに対し，選択交配や突然変異を行い，新たな翻訳ルールの生成を求める。そして，翻訳ルールの変数部分に単語の翻訳ルールを代入し，入力文と同じ英文が得られた場合，その英文に対する日本語訳文を翻訳結果とする [3]。

翻訳結果が生成された後，図2に示すシステム構成中のフィードバック部において，誤翻訳ルールに対する淘汰が行われる。以下に従来の淘汰手法の詳細を示す。

1）翻訳部で生成された翻訳結果と正しい翻訳結果が一致するかどうかを判定する。
2）一致する場合，その翻訳結果に使用した翻訳ルールを正しいものと判断し，正翻訳度数を1増加させる。一致しない場合には，誤ったものと判断し，誤翻訳度数を1増加させる。
3）以下の式（1）により適応度を求める。

$$適応度\ (\%) = \frac{正翻訳度数}{全使用回数} \times 100.0 \qquad (1)$$

全使用回数は，各々の翻訳ルールが翻訳処理に使用された際の回数である。

4）式（1）により得られた適応度に基づき淘汰を行う。淘汰の条件は，各翻訳ルールにおいて，全使用回数が5以上で，適応度が25%以下の場合である［3］。

このように，従来の淘汰手法では，翻訳に使用された回数に対し，どれだけの割合で誤翻訳結果の生成に使用されたのかにより，誤翻訳ルールを淘汰する。

次に，この従来の翻訳精度を用いた淘汰手法の問題点について述べる。

GA-ILMT では，翻訳精度を用いた淘汰手法においても，淘汰の対象となった翻訳ルールに対しては，比較的高い精度での淘汰が行われていた［3］。しかし，淘汰の対象となる翻訳ルールは，辞書中に存在する全翻訳ルールに対してごく一部の翻訳ルールである。従来の翻訳精度を用いた淘汰手法では，翻訳に使用されなければ式(1)の全使用回数が0のままであり，淘汰の対象にはならない。実際の翻訳では，辞書中に存在する全ての翻訳ルールが均等に使用されるわけではないため，多くの翻訳ルールが淘汰の対象から外れ，誤翻訳ルールが辞書中に残ったままとなる。

また，淘汰される誤翻訳ルールは，比較的少数の単語数より構成されている。GA-ILMT の翻訳処理では，原文に対して適用可能な翻訳ルールを選択する場合，図6の（2）に示すように翻訳ルールの英単語が原文の一部もしくは全体と完全に一致しなくてはならない。その結果，図4の（4）のように変数を持たず，英文において多くの単語数より構成されている翻訳ルールほど，原文と一致する可能性が低くなり，翻訳に使用されずに淘汰の対象外となる。

このように，従来の淘汰手法は，翻訳精度によって誤翻訳ルールの淘汰を行うものであり，帰納的学習によるものとは異なる。また，翻訳に使用された翻訳ルールのみが淘汰の対象となるために，淘汰される誤翻訳ルールが全誤翻訳ルールのごく一部となる。そのことが誤翻訳ルールの使用による誤翻訳の生成の原因になっている。また，多くの単語数より構成されている誤翻訳ルールを淘汰することが困難であるため，それらを用いた誤翻訳ルールの多段階の抽出処理が行われ，処理時間の増加を引き起こしている。

## 3．帰納的学習を用いた淘汰手法

### 3.1　基本的な考え方

本論文で提案する帰納的学習を用いた淘汰手法は，「翻訳ルールを構成している英文とその日本語訳文の対応関係が，与えられた正しい翻訳例中に存在しない場合，その翻訳ルールは誤っている可能性が高い。」というヒューリスティックスに基づいている。図7に与えられた翻訳例

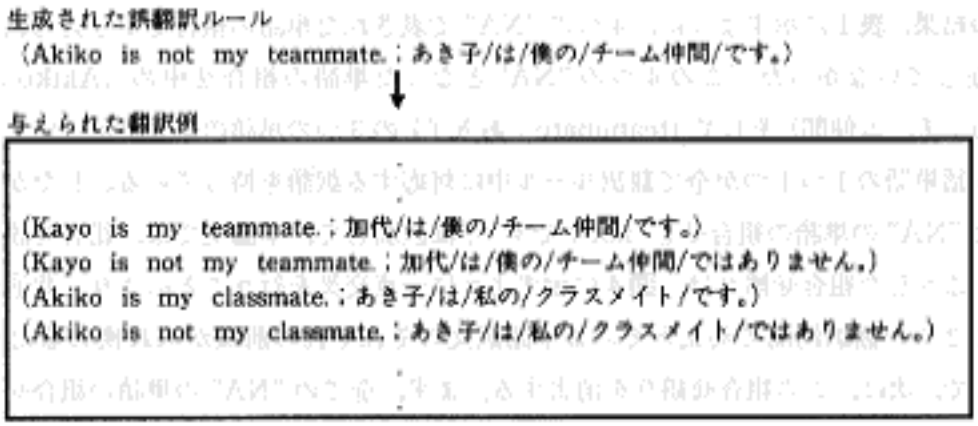


図7　与えられた翻訳例を用いた参照
**Fig. 7**　Reference utilizing given translationexamples.

を参照している具体例を示す。

図7に示すように，それまでに与えられた翻訳例より，誤翻訳ルール中の英文においては否定表現「not」が使用されているのに対し，日本語訳文では「です」という肯定表現が使用されていることが誤りであるとわかる。このように，与えられた翻訳例の英文とその日本語訳文の対応関係を利用することにより，翻訳ルール中に存在している誤りを抽出し，淘汰を行う。対応関係に着目する際には，翻訳ルールを構成している全ての単語の組合せを用いる。それは翻訳ルール中において，英文中のどの単語と日本語訳文中のどの単語が対応関係にあるのかをシステムには判断できないためである。

### 3.2　概要

図4の一点交叉により生成された誤翻訳ルール〈Akiko is not my teammate.；あき子/は/僕の/チーム仲間/です。〉を用いて，本手法の概要を述べる。初めに，この翻訳ルールを構成している英単語と日本語単語の全ての単語の組合わせを取り出し，それらが与えられた翻訳例中に存在しているのかどうかを判定する。表1に，その判定結果を示す。

表1　与えられた翻訳例による判定結果
**Table 1**　Result of decision utilizing given translation examples.

| | Akiko | is | not | my | teammate |
|---|---|---|---|---|---|
| あき子 | *AP* | *AP* | *AP* | *AP* | *NA* |
| は | *AP* | *AP* | *AP* | *AP* | *AP* |
| 僕の | *NA* | *AP* | *AP* | *AP* | *AP* |
| チーム仲間 | *NA* | *AP* | *AP* | *AP* | *AP* |
| です | *AP* | *AP* | *NA* | *AP* | *AP* |

*AP*：APpearance，　*NA*：Non-Appearance

判定の結果，表1に示すように，4つの"NA"で表された単語の組合せが与えられた翻訳例中に存在していなかった。この4つの"NA"となった単語の組合せ中の（Akiko；僕の），（Akiko；チーム仲間）そして（teammate；あき子）の3つの単語の組合せについては，英単語と日本語単語の1つ1つが全て翻訳ルール中に対応する訳語を持っている。したがって，このような"NA"の単語の組合せを（not；です）とは区別して，本論文では，組合せ誤りと位置付ける。こうした組合せ誤りは，図4に示すように一点交叉を行うことにより，共通部分を境界にして2つの翻訳例間で英文とその日本語訳文のそれぞれの前後が入れ換わるために生じる。そこで，次に，この組合せ誤りを消去する。まず，全ての"NA"の単語の組合せを集め初期集団を発生させる。そして，1つの単語の組合せを1個体として位置付け，全ての"NA"の単語の組合せ間で一点交叉を行う。図8に4つの単語の組合せに対して行われた一点交叉の実行前と実行後の一覧を示す。

そして，図8に示した交叉後の単語の組合せが，辞書中に翻訳ルールとして存在しているかどうかを検索する。その結果，（Akiko；あき子）と（teammate；チーム仲間）が共に存在していたため，交叉前の（Akiko；チーム仲間）と（teammate；あき子）を組合せ誤りであると位置付け，判定結果を"NA"から"AP"に変更する。表2に組合せ誤りの消去後の判定結果を示す。

