第3回MYCOM資料
2002年6月17日,18日

再帰チェーンリンク型学習の対訳コーパスへの適用

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
††† 北海学園大学大学院経営学研究科

E-mail echi@eli.hokkai-s-u.ac.jp†

概要 機械翻訳の分野では，人手で膨大な言語知識を構築し，それらを用いて原文および訳文を解析することにより翻訳を行う，解析的なアプローチが主流である．しかし，解析的なアプローチの大きな問題点として，膨大な言語知識の構築の困難さが指摘されている．この問題は，解析的なアプローチが文に内在する言語情報を詳細に抽出することを前提とした手法であることから生じる．それに対し，我々は，文に内在するあらゆる言語情報を過度に抽出するのではなく，ある程度は内在させたままの方が有効であると考え，対訳テキストを最大限に利用する．すなわち，対訳テキストの適度な一般化である．過度な一般化を行うと，対訳テキストを具象化する際に膨大な言語知識が必要になり，解析的なアプローチと同様の問題を抱えることになる．本報告で，我々は，適度に一般化された翻訳ルールを自動的に獲得することが可能な，再帰チェーンリンク型学習を提案する．再帰チェーンリンク型学習では，獲得された翻訳ルールに基づき他の対訳テキストからの切り離し部分を決定する．したがって，獲得された翻訳ルールが新たな翻訳ルールの獲得をもたらすことになり，翻訳ルールの獲得処理の連鎖が生じる．この再帰チェーンリンク型学習を用いることで，高い学習能力を有する機械翻訳システムを実現できると考えられる．本報告では，この再帰チェーンリンク型学習の基本的な考え方，そして，その有効性について述べる．

1 はじめに

1.1 従来手法とその問題点

高い品質を有する機械翻訳に対するニーズは，近年のインターネットの普及を背景に，より一層高まっている．現在，最も多く使用されている機械翻訳手法は，原文に対し人手で記述された言語知識を用いて解析し，その結果得られた原言語の構文構造を目的言語の構文構造に変換した上で訳文を生成する，解析的なアプローチの一つであるトランスファー方式の機械翻訳手法 [1] である．このトランスファー方式の機械翻訳手法は，解析，変換，生成の大きく３つのプロセスからなる．解析では，形態素解析，構文解析を順に行うことで，原文を原言語の構文構造に変換する．構文構造の表現形式には，句構造文法等がよく用いられる．変換では，原言語の構文構造に対応する目的言語の構文構造を変換規則を用いて決定すると共に，原文中の語に対する訳語選択を行う．そして，生成では，決定された目的言語の構文構造に基づき，訳文を生成する．図１にトランスファー方式を用いた翻訳例を示す．

ここで，原文「It starts in thirty minutes.」を解析的なアプローチで翻訳することを考える．解析的なアプローチでは，訳文「それは 30 分たてば始まります.」を得るためには，原文に内在している言語情報を適切に抽出することで，初めて正しい翻訳が可能となる．具体的には，品詞を決定する形態素情報，文の構造を決定する構文情報，そして，訳語の選択条件が必要となる．原文「It starts in thirty minutes.」は構文的には非常に単純な文ではあるが，正翻訳文に至るまには，様々な言語知識を要する．そして，それらの言語知識は，基本的には全て人手で構築しなければならない．どのような言語知識が必要になるのかについて具体的に考える．例えば，原文の単語 ”in ”に対する訳を決定する場合，そこには多義性が存在する．一般的な ”in ”の訳としては場所や位置を表す語と共起している場合の ”〜の中 ”がある．

