自然言語処理 117－8
（1997. 1. 20）

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淘汰処理の精度向上

越前谷 博† 荒木 健治†† 宮永 喜一† 栃内 香次†

†北海道大学工学部電子情報工学専攻
††北海学園大学工学部電子情報工学科

E-mail echi@hudk.hokudai.ac.jp

我々は，高い学習能力を持ち，ユーザあるいは対象分野に速やかに適応していく機械翻訳システムの実現に向けての研究を行なっている．その第一段階として，我々は，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）を提案し，その性能評価を行なってきた．その結果，与えられた翻訳例をより効率的に活用した翻訳が可能となった．しかし，大きな問題点として，誤った翻訳ルールに対する淘汰処理が遅く，それらが辞書中から十分に削除されないことが挙げられる．この問題を解決するために，我々は，与えられた翻訳例を利用することにより，淘汰処理の精度向上を試みた．本稿では，その改良手法の提案及び有効性を確認するために行なった性能評価実験について述べる．

An Improvement in the Selection Process Utilizing Translation Examples of Machine Translation Using Inductive Learning with Genetic Algorithms

Hiroshi Echizen-ya† Kenji Araki†† Yoshikazu Miyanaga† and Koji Tochinai†

†Division of Electronics and Information, Hokkaido University
††Dept. of Electronics and Information, Hokkai-Gakuen University

E-mail echi@hudk.hokudai.ac.jp

The goal of our research is to realize a machine translation system that is capable of adapting to the environment. For this purpose, we proposed a method of machine translation using inductive learning with genetic algorithms, and we evaluated the method. We confirmed the system based on this method can translate by using the given translation examples effectively. However, the system produces many erroneous translation rules that cannot be completely removed from the dictionary. To resolve this problem, we improved the selection process by utilizing translation examples. In this paper, we describe an improvement in the selection process and the results of evaluation experiments.

1 はじめに

実用的な機械翻訳システムの開発に向け，多くの研究が行なわれている．しかし，精度および翻訳品質の面において，ユーザが十分満足できるシステムの実現には至っていない．現在の機械翻訳システムに取り入られている翻訳手法の主流となっているのは，解析型の機械翻訳手法 [1][2] である．これらの手法では，人手により文法や辞書を作成し，それらに基づき翻訳を行なっている．したがって，有限個の文法や固定化された辞書を用いるため，例外的な文章や未登録語に対処することが困難となる．これらの問題を解決する手法として，近年，コーパスを利用した学習型の機械翻訳手法の研究が盛んに行なわれている．このコーパスを利用した学習型の機械翻訳手法には，大量の翻訳例を準備し，入力文と最も類似した構造を持つ翻訳例を利用する手法 [3][4] と大量のコーパスから翻訳規則を自動的に抽出し，その規則を利用する手法 [5] [6] が提案されている．これらの手法の問題点として，精度および翻訳品質を向上させるために，膨大な量のコーパスを必要とすることが挙げられる．

我々は，言語および知識獲得といった人間の持つ生得的な能力を計算機上で実現することを目的とした研究 [7][8] を行なっている．そうした立場より，翻訳例のみから，翻訳ルールを自動的に獲得し，それらを用いて翻訳を行なう帰納的学習による機械翻訳手法の提案 [9] とその性能評価 [10] を行なってきた．我々は，この提案手法が解析型の機械翻訳手法の持つ例外的な文章や未登録語における対処の困難さを解決できると考えている．しかし，この帰納的学習による機械翻訳手法では，他の学習型の機械翻訳手法と同様に，良質な翻訳を行なうために，類似した大量の翻訳例が必要になるという問題点が明らかとなった．そこで，我々は，少量の翻訳例のみを用いて，この問題を解決するために，帰納的学習による機械翻訳手法へ遺伝的アルゴリズムを適用した．そして，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（Machine Translation Using Inductive Learning with Genetic Algorithms，以下，GA-ILMT と記す.）の有効性を確認してきた [11][12]．遺伝的アルゴリズムの適用により，翻訳例のみを効率良く使用でき，その結果，精度および翻訳の質を向上させることが可能となった．