次いで，"NA"となった単語の組合せを，生成された翻訳ルール中に含まれている誤りであるとして，"false"に置き換える。その結果を表3に示す。

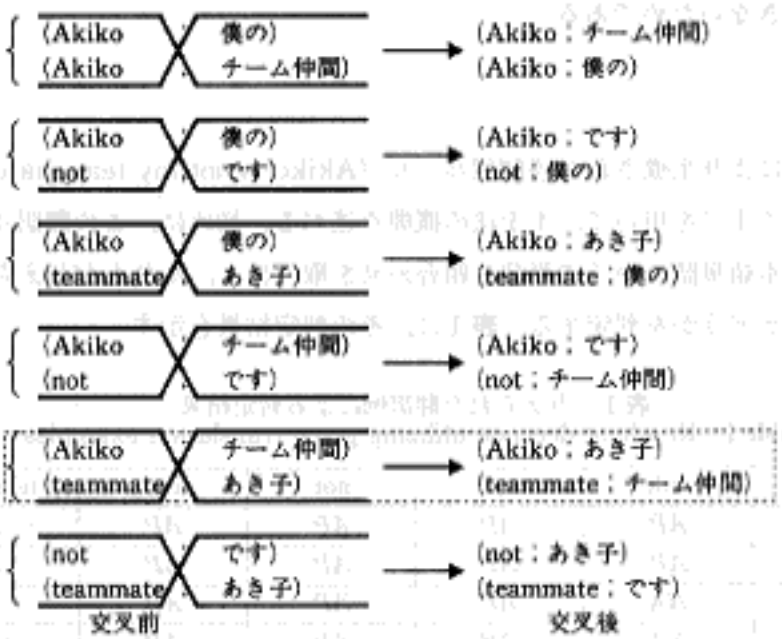


図8　組合わせ誤りの消去における一点交叉
Fig. 8　One-point crossover in deleting of erroneous combinations.

表2　判定結果
**Table 2**　Result of decision.

| | Akiko | is | not | my | teammate |
|---|---|---|---|---|---|
| あき子 | *AP* | *AP* | *AP* | *AP* | *AP* |
| は | *AP* | *AP* | *AP* | *AP* | *AP* |
| 僕の | *NA* | *AP* | *AP* | *AP* | *AP* |
| チーム仲間 | *AP* | *AP* | *AP* | *AP* | *AP* |
| です | *AP* | *AP* | *NA* | *AP* | *AP* |

*AP* : APpearance,　*NA* : Non-Appearance

表3　誤りの抽出結果
**Table 3**　Result of extraction for errors.

| | Akiko | is | not | my | teammate |
|---|---|---|---|---|---|
| あき子 | *AP* | *AP* | *AP* | *AP* | *AP* |
| は | *AP* | *AP* | *AP* | *AP* | *AP* |
| 僕の | *false* | *AP* | *AP* | *AP* | *AP* |
| チーム仲間 | *AP* | *AP* | *AP* | *AP* | *AP* |
| です | *AP* | *AP* | *false* | *AP* | *AP* |

*AP* : APpearance

そして，翻訳ルールの単語の組合せ総数に対する，"false" の単語の組合せ総数の占める割合を求め，それを誤り率とする。表3の場合，単語の組合せ総数が25，"false" の単語の組合せ総数が2であるため，誤り率は8.0%(2/25×100)となる。また，表3において，(Akiko：僕の)は「Akiko」と「僕の」共に，翻訳ルール中においてそれぞれ「あき子」と「my」が存在するため，本来，対応関係が得られるべきである。しかし，「my」に対して，「私の」と「僕の」といった異表記を認識できずに誤りと判断される。このように，本来，翻訳ルール中においては誤りではないが，最終的に誤りと判断されてしまう場合を考慮し，0.0%以外の誤り率を持った翻訳ルールを全て消去するのではなく，閾値を設け，その閾値を超えた誤り率を持つ翻訳ルールのみを消去する。この閾値は，予備実験により，5.0%が最適なものと判断された [15]。したがって，誤り率が8.0%であるこの翻訳ルールは誤翻訳ルールであるとして，辞書中から消去する。

### 3.3　処理過程

以下に，この帰納的学習を用いた淘汰手法の処理過程を示す。

1）誤翻訳ルールの決定

① 生成された翻訳ルールの英文とその日本語訳文の組から，構成している英単語と日本語単語の全ての単語の組合せを取り出す。翻訳ルールが変数を含む場合には，英文とその日本語訳文から，それぞれ変数を除いたうえで，単語の組合せを取り出す。

② 取り出された全ての単語の組合せが，与えられた翻訳例中に含まれているかどうかを判定する。

③ いずれかの翻訳例中に存在する場合，その単語の組合せに対して"AP(APpearance)"を与え，全く存在しない場合には，"NA(Non-Appearance)"を与える。

④ "NA"が与えられた単語の組合せにおいて，組合せ誤りが原因となっている単語の組合せを遺伝的アルゴリズムの基本操作を用いて消去する。以下にその手順を示す。

a）"NA"が与えられた各々の単語の組合せを1つの個体として全て取り出し，それらの集団を初期集団とする。

b）初期集団の中から2つの"NA"の単語の組合せを選択する。そして，英単語と日本語単語の境界を交叉位置として一点交叉を行い，英単語と日本語単語の組合せを変更する。

c）変更された単語の組合せが辞書中に翻訳ルールとして存在しているかどうかを検索する。

d）存在する場合，組合せ誤りであるとして，交叉前の単語の組合せの判定結果を"NA"から"AP"に変更する。

e）全ての"NA"の単語の組合せにおいて，b)以降の処理を繰り返す。

⑤"NA"である単語の組合せを生成された翻訳ルール中に含まれている誤りとし，"false"に変更する。

⑥①~⑤の処理により得られた結果から，以下の式（2）を用いて，翻訳ルールの誤り率を求める。

$$\text{誤り率}\ (\%)=\frac{\text{"false"の単語の組合せ総数}}{\text{単語の組合せ総数}}\times 100.0 \qquad (2)$$

2）誤翻訳ルールの消去

式（2）により得られた誤り率に対して，閾値$\alpha$を設け，以下の式（3）の条件を満たす場合，その生成された翻訳ルールを辞書中から消去する。

$$\text{誤り率}\geqq\alpha \qquad (3)$$

## 4．性能評価実験

### 4.1 実験方法

最初に，本論文で提案する帰納的学習を用いた淘汰手法をGA-ILMTのフィードバック部に導入し，実験システムを作成した。次いで，図2のシステム構成の流れに従って，中学1年生用教科書ガイド・ワンワールド［20］に掲載されている英文とその日本語訳文からなる翻訳例

500組に対して，1組ずつ逐次，翻訳と学習，そして，学習により得られた翻訳ルールに対する淘汰を行った。なお，本実験における辞書の初期状態は空である。

### 4.2 実験結果

実験の結果，帰納的学習を用いた淘汰手法により，多くの誤翻訳ルールを淘汰できることが確認された。表4に，従来の翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法における適合率と再現率を示す。

表4　適合率と再現率
Table 4　Precision and recall.

| | 適合率 | 再現率 |
|---|---|---|
| 翻訳精度を用いた淘汰手法 | 89.1% | 5.9% |
| 帰納的学習を用いた淘汰手法 | 92.8% | 65.0% |

適合率は，以下の式(4)により決定される。

$$\text{適合率}\ (\%)=\frac{\text{淘汰された誤翻訳ルール数}}{\text{淘汰された翻訳ルール数}}\times 100.0 \qquad (4)$$

また，再現率は，以下の式(5)により決定される。

$$\text{再現率}\ (\%)=\frac{\text{淘汰された誤翻訳ルール数}}{\text{辞書中の誤翻訳ルール数}}\times 100.0 \qquad (5)$$

表4より，適合率は89.1%から92.8%に，再現率は5.9%から65.0%に向上した。表5には，翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法における，それぞれの有効な翻訳率と無効な翻訳率を示す。

表5　有効な翻訳率と無効な翻訳率
Table 5　Effective translation rate and ineffective translation rate.

| | 有効な翻訳率 | 無効な翻訳率 |
|---|---|---|
| 翻訳精度を用いた淘汰手法 | 50.0% | 50.0% |
| 帰納的学習を用いた淘汰手法 | 52.0% | 48.0% |