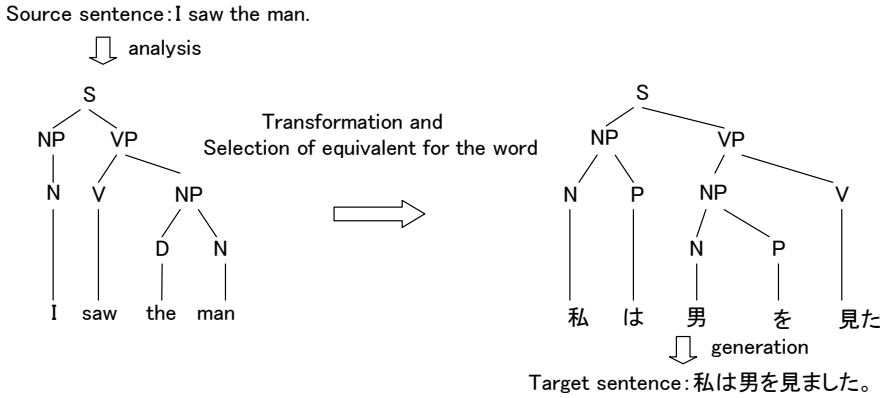

図 1: トラスファー方式による翻訳の具体例

更に，全体に対する割合を表現する場合の ”〜のうちで ”や時間を表す語と共起している場合の ”〜の間 ”もしくは ”〜たてば ”，”〜間 ”がある．原文「It starts in thirty minutes.」の場合は，時間を表す語 ”minutes ”と共起しているため，”〜の間 ”，”〜たてば ”もしくは ”〜間 ”のいずれかとなる．次に，この３つの中でどれが適切であるかを決定する．”〜間 ”という訳は「I haven’t seen him in three years.」などの「過去のある時点から現在までの間」という意味で使用されることを考えると適切ではない．したがって，原文「It starts in thirty minutes.」の ”in ”の訳としては，”〜の間 ”と ”〜たてば ”のいずれかが適切だと考えられる．この訳は動詞 ”starts ”との共起関係に基づき決定される．このように，”in ”の訳一つを取っても，文中の様々な語の関係を意味的に捉えることで初めて決定できる．このように非常に短く，構文的には単純な文であっても適切な訳文を導くためには，その文に内在している言語知識を正確に処理しなければならない．しかし，あらゆる言語現象を処理できる言語知識をあらかじめ人手で記述することは非常に困難である．そのため，新たな言語現象が出現する度に，それに対処するために新たな言語知識を人手で追加しなければならない状況が定常的に発生する．その結果，言語知識は複雑化し，それまで翻訳可能であったものが翻訳できなくなるという副作用を引き起こす．

1.2 提案手法の基本的な考え方

我々は，解析的なアプローチのように，文に内在している言語知識を詳細に抽出することを前提とするのではなく，高度な言語知識が要求される部分については，対訳テキストに内在させた状態のままの知識を利用することにより，より良い機械翻訳システムの実現を目指す．例えば，原文「It starts in thirty minutes.」において ”in ”の訳が ”starts ”と ”minutes ”に基づき決定されるのであれば，これらを切り離さずに「@0 starts in @1 minutes.」とし，それに伴い，訳文も「@0/は/@1/ 分/たて/ば/始まり/ます.」とし，これらの組を翻訳ルールとして持つことで，原文「It starts in thirty minutes.」に対する翻訳を行う.（@0 starts @1.；@0/は/@1/始まり/ます.）のように，前置詞句を切り離してしまうと,（in thirty minutes；30/分/の/間）,（in thirty minutes；30/分/たて/ば）,（in thirty minutes；30/分/間）等のように多義性が発生し，どれが適切であるかを決定するためには高度な言語知識が必要になる．したがって，この場合,（@0 starts in @1 minutes.；@0/は/@1/分/たて/ば/始まり/ます.）が最も有効な翻訳ルールであると考えられる．

更に我々は，学習能力の工学的な実現という立場から，このような適度に一般化された翻訳ルール（@0 starts in @1 minutes.；@0/は/@1/分/たて/ば/始まり/ます.）を自動的に獲得することを目指す．翻訳

