自然言語処理 135－23
（2000．1．28）

学習型機械翻訳手法GA-ILMTにおける状態遷移の導入について

越前谷 博† 荒木 健治†† 桃内 佳雄† 栃内 香次††

†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電子情報工学専攻
echi@eli.hokkai-s-u.ac.jp araki@media.eng.hokudai.ac.jp
momouchi@eli.hokkai-s-u.ac.jp tochinai@media.eng.hokudai.ac.jp

我々は，これまでに与えられた翻訳例のみから翻訳ルールを自動的に獲得することにより翻訳を行う学習型機械翻訳手法として，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．しかし，学習という観点から本手法は十分な能力を備えるまでには至っておらず，その結果として，翻訳精度もまた不十分であった．そこで，我々は，このGA-ILMTにおいて，解析的な知識を明示的に与えることなく，学習能力の向上という観点からの改良を試みた．即ち，システム自身が獲得した翻訳ルールを階層的に処理することにより翻訳を行う能力の実現である．そのために，我々は状態遷移を導入した．状態遷移を導入することにより，システム自身が翻訳結果の生成過程に着目した翻訳を行う．本稿では，GA-ILMTにおける状態遷移の導入とその有効性について述べる．

Using State Transition on GA-ILMT based Learning Capability

Hiroshi Echizen-ya† Kenji Araki†† Yoshio Momouchi† and Koji Tochinai††

†Dept. of Electronics and Information, Hokkai-Gakuen University
††Division of Electronics and Information, Hokkaido University
echi@eli.hokkai-s-u.ac.jp araki@media.eng.hokudai.ac.jp
momouchi@eli.hokkai-s-u.ac.jp tochinai@media.eng.hokudai.ac.jp

We previously proposed a method of machine translation using inductive learning with genetic algorithms (GA-ILMT) based on learning capability. However, its learning capability is not enough. As the result, its translation qulaity is still low. We used a state transition to improve the learning capability of GA-ILMT. GA-ILMT using the state transition can perform translation based on the process of a translation without using any analyitical knowledge. In this paper, we will describe the use of state transition on GA-ILMT and describe an effectiveness of the state transition.

1　はじめに

近年，インターネットの急速な普及に伴い，実用的な機械翻訳システムが強く求められている．そうした状況において，多くの機械翻訳システムが商品化されるようになった．しかし，それらの商用の機械翻訳システムは，いくつかの問題点を含んでいると考えられる．現在の商用機械翻訳システムの多くに採り入れられているのが，トランスファー方式や中間言語方式を用いた解析型機械翻訳手法 [1][2] である．この解析型機械翻訳手法では，有限個の文法規則を用いるため，多様な言語現象に対処することが困難である．また，それらの文法規則は人手で常に与えることが前提となるため，システムのカスタマイズの際には，大きな労力を伴う．こうした問題点を解決する手法として，実例型機械翻訳手法 [3][4] や確率的な手法に基づく機械翻訳手法 [5] などのコーパスに基づく機械翻訳手法が提案された．これらの手法は，基本的には翻訳例のみを用いて翻訳が行われる．一方，解析型機械翻訳手法とコーパスに基づく翻訳手法の両方の観点からの研究として，パターンベースの翻訳手法 [6] がある．また，学習という観点からは，コーパスから翻訳規則を統計的なアプローチに基づき自動獲得する手法 [7] がある．しかし，これらの手法もやはり，それぞれ問題点を抱えていると考えられる．コーパスに基づく機械翻訳手法では，翻訳精度の向上のためには膨大な量の翻訳例が必要になる．また，パターンベースの翻訳手法では，文脈自由文法（CFG）規則を使用しているため，解析型機械翻訳手法と同様，人手による文法規則の記述という煩わしさを抱えている．翻訳規則を自動獲得する手法においても，使用するコーパスに様々な解析的な知識を付与するため，解析的な知識に依存したものとなる．その結果，解析型機械翻訳手法の問題点を引き込む可能性を否定できない．