ここで，有効な翻訳は，以下の2つの条件のいずれかに該当する翻訳結果である。有効な翻訳率は，全翻訳数における，有効な翻訳の占める割合である。

① 未登録語を含まない正翻訳

② 未登録語を含むが未登録語に名詞句や形容詞などの単語を与えることにより容易に正翻訳が得られるもの

そして，無効な翻訳は，以下の3つの条件のいずれかに該当する翻訳結果である。無効な翻訳率は，全翻訳数における，無効な翻訳の占める割合である。

① 未登録語を含まない誤翻訳

② 未登録語を含み未登録語に名詞句や形容詞などの単語を与えても正翻訳が得られず誤翻訳となるもの

③ 入力文に対応する翻訳ルールが全く存在せずに翻訳不能となるもの

ここで，誤翻訳は，文として成立しない翻訳結果と，文としては成立するが訳として誤っている翻訳結果のいずれかである。

### 4.3 考察

#### 4.3.1 本手法の有効性

実験結果より，帰納的学習の導入が誤翻訳ルールに対する淘汰処理に有効であり，GA-ILMTを改善するために大きな役割を果たすことが確認できた。以下に帰納的学習を用いた淘汰手法の有効性について述べる。

(1)淘汰処理の精度向上について

従来の翻訳精度を用いた淘汰手法では，2.4項で述べたように2つの問題点を抱えていた。第1点は，淘汰される誤翻訳ルールが辞書中に存在する誤翻訳ルールに対し，ごく一部となっていることである。翻訳精度を用いた淘汰手法では，辞書中の全翻訳ルールに対し，評価の対象となる翻訳ルールは5.5%であった。図9に翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法における淘汰された誤翻訳ルール数の推移を示す。

図9より，翻訳精度を用いた淘汰手法では，生成される誤翻訳ルールに対し，淘汰処理が全く追い付いていないことがわかる。それに対し，帰納的学習を用いた淘汰手法では，誤翻訳ルールの増加に伴い，淘汰される誤翻訳ルールも増加していることが確認できる。帰納的学習を用いた淘汰手法において，翻訳例100組単位で淘汰された誤翻訳ルール数の占める割合を求めると，最初の100組は34.0%であったが，その後は常に55%～65%の割合で，生成される誤翻訳ルールに対する淘汰が行われている。したがって，帰納的学習を導入することにより，再現率を低下させることなく誤翻訳ルールを淘汰することが可能になったと考えられる。

また，第2点の問題点として，多くの単語数より構成されている誤翻訳ルールを淘汰することが困難であるという点が挙げられる。表6に，翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法のそれぞれの構成単語数別の再現率を示す。

構成単語数は誤翻訳ルール中の英単語数と日本語単語数の合計である。また，辞書中の翻訳ルールの平均構成単語数は6.9，辞書中の誤翻訳ルールの平均構成単語数は6.7であった。表6より，帰納的学習を用いた淘汰手法では，少ない単語数より構成されている誤翻訳ルールに偏ることなく，どのような構成単語数の誤翻訳ルールに対しても淘汰できていることが確認できる。また，表7に辞書中の正翻訳ルールの構成単語数の内訳を，表8には，実験に使用した翻

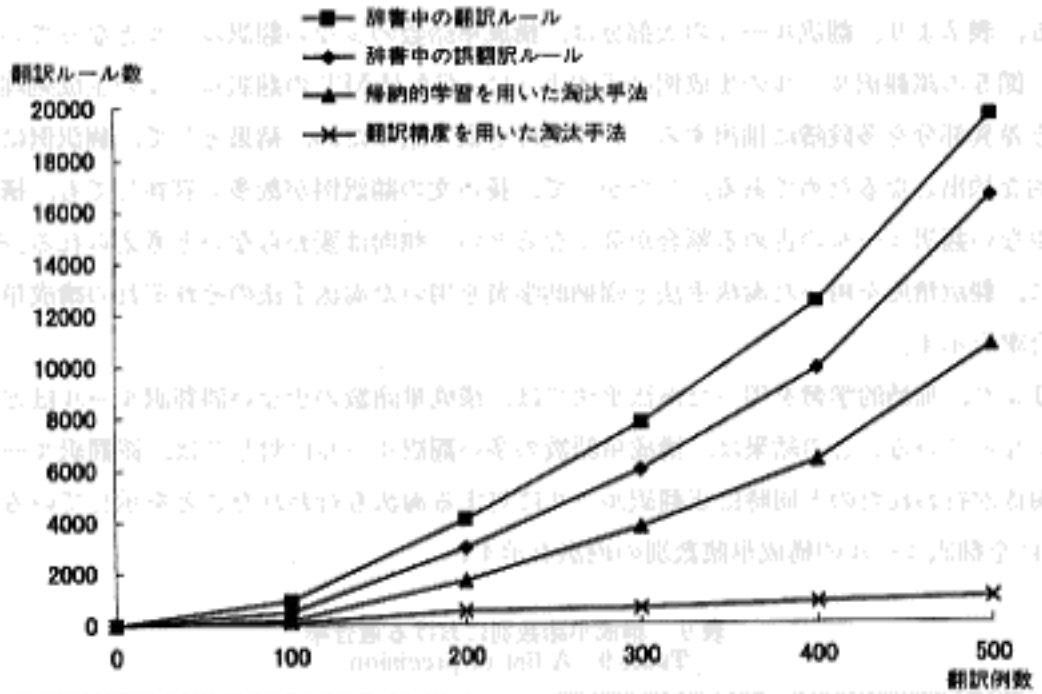


図9　淘汰された誤翻訳ルールの推移
**Fig. 9**　Changes in erroneous translation rules selected.

表6　構成単語数別における再現率
**Table 6**　A list of recall.

| 構成単語数 | 翻訳精度を用いた淘汰手法 | 帰納的学習を用いた淘汰手法 |
|---|---|---|
| 1～5 | 11.0%(922/8,354) | 60.1%(5,019/8,354) |
| 6～10 | 0.6%(41/6,433) | 66.5%(4,278/6,433) |
| 11～15 | 0.0%(0/1,435) | 82.0%(1,176/1,435) |
| 16～20 | 0.0%(0/235) | 96.6%(227/235) |
| 21～25 | 0.0%(0/2) | 100.0%(2/2) |

*(X/Y)*：Xは淘汰された誤翻訳ルール数，Yは辞書中の誤翻訳ルール数を示す．

表7　正翻訳ルールの構成単語数別の内訳
**Table 7**　A list of correct translation rules.

| 構成単語数 | 正翻訳ルール数 |
|---|---|
| 1～5 | 1,277 |
| 6～10 | 1,392 |
| 11～15 | 501 |
| 16～20 | 47 |
| 21～25 | 1 |

表8　翻訳例の構成単語数
**Table 8**　A list of translation examples.

| 構文単語数 | 翻訳例数 |
|---|---|
| 1～5 | 34 |
| 6～10 | 276 |
| 11～15 | 171 |
| 16～20 | 19 |

訳例の構成単語数の内訳を示す。

表6，表7より，翻訳ルールの大部分は，構成単語数の少ない翻訳ルールとなっている。これは，図5の誤翻訳ルールの生成例に示すように，GA-ILMTの翻訳ルールの生成処理が共通部分と差異部分を多段階に抽出するという処理を繰り返すため，結果として，翻訳例に対する部分的な抽出となるためである。したがって，長い文の翻訳例が数多く存在しても，構成単語数の少ない翻訳ルールの占める割合が高くなるという傾向は変わらないと考えられる。そして，表9に，翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法のそれぞれの構成単語数別の適合率を示す。

表9より，帰納的学習を用いた淘汰手法では，構成単語数の少ない誤翻訳ルールほど適合率は高くなっている。この結果は，構成単語数の多い翻訳ルールに対しては，誤翻訳ルールに対する淘汰が行われたのと同時に正翻訳ルールに対する淘汰も行われたことを示している。また，表10に全翻訳ルールの構成単語数別の内訳を示す。

表9　構成単語数別における適合率
Table 9　A list of precision.

| 構成単語数 | 翻訳精度を用いた淘汰手法 | 帰納的学習を用いた淘汰手法 |
|---|---|---|
| 1～5 | 89.3%(922/1,633) | 96.1%(5,019/5,224) |
| 6～10 | 85.4%(41/48) | 93.8%(4,278/4,563) |
| 11～15 | 0.0%(0/0) | 84.7%(1,176/1,388) |
| 16～20 | 0.0%(0/0) | 64.1%(227/354) |
| 21～25 | 0.0%(0./0) | 66.7%(2/3) |

"(X/Y)"：Xは淘汰された誤翻訳ルール数，Yは淘汰された翻訳ルール数を示す。

表10　全翻訳ルールの構成単語数別の内訳
Table 10　A list of all translation rules.