ルール（@0 starts in @1 minutes.；@0/は/@1/分/たて/ば/始まり/ます.）は対訳テキスト「It starts in thirty minutes.；それ/は/30/分/たて/ば/始まり/ます.」から（It；それ）と（thirty；30）を切り離すことで得られる．対訳テキスト「It starts in thirty minutes.；それ/は/ 30/分/たて/ば/始まり/ます.」から（thirty；30）を切り離すということは,（thirty；30）からすると，英文においては，”thirty ”の前後を，訳文においては ”30 ”の前後を対訳テキストから切り離しても良いということになる．したがって,（thirty；30）という翻訳ルールは,（thirty；30）を含む対訳テキストが他に存在する場合,（thirty；30）の両側を切り離すという情報を持っているとみなすことができる．一方，対訳テキスト「It starts in thirty minutes.；それ/は/30/分/たて/ば/始まり/ます.」から（thirty；30）を切り離すということは，対訳テキストから見ると，英文においては ”in ”の右側から ”minutes ”の左側，訳文においては ”は ”の右側から ”分 ”の左側の部分を，他の語で構成される対応関係，例えば（five；5）に置き換えても良いということになる．したがって,（@0 starts in @1 minutes.；@0/は/@1/分/たて/ば/始まり/ます.）は，他の対訳テキストにおいて原文中に ”in 〜 minutes ”という部分が存在し，かつ訳文中に ”は〜分 ”という部分が存在する場合，原文からは ”in ”と ”minutes ”で挟まれた部分を，訳文からは ”は ”と ”分 ”で挟まれた部分を切り離すという情報を持っているとみなすことができる．本報告では,（thirty；30）のように対訳テキスト中の部分を表現している翻訳ルールを部分翻訳ルールと呼ぶ．それに対し,（@0 starts in @1 minutes.；@0/は/@1/分/たて/ば/始まり/ます.）のように対訳テキストの全体を表現している翻訳ルールを文翻訳ルールと呼ぶ．部分翻訳ルールおよび文翻訳ルールが有する対訳テキストからの切り離しの位置情報を利用することにより，システムは品詞情報や構文情報などの解析的な知識を用いることなく，表層情報のみから適度に一般化された翻訳ルールを獲得することが可能になると考えられる．更に，学習機能に基づき翻訳ルールを対訳テキストから自動獲得する他手法 [3, 5, 6] が，膨大な量の類似した対訳テキストを要求するのに対し，本手法では，獲得済みの翻訳ルールに基づき新たな翻訳ルールを獲得するため，スパースな状態のデータからであっても，効率よく翻訳ルールを獲得することができる．

以上のことから，部分翻訳ルール，文翻訳ルールは共に他の対訳テキストからの切り離し部分の位置情報を有していると捉えることにより，様々な対訳テキストが与えられた場合，連鎖的に対訳テキストの一般化を行うことができる．すなわち，部分翻訳ルール A が存在することで，文翻訳ルール B が獲得され，更に，文翻訳ルール B の存在により部分翻訳ルール C が獲得されるということである．我々はこのような翻訳ルールの連鎖的な獲得のメカニズムを「対応付けされた 1 組の事物から，対応関係にある部分を抽出する能力」と位置づけ，これを再帰チェーン リンク型学習と呼ぶ．図 1 に再帰チェーン リンク型学習の処理の一般化した概略図を示す．

図 1 は，翻訳ルール A が与えられることにより，翻訳ルール B，C，D が連鎖的に獲得されていく過程を示している．本報告では，獲得された翻訳ルールにおいて原文から得られたものを原言語部（以下，原部と記す．）とし，それに対し，訳文から得られたものを目的言語部（以下，目的部と記す．）とする．また，翻訳ルール A のように，翻訳ルールの連鎖的な獲得処理の起点となる翻訳ルールは表層レベルで類似関係にある対訳テキスト対の差異部分と共通部分を抽出することにより獲得する．このような処理を行う手法を我々は遺伝的アルゴ リズムを適用した帰納的学習による機械翻訳手法（略して GA-ILMT と呼ぶ）[7] として既に提案している．GA-ILMT は本手法と同様に，翻訳に必要な知識を学習の観点より自動獲得する手法であるため，GA-ILMT を用いることによりシステム全体を学習機能に基づく機械翻訳システムとして構築することができる．

図 1 に示すように，再帰チェーン リンク型学習では，部分翻訳ルールと文翻訳ルールの獲得を交互に行なうことで，翻訳ルール獲得の連鎖が起こる．処理 1 では，部分翻訳ルール A が有する，原文または原部から “Z” が，訳文または目的部から “ζ” が抽出する部分であるという情報に基づき，対訳テキス