しかし，有効性と共に，いくつかの問題点も明らかとなった．最も大きな問題点は多くの誤った翻訳ルールが抽出され，それらを辞書中から即座に消去できないことである．この誤った翻訳ルールに対する淘汰処理の遅れが，誤翻訳の生成と処理時間の増加をもたらしていた．これは，GA-ILMT の淘汰処理が不十分であることが原因となっている．そこで，我々は解析的な知識を用いることにより生じるロバストネスの問題を避けるために翻訳例のみを利用し，GA-ILMT の淘汰処理の精度向上を図った．本稿では，淘汰処理の精度向上のための改良手法の提案とその有効性を確認するために行なった評価実験及び考察結果について述べる．

2 処理過程

GA-ILMT に基づき構築した英日機械翻訳システムの処理過程を図 1 に示す．

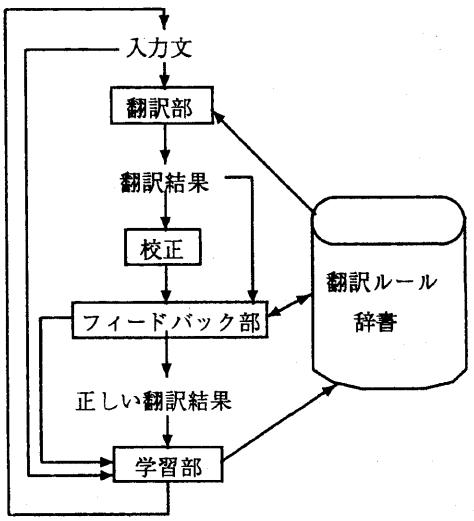

図 1: 処理過程

まず，入力文として英文を入力する．すると，

翻訳部において，それまでに抽出された翻訳ルールを使用し，最適な翻訳結果を生成する．その翻訳結果に対し，必要があれば人手による校正を行ない，正しい翻訳結果を与える．次いで，フィードバック部において，翻訳部で使用された各翻訳ルールに対し，適応度をそれぞれ決定し，その値に基づき淘汰処理を行なう．そして，学習部において，与えられた入力文と正しい翻訳結果からなる翻訳例と辞書中の翻訳例との間で交叉と突然変異を行ない，多様な翻訳ルールを抽出し，以後の翻訳処理に活用する [11].

3 淘汰処理の改良

3.1 従来の淘汰手法とその問題点

GA-ILMT の大きな問題点として，誤った翻訳ルールに対する淘汰処理の不十分さが挙げられる．GA-ILMT では，交叉や突然変異を行なうことにより多くの翻訳ルールを抽出する．その抽出された翻訳ルールには，多くの誤った翻訳ルールも含まれている [11]．そして，そのような誤った翻訳ルールは淘汰処理により，即座に消去されなければならない．しかし，従来の淘汰手法では，その精度が低いため，即座に消去できず，多くの誤った翻訳ルールが辞書中に残ったままとなっていた．図 2 に誤った翻訳ルールの具体例を示す．

(You like basketball. ; あなた/は/バレーボール/が/好き/ではないのです.)
(I don't like volleyball. ; 私/は/バスケットボール/が/好き/です.)
(You like @0. ; あなた/は/バレーボール/が/好き/@0.)
(volleyball ; です) , (basketball ; ではないのです)

図 2: 誤った翻訳ルール

図 2 に示すような誤った翻訳ルールに対し，これまでの淘汰手法では，適応度を決定し，その適応度の低い翻訳ルールを辞書中から消去していた [13]．適応度を求める際の計算式を以下に示す．

$$\text{適応度（\%）} = \frac{\text{正翻訳度数}}{\text{全翻訳度数}} \times 100 \qquad (1)$$

この適応度は，各翻訳ルールにそれぞれ設けられている．全翻訳度数は，それまでの翻訳において，その翻訳ルールが何回使用されたかを表している．正翻訳度数は，使用された中で，正翻訳を生成するために何回使用されたかを表している．したがって，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淘汰処理の対象となっていた．しかし，辞書中に存在する翻訳ルールの多くは使用されない．その結果，抽出された翻訳ルールに対する評価が非常に遅れ，誤った翻訳ルールを辞書中から十分に消去することができなかった．また，このような誤った翻訳ルールが辞書中に残ることにより，それを用いたさらなる翻訳ルールの抽出が行なわれ処理時間を増加させる原因となっている．