そうした機械翻訳手法の現状において，我々は，解析的な知識を使用することなく，翻訳例のみからの翻訳ルールの自動学習という観点より，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）[8][9] を提案している．本手法では，与えられた翻訳例のみから帰納的学習により，翻訳ルールを自動獲得する．また，翻訳例の不足を解消するために遺伝的アルゴリズムの基本操作を適用することで，より多くの翻訳例を自動生成する．そして，自動獲得された翻訳ルールを使用することにより翻訳を行う．したがって，GA-ILMT は，与えられた環境から言語知識を学習により獲得していく人間の学習能力を模倣しようとする試みとして位置付けられる．そして，このようなシステムの実現は，これまでの機械翻訳手法の抱える，解析的な知識を使用することの煩わしさや大量の翻訳例を必要とするといった問題点を解決するものになり得ると考えられる．

我々は，このような GA-ILMT において，翻訳精度の更なる向上を，学習能力の向上という観点より行う．即ち，翻訳率の向上を解析的な知識の導入により行うのではなく，新たな枠組みを GA-ILMT に取り込み，学習能力を向上させることにより行うということである．我々は，これまでの GA-ILMT の問題点が獲得した翻訳ルールを階層的に捉え，翻訳を行う能力の欠如にあるとし，獲得した翻訳ルールを階層化して翻訳する能力を実現することで，翻訳能力の向上を試みた．そのために，我々は状態遷移を導入した．状態遷移の導入により，獲得した翻訳ルールをシステム自身が階層的に捉え，処理するメカニズムを実現する．その結果，翻訳結果の生成過程に着目した翻訳処理が可能となる．状態遷移は，自然言語処理の分野において，様々な形で利用されてきた．その代表的なものとしては，統語解析に用いられる拡張遷移ネットワーク (ATN) [10] が挙げられる．また，機械翻訳システムにおける導入例としては，トラスファー方式に基づく Mu システム [11] が文法規則の適用順序を詳細に制御するために状態遷移を取り入れている．したがって，これまでの状態遷移の導入はシステム中の統語規則を対象にしたものとして位置付けることができる．それに対し，我々は，学習能力に基づくシステムを対象に翻訳ルールの階層的な利用を実現するために状態遷移を導入する．即ち，状態遷移を GA-ILMT へ導入することにより，GA-ILMT の学習能力を向上させ，その結果として翻訳精

度を向上させる．本稿では，学習型機械翻訳手法GA-ILMTに対する状態遷移の導入と，性能評価実験を通して得られた結果に基づき状態遷移の有効性について述べる．

2　基本的な考え方

獲得した知識を用いて問題解決を行う場合，そこには問題解決に至るまでの過程というものが伴う．それは，いくつかの知識を階層的に捉え，体系化して利用するということに相当すると考えられる．翻訳においては，与えられた原文に対し，翻訳知識を用いて訳文を作り出す際，一つ一つの翻訳知識を単独に利用することで訳文を得るのではなく，いくつかの翻訳ルールを階層的に捉えながら，体系化することにより訳文を導き出すメカニズムが働いていると考えられる．

そこで，我々は学習型機械翻訳手法GA-ILMTにおいて，自動獲得した翻訳ルールを階層的に捉え，体系化して利用する能力を組み込む．その結果，GA-ILMTでは，翻訳結果を導き出すまでの過程に着目した翻訳が可能になる．このような考え方に基づき，我々は，GA-ILMTに対して状態遷移の導入を行う．より汎用的な翻訳ルールを起点として，徐々に具象的な翻訳ルールを生成することにより訳文を導き出す過程は，翻訳ルールの状態遷移として表すことができる．したがって，我々は，獲得した翻訳ルールを階層的に捉え，体系化していくことで翻訳を行う能力の実現のために状態遷移を導入する．

3　GA-ILMTの概要

3.1　GA-ILMTの処理過程

図1にGA-ILMTに基づく英日機械翻訳システムの処理の流れを示す．原文として，英文を入力すると，翻訳部において，それまでに獲得した翻訳ルールを用いて，翻訳結果を生成する．

そこ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，翻訳に使用された翻訳ルールを対象に，個々の翻訳ルールの評価を行う．そして，学習部では，与えられた翻訳例間において，遺伝的アルゴリズムの基本操作である交叉を適用し，新たな翻訳例を生成する．更に，それらの翻訳例から帰納的学習により翻訳ルールを獲得する．