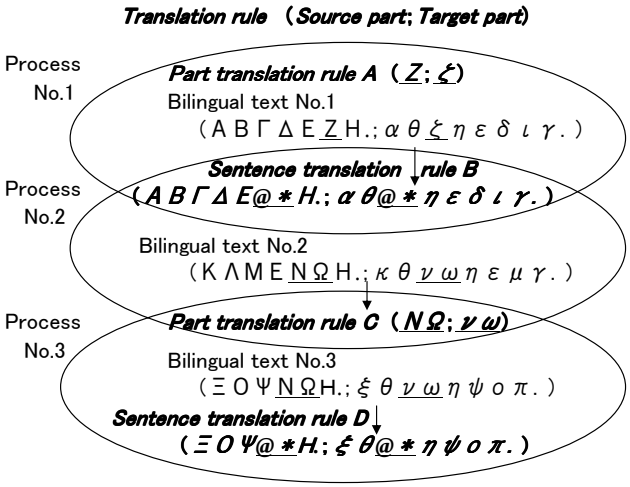

図 2: 再帰チェーンリンク型学習による翻訳ルールの獲得処理の概略図

ト１の原文から “Z” を，訳文から “ζ” を抽出する．そして，抽出された部分を変数 “@＊” に置き換えることで文翻訳ルール B を獲得する．獲得された文翻訳ルール B は，原文では “E” の右側から “H” の左側までが，訳文では “θ” の右側から “η” の左側までが抽出範囲であるという情報を有する．したがって，処理２では，文翻訳ルール B に基づき対訳テキスト２の原文から “NΩ” が，その訳文から “$\nu\omega$” が抽出され，部分翻訳ルール C（NΩ；$\nu\omega$）が獲得される．更に，処理３では，部分翻訳ルール C に基づき文翻訳ルール D が獲得される．抽出元である対訳テキスト自体も，変数を含まない文翻訳ルールとして登録される．また，これらの処理は表層レベルで共通部分と差異部分を決定することにより行なわれる．したがって，システムはプリミティブな能力として「二つの事物において同じ部分と異なる部分を判断する能力」[2] を有する．

2　再帰チェーンリンク型学習による機械翻訳手法

2.1　システムの概要

図３に再帰チェーンリンク型学習を備えた英日の翻訳を行う機械翻訳システムの構成図を示す．原文として英文が入力されると，翻訳部において，辞書中の翻訳ルールを用い翻訳結果として日本語訳文を生成する．生成された日本語訳文に誤りが含まれている場合には，人手により正しい訳文を与える．次いで，フィードバック部では，翻訳に使用された翻訳ルールの評価を行なう．その結果，正翻訳ルールと判断された翻訳ルールに対しては正確実度を１増加し，誤翻訳ルールと判断された翻訳ルールに対しては誤確実度を１増加する．学習部では，GA-ILMT と再帰チェーンリンク型学習により，与えられた翻訳例から翻訳ルールを自動獲得する．したがって，本システムは，学習により翻訳知識が増加することで，より良い翻訳システムへと成長するブートストラップ型の機械翻訳システムである．更に，再帰チェーンリンク型学習により学習能力の向上を図っている．

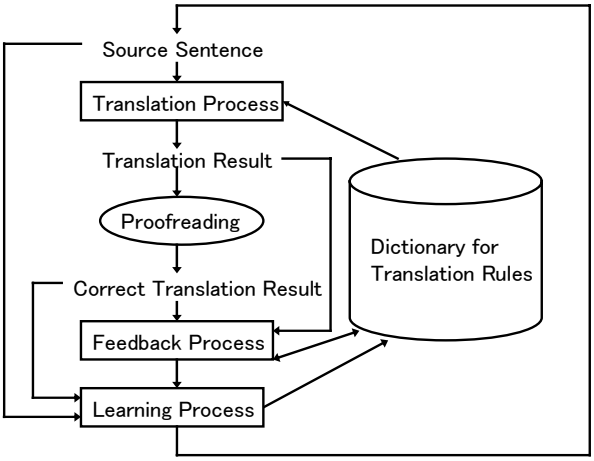

図 3: システム構成図

2.2 再帰チェーンリンク型学習による部分翻訳ルールから文翻訳ルールの獲得

再帰チェーンリンク型学習では，部分翻訳ルールと文翻訳ルールの獲得を交互に行なうことで，連鎖的な翻訳ルールの獲得を実現している．ここでは，文翻訳ルールによる部分翻訳ルールの獲得処理について述べる．以下に文翻訳ルールによる部分翻訳ルール獲得の処理過程を述べる．

（１）文翻訳ルールの原部と目的部のそれぞれにおいて，変数に隣接している部分と同じ部分を持つ対訳テキストを選択する．文翻訳ルールと対訳テキストとの間に存在する同じ部分である共通部分は１語以上で構成される．

（２）文翻訳ルールと選択された対訳テキストの原部間と目的部間のそれぞれにおいて，以下の処理のいずれかを行なう．

- 文翻訳ルールの原部または目的部の変数の両側に共通部分がある場合，対訳テクスト中の共通部分で挟まれている部分を原文またはその訳文から抽出する．
- 文翻訳ルールの原部または目的部の変数の右側のみに共通部分がある場合，対訳テキストの原文またはその訳文中の共通部分の左側から文の先頭までを抽出する．
- 文翻訳ルールの原部または目的部の変数の左側のみに共通部分がある場合，対訳テキストの原文またはその訳文中の共通部分の右側から文の末尾までを抽出する．