| 構成単語数 | 全翻訳ルール数 |
|---|---|
| 1～5 | 9,631 |
| 6～10 | 7,825 |
| 11～15 | 1,936 |
| 16～20 | 282 |
| 21～25 | 3 |

(2)帰納的学習を用いた淘汰手法の有効性について

次に，具体的に従来の翻訳精度を用いた淘汰手法では淘汰されなかった誤翻訳ルールがどのように淘汰されたのかについて述べる。翻訳精度を用いた淘汰手法において淘汰されない誤翻訳ルールは，その原因に基づき以下の2つに分類することができる。

① 翻訳処理に全く使用されない場合

② 翻訳処理に使用されたが淘汰の対象とならない場合

①の翻訳処理に全く使用されない場合は，適応度を求めるための式(1)における全使用回数が0であり，全く翻訳処理に使用されない誤翻訳ルールである。②の翻訳処理に使用されたが淘汰の対象とならない場合は，翻訳処理に使用されてはいるが，全使用回数が淘汰対象の条件で

ある5を下回るため，淘汰されない誤翻訳ルールである。表11に，この2つの分類の内訳とそれぞれの具体例，そして，それらの誤翻訳ルールが本手法により，どのように淘汰されたのかを示す。

表11　帰納的学習を用いた淘汰手法の有効性
**Table 11**　Effectiveness for method of selected using inductive learning.

| 分類 | 原因 | 割合 | 例 | |
|---|---|---|---|---|
| | | | 誤翻訳ルール | 誤翻訳ルールに対する淘汰 |
| ① | 翻訳処理に全く使用されない場合 | 82.2% | (Is that your good friend? ;彼/は/あなたの/姉さん/ですか?) | 帰納的学習により誤翻訳ルール中の誤り(that;彼)と(good friend;姉さん)を抽出，その結果，12.0%(3/25×100;誤り率)≧5.0%(閾値)となり辞書中から淘汰 |
| ② | 淘汰の対象とならない場合 | 17.8% | (your classmate ;僕の/クラスメイト) | 帰納的学習により誤翻訳ルール中の誤り(your;僕の)を抽出，その結果，25.0%(1/4×100;誤り率)≧5.0%(閾値)となり辞書中から淘汰 |

帰納的学習を導入することにより淘汰された誤翻訳ルールの占める割合は，翻訳処理に全く使用されない場合では67.7%，翻訳処理に使用されたが淘汰の対象とならない場合では52.2%であった。また，表11の2つの分類において，翻訳処理に全く使用されない場合が80%以上を占めていた。したがって，GA-ILMTの淘汰処理の精度を向上させるためには，このような翻訳処理に全く使用されない誤翻訳ルールを淘汰することが重要となる。本論文で提案する帰納的学習を導入することにより，翻訳精度に依存することなく，誤翻訳ルール中の誤りに基づき誤翻訳ルールを淘汰することが可能となった。

(3)誤翻訳ルールの判定について

本論文で提案する帰納的学習は，生成された誤翻訳ルール中に含まれている英文とその日本語訳文の対応関係における誤りを見つけ出し，辞書中から淘汰していくものである。したがって，本手法を評価する際には，誤翻訳ルールを淘汰できたかどうかということと同時に，正確に誤翻訳ルール中の誤りを抽出できているかどうかが重要となる。そこで，帰納的学習により誤翻訳ルールであると判断されたものの中で，その判断が誤翻訳ルール中に含まれている誤りを正確に抽出できたことによるものなのかどうかについて調査を行った。誤りの抽出対象となる淘汰された誤翻訳ルールは，その原因から以下の3つに分類することができる。

① 英文とその日本語訳文の対応関係の誤りのみが正確に抽出された誤翻訳ルール
② 英文とその日本語訳文の対応関係の誤りではない部分が抽出された誤翻訳ルール
③ 英文とその日本語訳文の対応関係の誤りと，誤りではない部分の両方が抽出された誤翻訳ルール

表12にこれら3つの分類に対する内訳とそれぞれの具体例を示す。

表12　帰納的学習を用いた誤翻訳ルールの判定
Table 12　Decision of erroneous translation rules using inductive learning.

| 分類 | 原因 | 割合 | 例 | |
|---|---|---|---|---|
| | | | 淘汰された誤翻訳ルール | 抽出結果 |
| ① | 誤りの抽出 | 71.9% | (I play tennis.<br>；私/は/テニス/が/好き/です。) | 正しく抽出された誤り<br>(play；が/好き/です) |
| ② | 正しい部分の誤抽出 | 1.4% | (do not like basketball<br>；が/好き/ではないのです) | 正しい部分の誤抽出<br>(do not；ではないのです) |
| ③ | 誤りの抽出と正しい部分の誤抽出の同時抽出 | 26.7% | (Kayo is not my classmate.<br>；加代/は/私の/クラスメイト/です。) | 正しく抽出された誤り<br>(not；です)<br>正しい部分の誤抽出<br>(Kayo；私) |

表12より，正しい部分の誤抽出のみを行い，その結果，誤翻訳ルールであると判断してしまう割合は1.4%であった。したがって，本論文で提案する帰納的学習が翻訳ルール中に含まれている誤りを抽出したうえで誤翻訳ルールを淘汰できていることを確認できた。表12に示すような誤翻訳ルール中の誤りは，それまでに与えられた翻訳例の英文とその日本語訳文の対応関係に基づいて抽出される。例えば，表12に示す誤翻訳ルール(I play tennis.；私/は/テニス/が/好き/です。)においては，(play；が/好き/です)の対応関係が誤りと判断されることにより誤翻訳ルールであると位置付けられた。これは，与えられた翻訳例に基づき，この誤翻訳ルールを評価した結果，英文では「play」が存在しているのにもかかわらず，「が好きです」が存在していることが誤りとして学習されたことに相当する。このように，帰納的学習を用いた淘汰手法では，それまでに与えられた翻訳例を活用することにより，解析的な知識を初期条件として与えることなく，自動的に誤翻訳ルール中に含まれている誤りを抽出し，辞書中から淘汰することが可能となる。

(4)意味的な誤りを含む誤翻訳ルールについて

本論文において，淘汰の対象となる誤翻訳ルールを，英文とその日本語訳文の対応関係が成立していない誤翻訳ルールとしている理由は，本論文の図5に示すように，それら自身が原因となり新たな誤翻訳ルールを生成するという点において，多大な悪影響を及ぼす存在となっているためである。我々は，英文とその日本語訳文の対応関係が成立しない誤翻訳ルールに対してだけではなく，意味的な誤りを含む誤翻訳ルールに対しても，同様のアプローチで解決していくことが可能であると考えている。その理由としては，本手法より，意味的に不自然な部分が，過去に与えられた翻訳例を参照することにって確率的に低くなり，誤りとして抽出されるためである。そこで，表13に帰納的学習を用いた淘汰手法における意味的な誤りを含む誤翻訳ルールの淘汰の精度を示す。

また，淘汰された意味的な誤りを含む誤翻訳ルールの具体例を図10に示す。

表13　意味的な誤りを含む誤翻訳ルールの淘汰の精度
Table 13　Precision of selection for erroneous translation ruleswhich have errors in meaning.

| | 精　度 |
|---|---|
| 変数なし | 70.2%(158/225) |
| 変数あり | 100.0%(4/4) |
| 合　計 | 70.7%(162/229) |

変数なし
（He is my mother.；彼/は/私の/母/です。）
（He is my classroom.；彼/は/僕の/教室/です。）
変数あり
（I am not my @0.；僕/は/僕の/@0/ではありません。）
（@0 name is two years old.；@0/名前/は/2/歳/です。）

図10　淘汰された意味的な誤りを含む誤翻訳ルールの具体例
Fig. 10　Examples of erroneous translation rules selected which have errors in meaning

表13より，意味的な誤りを含む誤翻訳ルールに対し，約70%の精度で淘汰できていることが明らかとなった。これは，英文とその日本語訳文の対応関係においては誤りが存在しなくとも，過去に与えられた翻訳例より，意味的に不自然な単語の組合せを抽出することができたためである。例えば，図10の中の誤翻訳ルール（He is my classroom.；彼/は/僕の/教室/です。）においては，（He；教室）と（classroom；彼）の単語の組合せが過去に与えられた翻訳例中に存在せず，誤りとして抽出された。これらの抽出は，過去に与えられた翻訳例では，英単語の「He」と日本語単語の「教室」，そして，「classroom」と「彼」の組合せが，翻訳例中の英文とその日本語訳文に同時に出現するのは，不自然であるということを学習したことに相当する。このように，本手法では，意味的な誤りを含む誤翻訳ルールに対しても，比較的高い精度で淘汰することが可能となることを確認できた。