3.2 改良手法

遺伝的アルゴリズムでは，交叉や突然変異を行なうことにより，多種多様な個体を作り出す．そして，淘汰処理において，環境に適するものは残り，適さないものは消去される．GA-ILMT においても交叉や突然変異を行なうことにより，多くの翻訳ルールを生成している．その中には，正しいものも誤ったものも存在する．そして，誤った翻訳ルールに対しては，淘汰処理を行なうことにより，その正誤を判定し，消去していく．したがって，誤った翻訳ルールを即座に消去するためには，淘汰処理の精度を向上させなければならない．また，我々は淘汰処理を改良するにあたり，解析的な知識をシステムに取り込むことは，解析型の機械翻訳手法の持つ例外的な文章や未登録語に対処することが困難であるというロバストネスの問題を抱え込むことになると考える．したがって，本稿で述べる改良手法では，解析的な知識を使用していない．我々は，学習機能をシステムに取り込むことで，例外的な文章や未登録語に対し，適応的に対処することが可能になると考えている．

淘汰処理では，各々の個体が環境に適応するかどうかを判定し，適応しないものは消去していく．GA-ILMTにおいて環境は翻訳例となる．そこで，この翻訳例の英文とその日本語訳文を構成している単語の組合せから，翻訳ルールを構成している単語の組合せが，翻訳例中に存在しているかどうかを調べる．存在していれば，その単語の組合せは確率的に正しいものと考えられる．例えば,"（I；私）"といった単語の組合せは，いずれかの翻訳例中に含まれている可能性が高く，確率的に正しい組合せと位置付けられる．それに対し"（volleyball；バスケットボール）"といった単語の組合せは翻訳例中に含まれる可能性が低く，訳語として誤ったものであると位置付けられる．その結果,"（volleyball ； バスケットボール)" の組合せを含んだ翻訳ルールの正当性は低くなる．このように，生成された翻訳ルールに対して，過去に与えられた翻訳例を直接的に参照することにより，その翻訳ルールがどの程度誤ったものなのかを評価する．以下にその具体的な手順を示す．

（1）翻訳ルールに対する評価

①生成された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組合せを取り出す．

②取り出された組合せが，与えられた翻訳例中に含まれているかどうかを調べる．

③いずれかの翻訳例中に存在していた場合，正しい組合せとしてtrueを与える．全く存在していない場合には，誤った組合せとしてfalseを与える．

④falseとなった単語の組合せ中に組合せ誤りがあるかどうかを検索し，除去する．

・falseとなった全ての組合せを取り出し，その組合せ同士で，英単語と日本語単語の組合せを変更する．

・変更された単語の組合せが，辞書中の翻訳ルールとして存在しているかどうかを検索する．

・存在している場合，組合せ誤りであるとして，変更前の単語の組合せをfalseからtrueに変更する．

⑤①～④の結果に従い，以下の計算式を用い，翻訳ルールの誤り率を求める．

$$\text{誤り率（\%）} = \frac{\text{誤った組合せの数}}{\text{組合せ総数}} \times 100 \quad (2)$$

（2）（1）で求めた誤り率に対して，閾値xを設け，以下の条件を満たすものを誤った翻訳ルールとして辞書から削除する．

$$\text{誤り率} \geq x \quad (3)$$

この手法において（1）の④を行なう理由は，falseとなる原因として，訳語誤りと組合せ誤りの2種類があり，その中の組合せ誤りを除去するためである．評価済みの翻訳ルールを再帰的に使用することにより，訳語を推定し，組合せ誤りを消去する．その場合に問題となるのは，辞書中の翻訳ルールに誤りが存在するために，訳語誤りであるものを組合せ誤りであると判断してしまい，trueを与えることである．そこで，訳語誤りを組合せ誤りであると判定してしまう割合を予備実験により調査した．その結果，辞書中の翻訳ルールを用いて検索を試みた場合に対する，誤った翻訳ルールと一致し，trueを与えてしまう場合の割合は，1.4%であった．この結果より，辞書中の翻訳ルールを再帰的に用いて組合せ誤りの検索を行なっても，訳語誤りを組合せ誤りと判断してしまう可能性は低いと考えられる．表1に，図2で示した誤った翻訳ルール"(I don't like volleyball.；私/は/バスケットボール/が/好き/です.）に対する評価結果を示す．