（３）原文とその訳文のそれぞれから抽出された部分の組を部分翻訳ルールとする．

（４）獲得された部分翻訳ルールに対し，文翻訳ルールと同じ値の確実度を与える．すなわち，確実度の高い文翻訳ルールを用いた場合，獲得された部分翻訳ルールの確実度も高くなる．

図４に文翻訳ルールによる部分翻訳ルール獲得の具体例を示す．図４では，文翻訳ルールの原部において変数の両側の部分 “in” と “minutes” が共に対訳テキストの英文に存在するため，これらが共通部分となる．したがって，対訳テキストの英文から共通部分で挟まれている部分 “thirty” が抽出される．更に，文翻訳ルールの目的部において変数の両側の部分 “は” と “分” が共に対訳テキストの日本語訳文に存在するため，これらが共通部分となる．したがって，対訳テキストの日本語訳文から共通部分で挟

まれている部分“３０”が抽出される．その結果，抽出された部分の組（thirty；３０）が部分翻訳ルールとして獲得される．

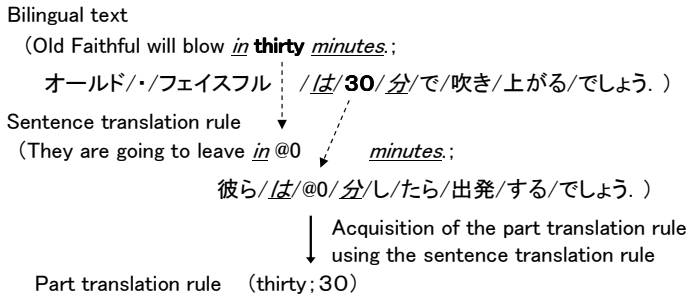

図 4: 文翻訳ルールによる部分翻訳ルールの獲得例

2.3 再帰チェーンリンク型学習による文翻訳ルールから部分翻訳ルールの獲得

2.2 で述べた文翻訳ルールによる部分翻訳ルールの獲得が行なわれると，その部分翻訳ルールを用いることで，新たな文翻訳ルールが獲得される．以下に部分翻訳ルールによる文翻訳ルールの獲得の処理過程を述べる．

（１）原部全体が対訳テキストの原文または文翻訳ルールの原部に存在し，かつ目的部全体が対訳テキストの訳文または文翻訳ルールの目的部に存在する部分翻訳ルールを選択する．

（２）対訳テキストの原文または文翻訳ルールの原部に対し，部分翻訳ルールの原部が存在する部分を変数に置き換える．また，対訳テキストの訳文または文翻訳ルールの目的部に対し，部分翻訳ルールの目的部が存在する部分を変数に置き換える．そして，変数が置き換えられた後の原部と目的部の組を文翻訳ルールとする．

（３）獲得された文翻訳ルールに対し，部分翻訳ルールと同じ値の確実度を与える．すなわち，確実度の高い部分翻訳ルールを用いた場合，獲得された文翻訳ルールの確実度も高くなる．

図５に，図４で獲得された部分翻訳ルール（thirty；３０）を用いた文翻訳ルールの獲得例を示す．図５では，部分翻訳ルール（thirty；３０）の原部と目的部が共に対訳テキストの原文とその訳文に存在するため，文翻訳ルール（It starts in @0 minutes.；それ/は/@0/分/たて/ば/始まり/ます.）が獲得される．そして，部分翻訳ルール（it；それ）を用いることにより，更に抽象化された文翻訳ルール（@1 starts in @0 minutes.；@1/は/@0/分/たて/ば/始まり/ます.）が再帰的に獲得される．

3 再帰チェーンリンク型学習を用いたシステムの性能評価

3.1 機械翻訳システム

我々は，図３のシステム構成図に基づき再帰チェーンリンク型学習を備えた英日の機械翻訳システムを構築し，本手法の性能評価実験を行った．その際には，システムに対し，中学 1，2 年生用の英語テキスト [8, 9, 10] に記載されている翻訳例 1,759 組を学習データとして与え，更に，中学 2 年生用の英語テ