(5)翻訳結果について

表5に示すように，有効な翻訳率は，翻訳精度を用いた淘汰手法と比べ，50.0%から52.0%の向上であった。これは，実験データ500文の翻訳において，無効な翻訳から有効な翻訳となったものが11文，有効な翻訳から無効な翻訳になったものが1文，結果として，有効な翻訳が10文増加したことに相当する。実験データに用いた500文は，図2に示すGA-ILMTのシステム構成に基づいて1文ずつ繰り返し行われた翻訳と学習の両方に使用している。図11に無効な翻訳から有効な翻訳へ移行した具体例を示す。また，図12には有効な翻訳から無効な翻訳へ移行した具体例を示す。

GA-ILMTでは，複数の翻訳結果が生成される場合には，生成された翻訳結果の上位1位から10位の間に有効な翻訳が存在する場合，その翻訳は有効な翻訳であるとしている[3]。した

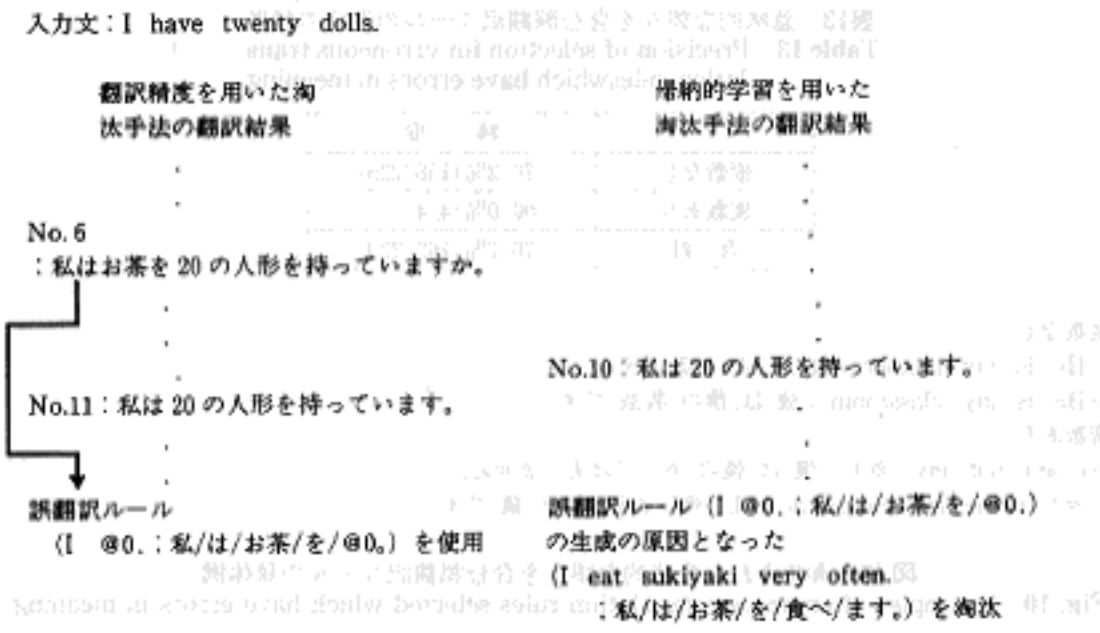


図 11　有効な翻訳結果の例
Fig.11　An example of effective translation result.

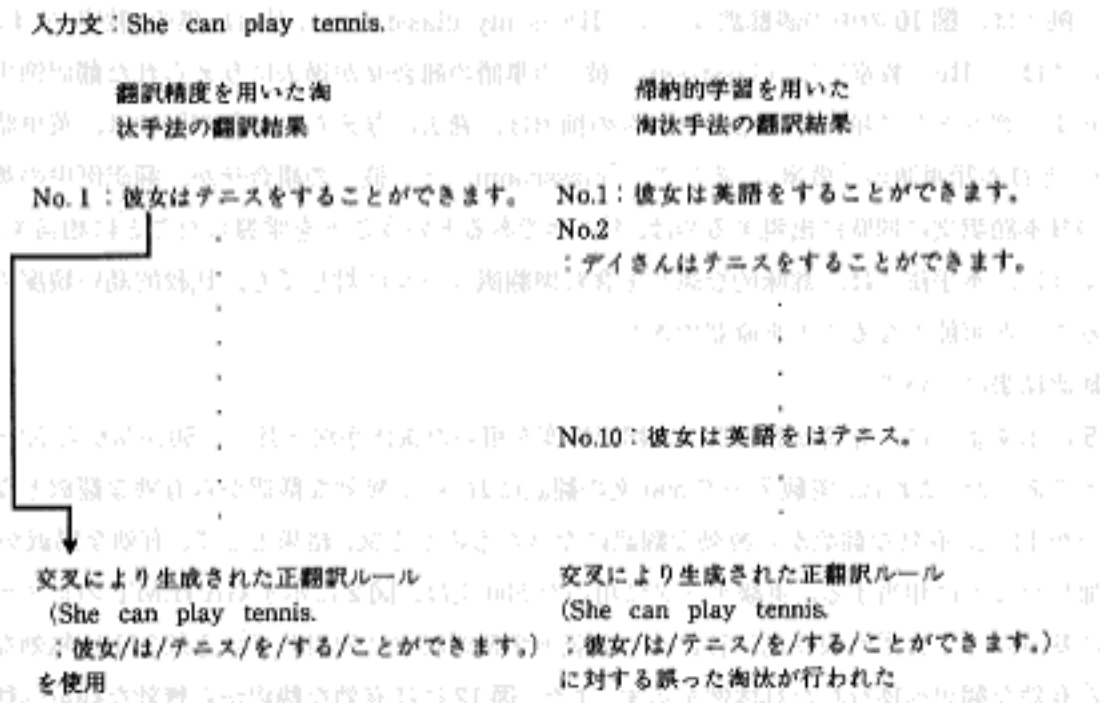


図 12　無効な翻訳結果の例
Fig. 12　An example of ineffective translation result.

がって，それまで10位以内に有効な翻訳が存在せずに無効な翻訳となっていたものが，誤翻訳ルールの減少により，10位以内に順位を上げ，無効な翻訳から有効な翻訳になった。実験の結果，複数の翻訳結果が得られたものの中で，その62.8%が順位を上げていた。それに対し，順位を下げた翻訳結果は，6.4%であった。また，上位1位のみを有効な翻訳とした場合には，有効な翻訳率は，39.4%から41.6%の向上となった。さらに，誤翻訳は翻訳精度を用いた淘汰手法に比べ，56.0%減少した。誤翻訳の大幅な減少が有効な翻訳率の向上をもたらさなかったのは，その大部分が11位以下に存在している誤翻訳の減少となったためである。

(6)淘汰処理の精度向上と有効な翻訳率の関係

図13に従来の翻訳精度を用いた淘汰手法と本論文で提案する帰納的学習を用いた淘汰手法における，淘汰処理後に辞書中に残された誤翻訳ルールの占有率の推移を示す。

占有率は，以下の式により得られる。

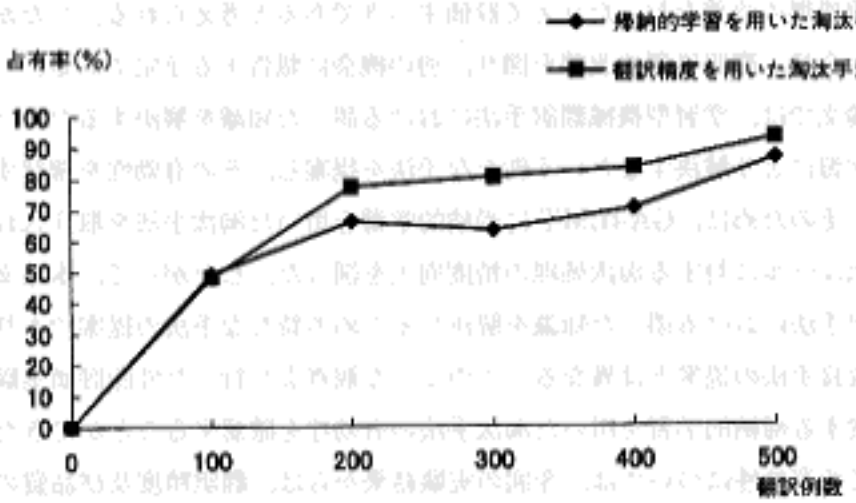


図13　淘汰後の誤翻訳ルールの占有率の推移
Fig. 13　Changes of erroneous translation rules rates in dictionary after selection.