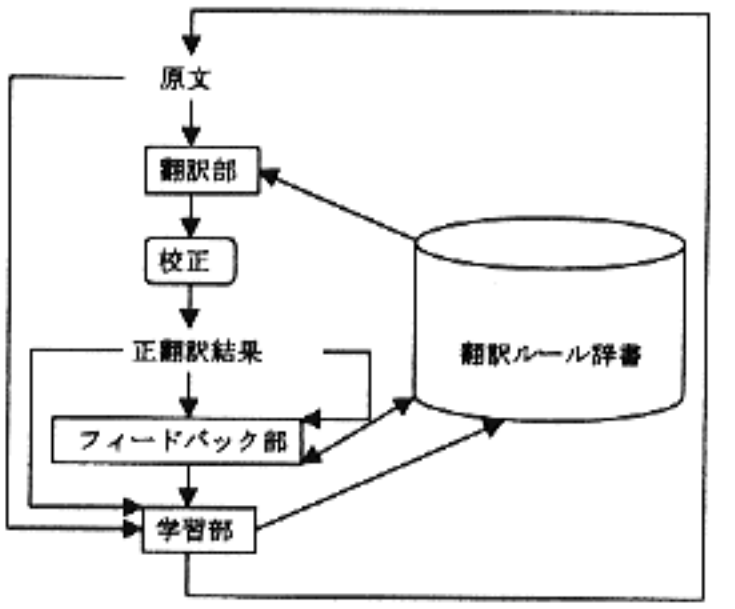

図1: GA-ILMTの処理過程

3.2　学習部における翻訳ルールの獲得

翻訳ルールの獲得は学習部において，帰納的学習により行われる．本稿で述べる帰納的学習とは，翻訳例からそこに内在する一般的な規則を抽出することである．その方法として，GA-ILMTでは，翻訳例間の表層的な共通部分と差異部分を多段階に抽出する．図2に翻訳ルール獲得の具体例を示す．

図2では，翻訳例中の差異部分がそれぞれ，"(tennis; テニス)"と"(tea ; お茶)"となるため，これらが存在していた箇所を変数化することで，共通部分としては"(He likes @0. ; 彼 は @0 が 好き です.)"が抽出されることになる．

翻訳例

(He likes tennis. ; 彼 は テニス が 好き です。)

交叉により生成された翻訳例

(He likes tea. ; 彼 は お茶 が 好き です。)

↓ 共通部分と差異部分の抽出

共通部分

(He likes @0. ; 彼 は @0 が 好き です。)

差異部分

(tennis ; テニス)

(tea ; お茶)

図 2: 帰納的学習による翻訳ルールの獲得例

4 GA-ILMTへの状態遷移の導入

4.1 状態遷移に基づく翻訳

翻訳部では，翻訳ルールを体系化し，翻訳結果の生成過程に着目した翻訳を行うために状態遷移に基づいた翻訳を行う．まず，個々の具象的な翻訳ルールに対して，それぞれ状態遷移を生成する．更に，尤度評価関数を用いて最も良質な状態遷移を選択する．そして，選択された状態遷移の最終状態に該当する翻訳ルールを翻訳処理に使用する．以下にその処理過程を述べる．

(1) 最も抽象的な翻訳ルールを初期状態に，そして，原文との類似性が高い，最も具象的な翻訳ルールを最終状態として，状態遷移を生成する．

(2) 以下の尤度評価関数を用いて，個々の状態遷移の評価値を決定する．

$$\sum_{j=1}^{N} \alpha \times CF - \beta \times EF + \gamma \qquad (1)$$

(3) 式(1)より計算された評価値に基づき，その値が最も高い状態遷移を選択する．そして，選択された状態遷移の最終状態の翻訳ルールを，翻訳結果の生成過程から求められた翻訳処理に最適な翻訳ルールであると位置付ける．