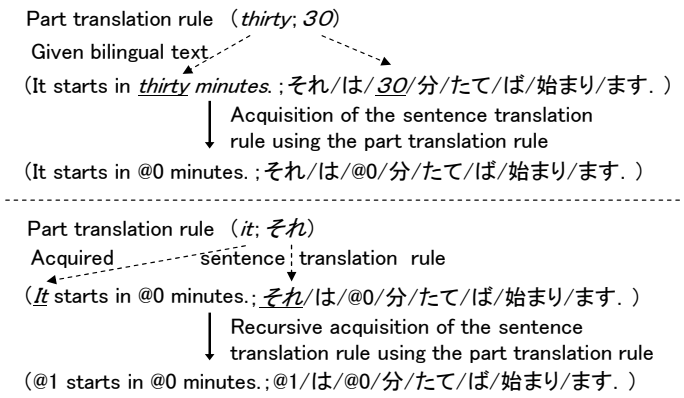

図 5: 部分翻訳ルールによる文翻訳ルールの獲得例

キスト [11, 12] に記載されている英文 1,097 文を評価データとして与えた．その結果，評価データに対する有効な翻訳率としては 61.1%が得られた．また，学習文が過不足なく与えられた場合の有効な翻訳率としては 85.0%が得られた．

3.2 辞書未登録語の獲得

再帰チェーンリンク型学習により獲得された部分翻訳ルールのみに着目した場合，そこには数多くの対訳語ペアが含まれている．したがって，本システムは対訳語ペアを自動獲得し，辞書未登録語を解消する手法と位置付けることもできる．我々は，始めに英語のテキスト [8]〜[12] に記載されている英文 2,856 を商用の機械翻訳システムに与えることにより，訳語が得られずアルファベットのまま出力された英単を抜粋した．すなわち，辞書未登録語の検出を行った．そして，検出された辞書未登録語 37 に対し，再帰チェーンリンク型学習を備えたシステムがどれだけの訳語を獲得できたのかを調査した．その結果，23 個の未登録語に対し訳語を得ることができ，その精度は 62.2%であった．また，獲得された対訳語ペアの 65.2%が対訳コーパス中に 1 度もしくは 2 度しか出現していないものであった．したがって，本システムは，大量の対訳テキストを必要とすることなく，対訳テキストを有効利用することにより，辞書未登録語に対処できることが確認された．

4 おわりに

機械翻訳において，我々は，文に内在する言語情報を詳細に抽出することを前提とする解析的なアプローチに対し，高度な言語知識が要求される部分については，対訳テキストに内在させた状態のままの知識を利用するという観点から，かつ，そのような知識をシステムが自動的に獲得するという観点から，対訳テキストを適度に一般化することのできる再帰チェーンリンク型学習について述べた．再帰チェーンリンク型学習では，様々な対訳テキストから既に獲得済みの翻訳ルールに基づき，新たな翻訳ルールを連鎖的に獲得する．したがって，他の対訳テキストから翻訳ルールを自動獲得する手法と比べ，与えられた対訳テキストをより有効に活用することが可能である．すなわち，膨大な量の対訳テキストを要求しない．更に，表層情報のみを利用するため，解析的な知識に依存しない．我々は，この再帰チェーンリンク型学習をブートストラップ型のシステムにインプリメントし，その有効性を性能評価実験を通

して確認した．

参考文献

[1] 田中穂積（監），自然言語処理—基礎と応用—,（社）電子情報通信学会，東京，1999．

[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. J79-D-ll, No. 3, pp. 391–402 (1996).

[3] C. Malavazos and S. Piperidis，“Application of analogical modelling to example based machine translation，”Proc. Coling2000, pp.516–522, Saarbücken, Germany (2000).

[4] C. Malavazos, S. Piperidis and G. Carayannis, “Towards memory and template based translation synthesis，”Proc. Machine Translation and Multilingual Applications in the New Millennium, pp.1-1–1-8, Exeter, England (2000).

[5] H.A. Güvenir and I. Cicekli，“Learning translation templates from examples，”Information Systems, vol.23, no.6, pp.353–363, 1998.

[6] K. McTait，“Linguistic knowledge and complexity in an EBMT system based on translation patterns，”Proc. Workshop on EBMT, Machine Translation Summit VIII, pp.23–34, Santiago de Compostela, Spain (2001).

[7] 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).

[8] 教科書ガイド教育出版版ワンワールド１，日本教材，東京，2001．

[9] 教科書ガイド教育出版版ワンワールド２，日本教材，東京，2001．

[10] 教科書システム問題集２，朋友出版，東京，2001．

[11] 教科書ワーク２，文理，東京，2001．

[12] 教科書トレーニング２，新興出版社，大阪，2001．