$$\text{占有率}\ (\%) = \frac{\text{淘汰後の辞書中の全誤翻訳ルール数}}{\text{淘汰後の辞書中の全翻訳ルール数}} \times 100.0$$

図13より，翻訳例数200組から300組の間では，帰納的学習を用いた淘汰手法の誤翻訳ルールの占有率は，66.0%から63.1%に減少している。したがって，誤翻訳ルールの占有率は，翻訳例数の増加に比例して，必ずしも増加していくわけではない。誤翻訳ルールの占有率が増加している100組から200組の翻訳例は，それまでのbe動詞の文に，新たに一般動詞の文が加わっている。それに対し，誤翻訳ルールの占有率が減少している200組から300組の翻訳例には，100組から200組の翻訳例で学習した一般動詞の文が数多く存在している。したがって，翻訳例数100組から200組の間の誤翻訳ルールの占有率は，新たな文法事項が加わったことによ

り，学習が不十分となり増加したと考えられる。また，図 14 には，従来の翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法における有効な翻訳率の推移を示す。

図 13 と図 14 より，誤翻訳ルールの占有率が異なる場合でも，有効な翻訳率には大きな違いが見られなかった。図 14 において，誤翻訳ルールの占有率の差が最大となる翻訳例数 200 組から 300 組の間の有効な翻訳率は，翻訳精度を用いた淘汰処理の場合で 64.1%，帰納的学習を用いた淘汰手法の場合では 62.1%となり，わずかな減少となっている。したがって，今回の実験結果においては，誤翻訳ルールの占有率の増加に伴って，有効な翻訳率が大幅に減少するという現象は見られなかった。また，この結果から，誤翻訳ルールの占有率が減少しても，有効な翻訳率は大幅には向上しないということが推測される。しかし，この原因は本手法の導入による翻訳ルールの精度向上を有効な翻訳率に反映させるような，翻訳処理に対する改善が行われていないことにある。このような点から，誤翻訳ルールの占有率と有効な翻訳率との関係については，翻訳処理の改善を行ったうえで評価すべきであると考えられる。したがって，この点については，今後，翻訳処理の改善を図り，別の機会に報告する予定である。

また，本論文では，学習型機械翻訳手法における誤った知識を解決するために，システム自身が帰納的学習により解決するという新たな手法を提案し，その有効性を確認することを目的としている。そのために，GA-ILMT に帰納的学習を用いた淘汰手法を取り入れ，誤った知識である誤翻訳ルールに対する淘汰処理の精度向上を図った。したがって，本論文の目的は，学習型機械翻訳手法における誤った知識を解決するための新たな手法の提案であり，機械翻訳手法としての改良手法の提案とは異なる。このような観点より行った性能評価実験においては，本論文で提案する帰納的学習を用いた淘汰手法の有効性を確認することができた。機械翻訳システムとしての有効性については，今回の実験結果からは，翻訳精度及び品質の大幅な向上は

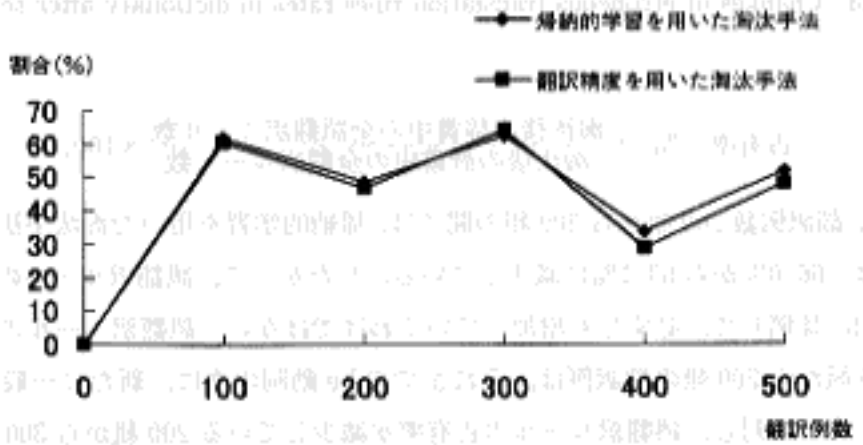


図 14　有効な翻訳率の推移
Fig. 14　Changes of effective translation rates.

見られなかったが，多くの誤翻訳ルールが淘汰されたことにより，誤翻訳結果数が56.0%減少した。すなわち，本手法は，GA-ILMTの淘汰処理を行うフィードバック部のみの精度向上をもたらすための手法として位置付けられる。このような観点より，機械翻訳手法としての有効性については，本論文の目的が誤った知識を解決するための手法を提案することにあるため，翻訳処理に対する改良は行っておらず，ほぼ従来の翻訳ができるだけとなった。しかし，GA-ILMTは，従来の解析型機械翻訳手法を超える可能性を持っており，そのことは，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 [3] で，その見通しについて述べている。また，長い複雑な構造を持った文に対する翻訳については，膨大な処理時間を必要とするため，現時点では，物理的に評価実験が困難である。しかし，理論的には，翻訳ルールの変数部分に対して，単文や文節を代入することにより翻訳可能になると考えられる。省略を含んでいる文などの文脈に依存した文の翻訳については，旅行用英会話文を用いた性能評価を通して，解析型機械翻訳手法との比較 [21, 22] を行っているが，今後，更に検討を進め，別の機会に報告する予定である。

(7)処理時間について

帰納的学習を用いた淘汰手法により，処理時間が大幅に減少した。図15に翻訳精度を用いた淘汰手法と帰納的学習を用いた淘汰手法のそれぞれの処理時間を示す。

処理時間は，翻訳精度を用いた淘汰手法と比べ，全体で52.1%減少した。これは，誤翻訳ルールが減少したことにより，翻訳処理において翻訳結果生成のための誤翻訳ルールを用いた組合せ処理が減少したためである。また，誤翻訳ルールを生成する原因となった親となる誤翻訳ルールが数多く淘汰されたために，それらの誤翻訳ルールを用いた2次的な誤翻訳ルールの生成処

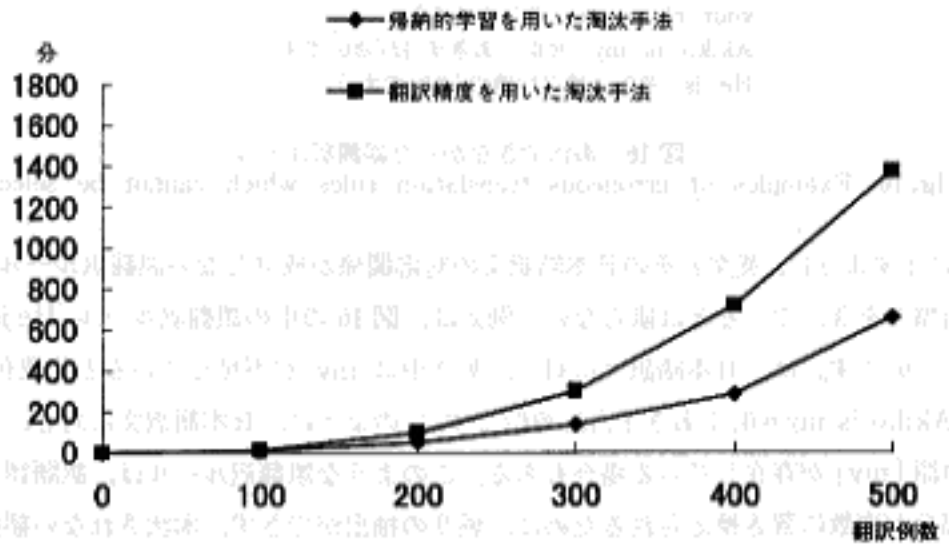


図15　処理時間の推移
Fig. 15　Changes in CPU times.