4 実験結果

4.1 実験方法

初期状態は辞書が空の状態である．そして，中学1年生用教科書ガイド・ワンワールド[14]に掲載されている英文とその日本語訳文の395組の翻訳例を用いて，図1で示した処理過程に従い，翻訳ルールを抽出した．辞書中の翻訳ルールに対しては，人手によりその正誤を判定した．また，本実験は，いくつかの閾値を用いて行なった．

表 1: "(I don't like volleyball.; 私/は/バスケットボール/が/好き/です.)"の評価結果

	I	don't	like	volleyball
私	true	true	true	true
は	true	true	true	true
バスケットボール	true	true	true	false
が	true	true	true	true
好き	true	true	true	true
です	true	false	true	true

この場合，誤り率は 8.3%(2/24×100) となる．

表 2: 淘汰処理における適合率と再現率

	適合率	再現率
従来手法	86.1%	7.5%
改良手法 (5%)	89.9%	57.2%
改良手法 (10%)	93.3%	47.2%
改良手法 (15%)	95.1%	37.8%

4.2 実験結果

表 2 に従来手法と本稿で提案した改良手法の淘汰処理における適合率と再現率を示す．適合率は，淘汰された翻訳ルールにおける，正しく淘汰された翻訳ルールの占める割合である．再現率は，辞書中に存在する誤った翻訳ルールにおける，正しく淘汰された翻訳ルールの占める割合である．また，改良手法の ( ) 内の数値は，4.1 節で述べた閾値である．表 3 には，従来手法と改良手法における，それぞれの処理時間を示す．そして，表 4 には，改良前と改良後の GA-ILMT における翻訳率を示す．表 3 と表 4 における改良手法の実験結果は，閾値が 5%の場合である．

表 3: 処理時間

	従来手法 (秒)	改良手法 (秒)
100 文	11:21	10:00
200 文	71:19	45:12
300 文	2:42:21	1:41:30
395 文	5:05:58	2:40:58
合計	9:10:59	5:17:40

表 4: 翻訳率

	従来手法	改良手法
有効な翻訳	52.0%	53.3%
無効な翻訳	48.0%	46.7%

4.3 考察

4.3.1 閾値の決定

表 2 の閾値を変更した場合の実験結果において，適合率は，いずれの閾値においても 90%前後となった．それに対し，再現率は，5%の場合に最も高い精度を示している．改良手法では，翻訳ルールを構成している単語の組合せが，与えられた翻訳例のいずれかに存在していれば true となるため，3.2 節で述べた誤り率はかなり低くなる．そのため，閾値が比較的低い場合に，多くの誤った翻訳ルールを淘汰でき，適合率をさほど低下させることなく，再現率を向上させることが可能となる．この結果より，最適な閾値は，5%となった．

4.3.2 本手法の有効性

(1) 従来手法との比較

表 2 より，従来手法と比べ，適合率は 3.8%，再現率は 49.7%増加している．適合率は，従来手法においても比較的高い精度を示しているが，改良手法によって，さらに，向上させることができた．また，再現率は大幅に向上した．従来手法では，翻訳に使用された翻訳ルールのみが淘汰の対象となるのに対し，改良手法では，翻訳例を利用

することにより，全ての生成された翻訳ルールが淘汰の対象となる．その結果，翻訳ルールに対する直接的な正誤の判定が可能となり，再現率の向上の大きな原因となった．

また，従来手法では，淘汰の対象となる翻訳ルールは少ない単語数より構成されているものが大部分を占め，多くの単語数より構成されている翻訳ルールを評価することが困難であった．これは，構成されている単語数の多い翻訳ルールほど，入力文に対して適用可能な翻訳ルールを選択する際に行なわれるマッチング処理 [11] において，入力文と表層レベルで一致する可能性が低くなり，使用される回数が少なくなるためである．それに対し，改良手法では，翻訳ルールの構成単語数に依存することなく，翻訳ルールを評価することが可能となった．