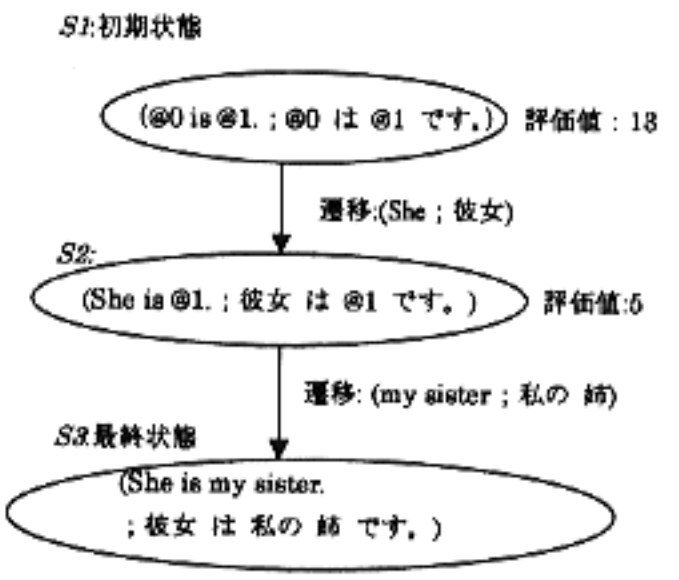

図 3: 状態遷移に基づく翻訳処理の具体例

式(1)において，Nは状態遷移を構成している翻訳ルールの数である．α，βは係数である．γは，正翻訳に使用された翻訳ルールに与えられる定数である．また，CFは正確実度であり，その翻訳ルールの英日の対応関係の信頼性が高いことを示している．EFは誤確実度であり，その翻訳ルールの英日の対応関係の信頼性が低いことを示している．これらは，次節で述べる状態遷移に基づく翻訳ルールの評価方法により決定される．

図3に，状態遷移に基づく翻訳処理の具体例を示す．図3において，S_iは状態であり，状態間を結んでいる矢印が遷移となる．最終状態の翻訳ルールが"(She is my sister. ; 彼女 は 私の 姉です.)"に対する評価値は18，また，最終状態の

翻訳ルールが"(She is my sister.;彼女 は 私の 姉.)"に対する評価値は4であった．したがって，最終状態の翻訳ルール（She is my sister.;彼女 は 私の 姉 です.）の状態遷移が選択され，その最終状態である翻訳ルールが翻訳処理に使用される．

4.2 状態遷移を用いた翻訳ルールの評価

翻訳ルールの評価はフィードバック部で行う．これまでのGA-ILMTの翻訳ルールの評価では正翻訳の場合には，それらを構成している全翻訳ルールの正翻訳度を，また，誤翻訳の場合には，それらを構成している全翻訳ルールの誤翻訳度を増加させる．したがって，これまでの翻訳ルールに対する評価は，組み合わせ結果に基づく評価である．しかし，その場合，誤翻訳ルール同士での組み合わせからでも正翻訳を生成することがあるため，誤翻訳ルールに対して，正翻訳度を増加するという問題点が生じる．状態遷移を利用した翻訳ルールの評価では，状態遷移より正翻訳の生成過程を保持しておくことにより，翻訳ルールの組み合わせ過程に基づき翻訳ルールを評価することができる．以下にその処理過程を述べる．以下の処理は，過去の翻訳において正翻訳に対する状態遷移を保持しておくことで実現される．

(1) 正翻訳のみに使用された変数を含まない翻訳ルールを正翻訳ルール（CTR）として，その翻訳ルールのCFを1増加させる．また，誤翻訳のみに使用された変数を含まない翻訳ルールを誤翻訳ルール（ETR）として，その翻訳ルールのEFを1増加させる．

(2) 上記の処理(1)より決定された正翻訳ルールまたは誤翻訳ルールを含む正翻訳の状態遷移を選択し，更にその中から現在の状態，遷移，次の状態を組とした1つの遷移過程を取り出す．

(3) 取り出された遷移過程に対して，表1に示す翻訳ルールの評価規則を適用することにより，変数を含む翻訳ルールを評価する．その結果，正翻訳ルールと判断された翻訳ルールのCFを1増加させる．また，誤翻訳ルールと判断された翻訳ルールのEFを1増加させる．表1にある"*"は翻訳ルールの正誤が決定されていないことを表している．