理が大幅に減少したことも大きな原因である。誤翻訳ルールの生成の原因となった親の誤翻訳ルールが淘汰され，その生成を未然に防ぐことのできた誤翻訳ルールは，淘汰された誤翻訳ルールの 73.7%を占めていた。また，図 15 の処理時間は，翻訳時間と学習時間の合計である。翻訳時間においては，他の実例型機械翻訳手法のうち，古瀬らの手法 [6] が，その詳細を述べている。古瀬らの手法では，平均単語数 9.2 語の 825 文に対し，翻訳時間を計測した結果，平均 1.9 秒が得られたと示されている。今回の実験では，英文の平均単語数は 4.2 語であり，翻訳時間と学習時間を合わせた処理時間の平均は，翻訳精度を用いた淘汰手法では約 2 分 45 秒，帰納的学習を用いた淘汰手法では約 1 分 18 秒となった。処理時間については，本手法を導入したことの結果として，大幅に減少したことを述べている。したがって，本論文では，絶対的な時間の短縮を図るための工夫は行っていない。処理時間の短縮を図るためには，翻訳ルールの階層化を行い，効率的な翻訳ルールの組合せ処理を取り入れることなどが考えられる。

### 4.3.2 問題点

実験結果から，帰納的学習を用いた淘汰手法が GA-ILMT の淘汰処理の精度向上を可能にすることを確認できた。しかし，有効性の確認と共に，いくつかの問題点も明らかとなった。以下に帰納的学習を用いた淘汰手法の問題点について述べる。

(1)淘汰できなかった誤翻訳ルールについて

表 4 より，帰納的学習を用いた淘汰手法の再現率は 65.0%であった。したがって，辞書中に存在する誤翻訳ルールの 35.0%を淘汰することができなかったことになる。図 16 に淘汰できなかった誤翻訳ルールの具体例を示す。

(sister；私の/姉)
(your classmate；クラスメイト)
(Akiko is my @0.；あき子/は/@0/です。)
(He is @0.；彼/は/僕の/@0/です。)

図 16　淘汰できなかった誤翻訳ルール
**Fig. 16** Examples of erroneous translation rules which cannot be selected.

図 16 に示すように，英文とその日本語訳文の対応関係が成立しない誤翻訳ルールは，その全てが訳語誤りを含んでいるとは限らない。例えば，図 16 の中の誤翻訳ルール (He is@0.；彼/は/僕の/@0/です。)は，日本語訳文に対し，英文中に「my」が不足していると位置付けられる。また，(Akiko is my@0.；あき子/は/@0/です。) のように，日本語訳文に対し，英文中に不必要な単語「my」が存在している場合もある。このような誤翻訳ルールは，訳語誤りの存在している部分が変数に置き換えられるために，誤りの抽出ができず，淘汰されない誤翻訳ルールとなる。図 17 にその具体例を示す。

図 17 に示すように，[2] の誤翻訳例においては，(your；僕の) が訳語誤りとなっているた

与えられた翻訳例
[1] (He is my teammate.；彼/は/僕の/チーム仲間/です。)
生成された誤翻訳例
[2] (He is your classmate.；彼/は/僕の/クラスメイト/です。)

↓

差異部分
[3] (my teammate；チーム仲間)
[4] (your classmate；クラスメイト)
共通部分
[5] (He is @0.；彼/は/僕の/@0/です。)

図17　淘汰できなかった誤翻訳ルールの生成例
Fig. 17　Examples of erroneous translation rules which cannot be selected.

め，[1] の与えられた翻訳例と比較した場合，日本語訳文においては「クラスメイト」が差異部分となるが，英文においては「your classmate」が差異部分となり，変数に置き換えられてしまう。したがって，生成された [5] の翻訳ルールからは「your」が存在しなくなり，訳語誤りが消去されてしまう。このような誤翻訳ルールを本手法により評価すると，生成された [5] の誤翻訳ルールの変数を除いた英文「He is」とその日本語訳文「彼/は/僕の/～/です。」共に，[1] の与えられた翻訳例と一致するため，誤りが含まれていないと判断される。このように，変数を含む誤翻訳ルールの多くは誤りの抽出が困難となり，辞書中に残ったままとなる。表14に変数別の適合率と再現率を示す。

表14より，変数を含む誤翻訳ルールに対する再現率が低いことを確認できる。また，適合率においても，変数を含む誤翻訳ルールが低いことを確認できる。これは，親となる正翻訳ルールが誤って淘汰されたために，生成されなくなった変数を含む正翻訳ルールが存在したためである。

(2)淘汰された正翻訳ルールについて

帰納的学習を用いた淘汰手法の適合率は92.8%であった。したがって，正翻訳ルールであるのにもかかわらず，淘汰されたものは7.2%存在したことになる。その結果，淘汰された正翻訳

表14　適合率と再現率の内訳
Table 14　A list of precision and recall.

| | 適合率 | 再現率 |
|---|---|---|
| 変数なし | 93.4%(8,083/8,654) | 71.4%(8,083/11,313) |
| 変数1つ | 91.7%(2,056/2,243) | 50.4%(2,056/4,083) |
| 変数2つ | 89.7%(506/564) | 51.7%(506/978) |
| 変数3つ | 80.3%(57/71) | 67.9%(57/84) |
| 変数4つ | 0.0%(0/0) | 0.0%(0/1) |
| 合　計 | 92.8%(10,702/11,532) | 65.0%(10,702/16,459) |

ルールは，辞書中に存在する正翻訳ルールの25.8%であった。正翻訳ルールが淘汰された最大の原因は，正しい部分の誤抽出が行われ，誤りを含んでいると判断されたためである。例えば，正翻訳ルール（He is a student.；彼/は/生徒/です。）を，本手法により評価した結果，与えられた翻訳例中に，(He；生徒)，(a；彼) そして，(student；彼) の3つの単語の組合せが存在していなかった。しかし，(He；生徒) と (student；彼) は組合せ誤りと判断され，結果的に，(a；彼) が翻訳ルール中の誤りと判断された。この場合，(a；彼) の「a」は日本語訳文に訳語は存在しないが，翻訳ルール中における誤りではない。しかし，組合せ誤りを消去する際に，1単語間で組合せの変更を行っているために (a；彼) が残ってしまい，翻訳ルール中の誤りと判断された。また，正翻訳ルール(Andy is your sister.；アンディ/は/あなたの/姉さん/です。)に対しては，(Andy；姉さん)と(sister；アンディ)が誤っていると判断された。これは，個々の英単語と日本語単語がそれぞれ翻訳ルール中に対応する単語を含んでいるため，組合せ誤りとなるべきであるが，(Andy；アンディ) と (sister；姉さん) の中の (Andy；アンディ) が辞書中に翻訳ルールとして登録されていなかったため，(Andy；姉さん) と (sister；アンディ) が翻訳ルール中の誤りと判断された。このように，正しい部分の誤抽出が原因となり，正翻訳ルールが淘汰された。その他の原因としては，正翻訳ルールの生成の原因となった親となる正翻訳ルールが淘汰されたことが挙げられる。したがって，帰納的学習を導入することにより淘汰された正翻訳ルールは，その原因に基づき以下の2つに分類することができる。

① 翻訳ルール中の誤りではない部分が抽出された正翻訳ルール

② 翻訳ルールを生成する原因となった親となる正翻訳ルールが淘汰されたため生成されなくなった正翻訳ルール

表15に，これら2つの分類に対する内訳を示す。

表15より，淘汰された正翻訳ルールの最大の原因は，翻訳ルール中の誤りではない部分に対する誤抽出であることが明らかとなった。このような正翻訳ルールに対する誤った判断を防ぐ

表15　淘汰された正翻訳ルールの内訳
**Table 15**　A list of correct translation rules selected.

| 分類 | 原因 | 割合 | 例 | |
|---|---|---|---|---|
| | | | 淘汰された誤翻訳ルール | 抽出結果 |
| ① | 正しい部分の誤抽出 | 64.0% | (Andy is your sister.<br>：アンディ/は/あなたの/姉さん/です。) | 正しい部分の誤抽出<br>(Andy：姉さん)<br>(sister；アンディ) |
| ② | 生成されなくなった正翻訳ルール | 36.0% | (Kayo your：加代/は/きみの) | 生成の原因となった親となる正翻訳ルール<br>[1] (my brother；私の/兄)<br>[2] (Kayo your brother<br>：加代/は/きみの/兄)<br>の [2] の正翻訳ルールを本手法により淘汰 |

ためには，組合せ誤りを消去する際の組合せの変更を，常に，1単語のみで行うのではなく「a student」のように複数の単語に対しても行うことや，確率的な手法を用いて誤りを決定することなどが考えられる。