(2) 翻訳結果について

GA-ILMT では，複数の翻訳結果が存在する場合，生成された翻訳結果の上位１番から１０番までに正翻訳が含まれていれば，その翻訳結果を有効な翻訳として評価している [11]．本手法の導入により，有効な翻訳率は 52.0% から 53.3%となり，大きな向上は見られなかった．しかし，有効な翻訳結果が存在していたにもかかわらず，その翻訳結果における優先順位が１位とならなかったものの 54.5%が順位を上げていた．その結果，上位１番のみを有効な翻訳として評価した場合には，有効な翻訳率は 42.3%から 44.6%に向上した．この結果は，誤った翻訳ルールが減少したため，誤翻訳の生成を抑制できたことを示している．

(3) 処理時間について

表３より，処理時間は約 42.3%減少した．これは，従来手法と比べ，多くの単語数より構成されている誤った翻訳ルールを消去することが可能となり，それらを用いたさらなる誤った翻訳ルールの抽出処理が大幅に減少したためである．

4.3.3 改良後の淘汰手法の問題点

改良手法では，4.3.2 節で述べたように，従来手法とは対照的に，多い単語数より構成されている翻訳ルールに対する淘汰が可能となった．しかし，少ない単語数より構成されている翻訳ルールに対する，正しい評価が依然として困難である．例えば,"(brother ; きみの/兄)" といった翻訳ルールに対しては,"(brother ; きみの)" と"(brother ; 兄)" を用いて，翻訳例との照合を行なう．その結果，翻訳例中に"(Is Makoto your brother? ; 真/は/きみの/兄/ですか？)" が存在すれば，この翻訳ルールの誤り率は 0.0 となってしまう．このように改良手法では，一度でも過去に与えられた翻訳例中に，単語の組合せが存在していれば正しいと判断されてしまうため，構成単語数が少なく，頻繁に出現する単語を持つ誤った翻訳ルールに対しては，正しい判断が困難となる．それに対し，従来手法では，少ない単語数より構成されている翻訳ルールに対する淘汰が可能である．しかし，GA-ILMT により作り出される翻訳ルール数に対し，そのような翻訳ルールは少数である．そのため，少ない単語数より構成されている誤った翻訳ルールを十分に淘汰できているとはいえず，この問題点を解決するものとはなっていない．したがって，改良手法では，少ない単語数より構成されている誤った翻訳ルールに対する淘汰の困難さが依然として問題となっている．

5 おわりに

本稿では，GA-ILMT における淘汰処理の精度向上のための改良手法とその有効性について述べた．遺伝的アルゴリズムでは，交叉や突然変異により，多種多様な個体が生成される．その中には，誤ったものも存在する．そのような誤った個体に対しては，淘汰を行なうことにより消去していく．GA-ILMT においても誤った翻訳ルールに対しては，淘汰処理を行なうことにより消去していくが，その淘汰処理の精度が低いため，即座に誤った翻訳ルールを消去できず，そのことが大きな問題となっていた．また，この問題を解決するにあたり，解析的な知識を使用することは，例外的なものや未登録語に対処することが困難であるという問題を抱え込むことになる．したがって，我々は，翻訳例のみから，学習的手法を用いて淘汰処理の精度を向上させる手法を提案した．その結果，評価実験により，適合率は 89.9%，再現率

は57.2%となった．また，処理時間は約42.3%減少した．我々は，評価実験を通し，本提案手法が淘汰処理の精度向上のために，有効な手法となることを確認できた．今後は，より実用的なデータを用いて，GA-ILMTにおける性能向上のための研究を進める予定である．

参考文献

[1] 長尾真：機械翻訳サミット，オーム社 (1989).

[2] 野村浩郷（編）：言語処理と機械翻訳，講談社 (1991).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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. J75-D-II, No. 7, pp. 1213–1221 (1992).

[8] 荒木健治，高橋祐治，桃内佳雄，栃内香次：帰納的学習によるべた書き文のかな漢字変換手法の適応能力の評価，電子情報通信学会信学技報，NLC 94–3, pp. 17–24 (1994).

[9] 荒木健治，栃内香次：多段階共通パターン抽出法を用いた翻訳例からの帰納的学習による翻訳，情報処理北海道シンポジウム’９１，pp. 47–49 (1991).

[10] 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).

[11] 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).

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

[13] 越前谷博，荒木健治，桃内佳雄：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，平成６年度電気関係学会北海道支部連合大会講演論文集，No. 162.

[14] 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).