表 1: 翻訳ルールの評価規則

No.	現在の状態	遷移	次の状態	*
1	CTR	CTR	*	CTR
2	ETR	CTR	*	ETR
3	*	CTR	CTR	CTR
4	*	CTR	ETR	ETR
5	*	ETR	CTR	ETR
6	CTR	*	CTR	CTR
7	CTR	*	ETR	ETR
8	ETR	*	CTR	ETR

図4に，状態遷移に基づく翻訳ルールの評価の具体例を示す．図4では，変数を含まない翻訳ルールとして"(He is Andy.;彼 は アンディ です.)"，"(Andy;アンディ)"そして，"(He;彼)"が正翻訳のみに使用されたとして，正翻訳ルールとなる．次いで，遷移過程Aにおいて変数を含む翻訳ルール"(He is @1.;彼 は @1 です.)"が表1の規則3に基づき正翻訳ルールとして，評価される．そして，その結果より遷移過程Bにおいて，更に抽象化されている翻訳ルール"(@0 is @1.;@0 は @1 です.)"が，同じく表1の規則3に基づき正翻訳ルールとして評価される．

それに対し，組み合わせ結果に基づく翻訳ルールの評価を行っていた従来のGA-ILMTでは，翻訳ルール"(@0 is @1.;@0 は @1 です.)"を誤翻訳ルールとして評価する場合がある．例えば，"He is not Andy."といった否定文に使用された場合，"(not Andy;アンディではありません)"との組み合わせにより，翻訳結果は"～ではありませんです"となる．その結果，組み合わせ結果に基づく翻訳ルールの評価では，翻訳ルール"(@0 is @1.;@0 は @1 です.)"は誤翻訳ルールとして評価されてしまう．