### 4.3.3　翻訳例の増加に伴う問題点とその解決法

本手法では，翻訳例が増加することによって，過去に与えられた翻訳例中に，(not；です)という単語の組合せを含む翻訳例が存在した場合，本論文の3.2項に示す誤翻訳ルール(Akiko is not my teammate.；あき子/は/僕の/チーム仲間/です。)から誤りとして(not；です)を抽出できずに，淘汰が困難になる。これは，本手法では，判定の対象となる単語の組合せが過去に与えられた翻訳例中に一度でも存在していた場合，誤った単語の組合せである可能性が全くないと判断されるためである。今回の実験では，単語数の少ない単文のみから構成されている翻訳例を用いていることから，(not；です)という単語の組合せを持つ翻訳例が出現する確率は非常に低い。したがって，過去に一度でも存在した単語の組合せは誤っている可能性が全くないとする方法を用いた。上記のような問題点に対しては，誤りの抽出を確率的に行う手法を取り入れることにより解決できると考えられる。以下に，確率的に誤りを抽出する手法について検討した結果を述べる。具体的には，誤翻訳ルール中の単語の組合せが過去の翻訳例中にどれだけの頻度で出現したのかにより判断する方法が考えられる。例えば，以下の式を用いて出現率を求める。

$$\text{出現率}\ (\%) = \text{単語の組合せの出現回数}\quad \text{英単語の出現回数} \times 100.0 \qquad (7)$$

そして，出現率の高い単語の組合せをAPとし，出現率の低い単語の組合せをNAとする。ここで，翻訳例に，(not；です)を含む正しい翻訳例として，(Not Akiko but Kayo is my classmate.；あき子/ではなく/加代/が/私の/クラスメイト/です。)を追加し，誤翻訳ルール(Akiko is not my teammate.；あき子/は/僕の/チーム仲間/です。)に対する出現率を求めた。その結果を表16に示す。

表16より，(not；です)の単語の組合せの出現率は8.3%と低いものとなった。また，表16には，表1においてNPとなった単語の組合せ以外にも出現率の低い単語の組合せがいくつか

表16　出現率の一覧
**Table 16**　A list of rate of appearance.

| | Akiko | is | not | my | teammate |
|---|---|---|---|---|---|
| あき子 | 37.5 | 8.1 | 16.7 | 10.5 | 0.0 |
| は | 100.0 | 97.3 | 100.0 | 100.0 | 100.0 |
| 僕の | 0.0 | 16.7 | 25.0 | 31.6 | 100.0 |
| チーム仲間 | 0.0 | 8.1 | 8.3 | 15.8 | 100.0 |
| です | 75.0 | 75.7 | 8.3 | 63.2 | 66.7 |

存在している。しかし，このような単語の組合せは，学習の進行に伴い，翻訳例数が増加するにしたがって，徐々に高くなると考えられる。

また，長文の翻訳例が増加すると，上述した確率的に誤りを抽出する手法を用いても，正確に誤りの抽出を行うことが困難になると考えられる。この問題に対しては，誤翻訳ルールの誤りを抽出する際に使用する翻訳例を比較的短い文の翻訳例とし，他の翻訳例とは区別して利用することなどが考えられる。

## 5．おわりに

本論文では，翻訳例から知識を自動抽出する際に生じる，誤った知識に対する解決策として，与えられた翻訳例より帰納的学習を用いて消去する手法を提案した。そして，その有効性を GA-ILMT の淘汰処理に導入することにより述べた。GA-ILMT では，これまでに誤った知識として多くの誤翻訳ルールが生成されていた。その誤翻訳ルールに対する淘汰手法として，従来の淘汰手法では，翻訳に使用された翻訳ルールのみが翻訳精度に基づいて淘汰されていた。しかし，このような淘汰手法では，翻訳に使用されない多くの誤翻訳ルールが淘汰の対象にならずに辞書中に残ったままとなる。この問題点に対し，我々は，解析的な知識を初期条件として与えるのではなく，ロバストネスを保持するため，与えられた翻訳例を用いた帰納的学習を導入することによりその解決を図った。実験の結果，適合率は 89.1%から 92.8%，再現率は 5.9%から 65.0%に向上した。これらの大幅な改善は，帰納的学習を導入したことにより，翻訳精度に依存した淘汰ではなく，誤翻訳ルール中の誤りに基づいた淘汰が可能になったためである。さらに，その結果，誤翻訳は 56.0%，処理時間は 52.1%減少した。これらの実験結果より，誤った知識に対する解決策として，帰納的学習の導入が有効であり，学習型機械翻訳システムの改善に大きな役割を果たすことが確認できた。

今後は，より実用的な文章を翻訳するために，翻訳ルールの階層化などを進め，翻訳ルールに対する改良を行う。そして，より実用的な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。

## 6．謝　辞

本研究の一部は，文部省科学研究費補助金（第 10680367 号，第 09878070 号）及び北海学園大学ハイテク・リサーチ・センター研究費による補助のもとで行われた。

## 参考文献


[1] 荒木健治，栃内香次：帰納的学習による語の獲得および確実性を用いた語の認識，電子情報通信学会論文誌，Vol. J 75-D-II，No. 7，pp. 1213－1221 (1992)．

[2] 荒木健治，高橋祐治，桃内佳雄，栃内香次：帰納的学習を用いたべた書き文のかな漢字変換，電子情報通信学会論文誌，Vol. J 79-D-II，No. 3，pp. 391－402 (1996)．

[3] 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)．

[4] Echizen-ya, H., Araki, K., Momouchi, Y. and Tochinai, K. Machine Translation Method Using Inductive Learning with Genetic Algorithms. In *Proceedings of the Coling '96*, pages 1020－1023, Copenhagen, Denmark, August(1996).

[5] 佐藤理史：ＭＢＴ２：実例に基づく翻訳における複数翻訳例の組合せ利用，人工知能学会誌，Vol. 6，No. 6，pp. 861－871 (1991)．

[6] 吉瀬謙，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)．

[7] 野美山浩：事例の一般化による機械翻訳，情報処理学会論文誌，Vol. 34，No. 5，pp. 905－912 (1993)．

[8] 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)．

[9] 新納浩幸，井佐原均：語義の特異性を利用した慣用表現の自動抽出，情報処理学会論文誌，Vol. 36，No. 8，pp. 1845－1853 (1995)．

[10] 野村浩郷（編）：言語処理と機械翻訳，講談社 (1991)．

[11] 長尾真（編）：自然言語処理，岩波書店 (1996)．

[12] Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley (1989).

[13] 北野宏明：遺伝的アルゴリズム，産業図書 (1993)．

[14] 安居院猛，長尾智晴：ジェネティックアルゴリズム，昭晃堂 (1993)．

[15] 越前谷博，荒木健治，宮永喜一，栃内香次：遺伝的アルゴリズムを用いた帰納的学習による機械翻訳手法（GA-ILMT）における翻訳例を利用した構文処理の精度向上，情報処理学会研究報告，NL 117-8，pp.51－57 (1997)．

[16] Echizen-ya, H., Araki, K., Miyanaga, Y. and Tochinai, K. Improvement in Selection Process Based on Translation Examples of GA-ILMT. In *Proceeding of the IASTED International Conference*, pages 121－124, Banff, Canada, July－August (1997).

[17] Echizen-ya, H., Araki, K., Miyanaga, Y. and Tochinai, K. An Improvement of the Method for Removing Erroneous Translation Rules in GA-ILMT. In *Proceeding of the PACLING'97*, pages 105－112, Tokyo, Japan, September (1997).

[18] 越前谷博，荒木健治，宮永喜一，栃内香次：遺伝的アルゴリズムを用いた帰納的学習による機械翻訳手法の性能向上のための改良，1996年電子情報通信学会総合大会，D-54 (1996)．

[19] 内山智正，荒木健治，宮永喜一，栃内香次：帰納的学習による機械翻訳手法の評価実験，情報処理学会研究報告，NL 93-4，pp. 23－30 (1993)．

[20] 教科書ガイド教育出版版ワンワールド１，日本教材，東京 (1991)．

[21] 荒木健治，越前谷博，栃内香次：GA-ILMT の旅行用英会話文を用いた適応性能の評価，電子情報通信学会信学技報，NLC 96－63，pp. 53－60 (1997)．

[22] Araki, K., Echizen-ya, H. and Tochinai, K. Performance Evaluation in Travel English for GA-ILMT. In *Proceeding of the IASTED International Conference*, pages 117－120, Banff, Canada, July－August (1997).