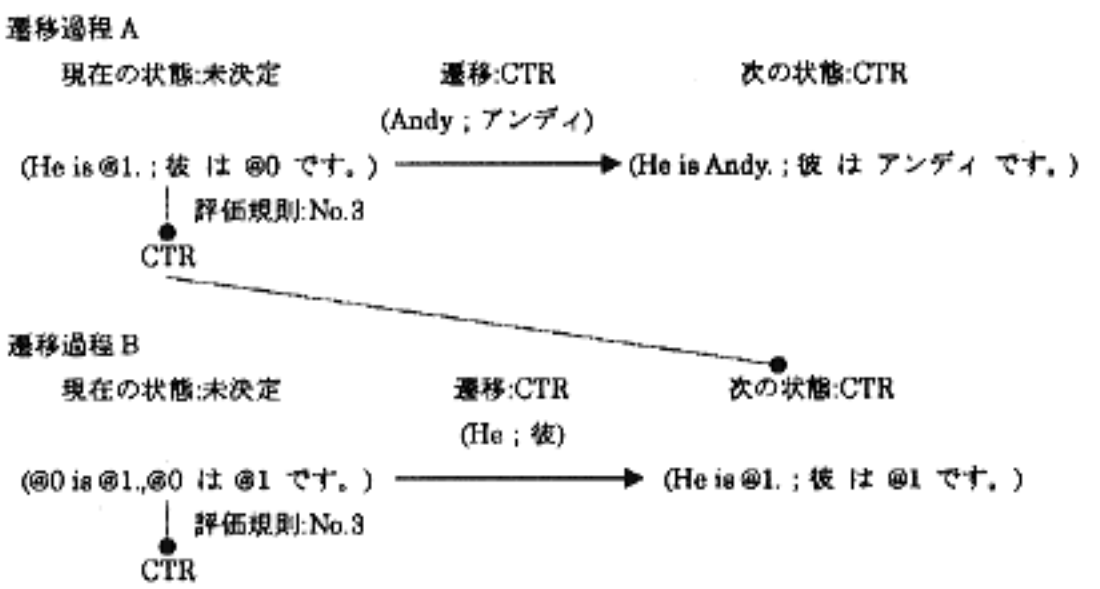

図 4: 状態遷移に基づく翻訳ルールの評価の具体例

また，従来では，本来誤翻訳ルールであるものが，正翻訳ルールとして評価される場合もある．例えば，"(Andy ; アンディです)"は"He is Andy."という原文に対し，翻訳ルール"(He is @0. ; 彼 は @0.)"との組み合わせから正翻訳を導き出すため，正翻訳ルールとして評価されてしまう．しかし，状態遷移を利用した翻訳ルールの評価では，それまでの状態遷移中に"(That is @0. ; あれ は @0.)","(Andy ; アンディ です)"そして，"(That is Andy. ; あれ は アンディ です.)"という遷移過程があり，"(That is @0. ; あれ は @0.)"が誤翻訳ルール，また，"(That is Andy. ; あれ は アンディ です.)"が正翻訳ルールであると評価されると表 1 の規則 8 より，"(Andy ; アンディ です)"を誤翻訳ルールとして評価することができる．このように，状態遷移を用いた翻訳ルールの評価では，組み合わせの履歴を参照することにより，評価精度を向上させることが可能となる．

4.3 確実性の高い翻訳ルールを用いた翻訳ルールの獲得

4.2 節で述べた状態遷移に基づく翻訳ルールの評価方法により，翻訳ルールに対する評価精度の向上を図ることが可能となる．学習部では，その結果を用いて，良質な翻訳ルールの獲得を行う．以下にその処理過程を示す．

(1) 以下の式に基づいて，単語の翻訳ルールの信頼度を求める．

$$\text{信頼度} = \frac{CF}{CF + EF} \qquad (2)$$

(2) 信頼度がある閾値以上の単語の翻訳ルールが，与えられた翻訳例に完全に含まれている場合，翻訳例中の単語の部分を変数化し，翻訳ルールを獲得する．

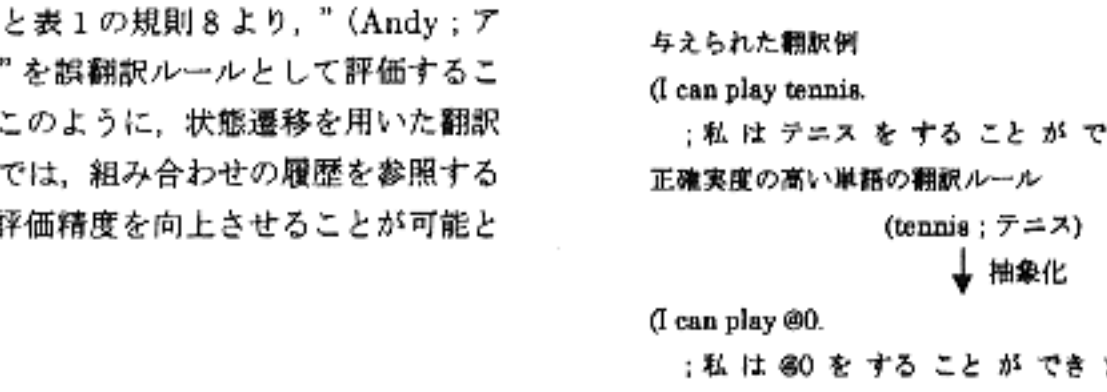

図 5: 信頼度の高い単語の翻訳ルールを用いた翻訳ルールの獲得

図 5 に信頼度の高い単語の翻訳ルールを用いた翻訳ルール獲得の具体例を示す．図 5 では，単語

の翻訳ルール”(tennis ; テニス)”の信頼度がある閾値以上であり，与えられた翻訳例に完全に含まれているため，その部分を変数に置き換えることで翻訳ルールを獲得する．

信頼度の高い翻訳ルールは，英日の対応関係が正しいということを意味する．したがって，翻訳例中のその部分を抽象化しても，残された部分の対応関係を崩すことなく，良質な翻訳ルールを獲得することが可能となる．

5 性能評価実験

5.1 評価方法

生成された翻訳結果については，以下の分類に該当するものを有効な翻訳結果とした．

(1) 未登録語を含まない正翻訳結果

(2) 未登録語を含んでいるが，未登録語が名詞句または形容詞句であるため単語の対訳を与えることで，容易に正翻訳結果となるもの

また，1文の翻訳において複数の翻訳結果が存在する場合には，上位1のみの翻訳結果を評価の対象とした．

5.2 実験方法

実験は，中学1年生用教科書ガイド・ワンワールド[12]に掲載されている英文700文を1文ずつ翻訳させ，その都度，文献[12]に掲載されている英文に対応する正しい日本語訳文を与えることで，校正及び学習を繰り返し行った．その際の翻訳ルール辞書の初期状態は空である．また，式(1)における係数α，βはそれぞれ，予備実験の結果に基づき2.0，1.0とした．定数γは5.0とした．4.3節で述べた信頼度の高い単語の翻訳ルールを用いた翻訳ルールの獲得において，信頼度の高い単語の翻訳ルールかどうかを決定する際の閾値としては，0.6を用いた．

5.3 実験結果

表2に従来のGA-ILMTの有効な翻訳率と状態遷移を導入したGA-ILMTの有効な翻訳率を示す．表2における()内の数は翻訳結果数である．

5.4 考察

表2より，有効な翻訳率は29.6%から41.9%へと12.3ポイント増加した．状態遷移の導入により，無効な翻訳から有効な翻訳となったものは118文であった．また，逆に，有効な翻訳から無効な翻訳となったものは32文であった．したがって，有効な翻訳は86文増加したことになり，状態遷移の有効性を確認することができた．状態遷移の導入が，生成過程に着目した翻訳結果の生成，システム自身による確実性の高い翻訳ルールの決定，そして，新たな確実性の高い良質な翻訳ルールの獲得をもたらした．解析的な知識を与えることなく，このような効果が得られたということは，状態遷移の導入によりGA-ILMTの学習能力が向上したたことを意味する．

従来のGA-ILMTでは，翻訳の際に翻訳結果の生成過程に着目することなく，原文との類似性の高い具象的な翻訳ルールを選択していた．しかし，その場合，良質な翻訳ルールの選択の精度は十分とはいえず，最終的に誤翻訳ルールを選択し，それを翻訳に用いることにより誤翻訳を生成する場合が生じる．それに対し，状態遷移を導入したGA-ILMTでは，いくつかの翻訳ルールを階層的に捉え，翻訳結果の生成過程に着目した翻訳処理を行うようになった．そのため，どのような翻訳ルールを，どのような過程で用いると有効な翻訳を導き出すことになるのかをシステム自身に判断させることが可能となった．

6 おわりに

我々は，人間が知識を階層的に捉え，体系化しながら翻訳処理を進めて行くという考え方に基づき，学習型機械翻訳手法GA-ILMTにおいて，自動獲得した翻訳ルールを階層的に捉えることで，

表 2: 従来の GA-ILMT と状態遷移を導入した GA-ILMT の有効な翻訳率

従来の GA-ILMT の有効な翻訳率		状態遷移を導入した GA-ILMT の有効な翻訳率	
29.6%(207)		41.9%(293)	
未登録語なし	未登録語あり	未登録語なし	未登録語あり
76.3%(158)	23.7%(49)	67.3%(197)	32.7%(96)

翻訳結果の生成過程に着目した翻訳を行うメカニズムを状態遷移の導入により実現した．そして，その有効性を確認するために性能評価実験を行った．その結果，正翻訳率は 29.6%から 41.9%へと 12.3 ポイント増加した．これは，翻訳ルールを階層的に捉えながら，翻訳結果の生成過程に着目した翻訳をもたらす状態遷移の有効性を示すものである．今後は更なる学習能力の向上を行うための研究を進めていく予定である．

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に行われた．

参考文献

[1] 野村浩郷（編），言語処理と機械翻訳，講談社 (1991).

[2] 田中穂積（監），自然言語処理—基礎と応用，コロナ社（1999）.

[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] Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer : The Mathematics of Statistical Machine Translation: Parametric Estimation, In *Computational Linguistics*, Vol. 9, No. 2(1993).

[6] Takeda K : Pattern-based Context-Free Grammars for Machine Translation, In *proceedings of the 34th ACL*(1996).

[7] 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).

[8] 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).

[9] 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*, Copenhagen, Denmark, pp.1020–1023(1996).

[10] W. A. Woods : Transition network grammars for natural language analysis, *CACM*, Vol. 13, No. 10(1970).

[11] Nakamura, J., Tsujii, J. and Nagao, M : Grammar writing system(GRADE) of Mumachine Translation Project and its Characteristics, In *proceedings of the Coling'84* (1984).

[12] 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).