社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (19)
2008－ＮＬ－183 (19)
2008／1／22

学習型機械翻訳手法における省略可能性を用いた翻訳ルールの自動獲得とその有効性

寺島 涼† 越前谷 博†† 荒木 健治†

†北海道大学大学院情報科学研究科
††北海学園大学工学部
E-mail:†{terashima,araki}@media.eng.hokudai.ac.jp
††echi@eli.hokkai-s-u.ac.jp

我々は学習の観点より，翻訳例から翻訳に必要な翻訳ルールを学習により自動獲得する，学習型機械翻訳手法を提案している．そこでは，効率的に翻訳ルールを獲得することが重要となる．即ち，データスパースネスなコーパスに対処する必要がある．この問題を解決するために，本稿では，翻訳例中の省略可能な部分に着目することにより，より効率よく翻訳ルールを自動獲得するための手法を提案する．本手法では，翻訳例中の句に相当する，対応関係が明確な部分を抽出することにより翻訳ルールを獲得する．そして翻訳例からそれらの対応関係を抽出するための範囲情報に関する知識を省略可能性の観点から自動獲得する．性能評価実験の結果，省略可能性による翻訳ルールの獲得処理を適用することで有効な翻訳率が 5.5 ポイント増加し，本手法の有効性が確認された．

Automatic Acquisition of Translation Rules Focused on the Deletion Possibility in Translation Examples for Machine Translation Method and Its Effectiveness

Ryou Terashima†, Hiroshi Echizen-ya†† and Kenji Araki†

†Graduate School of Information Science and technology, Hokkaido University
††Faculty of Engineering, Hokkai-Gakuen University
E-mail:†{terashima,araki}@media.eng.hokudai.ac.jp
††echi@eli.hokkai-s-u.ac.jp

From the view of learning, we have proposed a method for machine translation that acquires the translation rules from the translation examples. In this method, it is important to acquire the translation rules efficiently because of the sparse data problem in a corpus. In this paper, we propose a method that efficiently acquires the translation rules focusing on the deletion in the translation examples. In our method, the translation rules are acquired by extracting the correspondence parts between the source language sentences and target language sentences, such a phrase, in the translation examples. Based on the view of the deletion possibility, our method acquires the knowledge about the scope information that is used to extract the correspondence parts from the translation examples. As the result of evaluation experiments, the effective translation rate improved 5.5 points by using our method. Therefore, we confirmed the effectiveness of our method.

1．はじめに

近年，多くの機械翻訳システムを利用することが可能となっている．しかし現状では十分な翻訳品質をユーザに提供するには至っていないと考えられる．解析型の機械翻訳手法[1]は，人手であらかじめ文法規則や変換規則などを記述し，解析的に翻訳を行う．この手法は多様な言語現象を規則として記述することが困難であること，また規則の追加などの改良により副作用が生まれてしまうといった問題点がある．

一方，コーパスに基づく機械翻訳手法は，統計に基づく機械翻訳手法[2]と実例に基づく機械翻訳手法[3],[4]の二つに大きく分類される．これらの手法は翻訳例そのものを有効利用することにより，自然な訳文を生成することができる．しかし，統計に基づく機械翻訳手法では翻訳のために，大規模なコーパスが不可欠である．また，実例に基づく機械翻訳手法では，構文解析ツールなどの解析的な知識に依存しているため，様々な言語への適用が容易ではない．

そこで我々は学習の観点より，翻訳例から翻訳に必要な翻訳ルールを学習により自動獲得する，学習型機械翻訳手法[5],[6]を提案している．しかし，データスパースネスなコーパスに対しては，効率的な翻訳ルールの獲得という点において，十分とはいえない．この問題を解決するために，本稿では，翻訳例中の省略可能な部分に着目することにより，より効率よく翻訳ルールを自動獲得するための手法を提

案する．本稿における翻訳ルールとは，翻訳例中に内在する規則であり，翻訳例を一般化することで得られる．本手法では，翻訳ルールを得るうえで，翻訳例中の一般化する部分を決定するための範囲情報を含む抽出ルールを自動獲得する．この抽出ルールは，翻訳例中の省略可能な部分に着目することにより得られる．例えば，翻訳例(I'd like a quiet room with a bath. ；バス/付き/の/静か/な/部屋/を/お願い/し/ます．)に対して，翻訳ルール(a room ；部屋)が存在した場合，"a"と"room"に挟まれた"quiet"と，"部屋"の左側の"静か/な"を省略可能と位置づけ，抽出ルールとして(a @0 room ；@0/部屋)を獲得する．この抽出ルールは翻訳例の原文における"a"から"room"までの部分と，目的言語文における"部屋"を含みその左側に存在する"a ～ room"に対応する部分の組が，句レベルの対応関係にあることを表している．また，この抽出ルールを一般化することにより，(a @0 ；@0)のような新たな抽出ルールを獲得する．この抽出ルールは原文における"a"とその右側に隣接する部分の組と，その組に対応する目的言語文中の部分が句レベルの対応関係にあることを表わしている．更に，このような抽出ルールを翻訳例に適用することにより，より多くの抽出ルールを再帰的に獲得する．例えば(～with a double bed ?；ダブル/ベッド/付き/の～)を含む翻訳例に対して，抽出ルール(a @0 ；@0)を適用することで，(～with @0 ；@0/付き/の～)を含む翻訳ルールが獲得される．この翻訳ルールに基づき，さらに抽出ルールとして(with @0 ；@0/付き/の)を獲得する．この抽出ルールは翻訳例の原文における"with"の右側に隣接する部分と，目的言語文における"付き/の"の左側に隣接する部分の組が句レベルの対応関係にあることを表している．本稿では，省略可能性による翻訳ルールの獲得処理とその有効性について述べる．

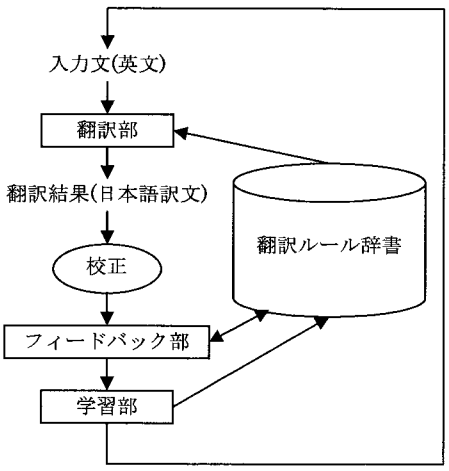

図１　処理過程

２．処理過程

本手法を用いた英日機械翻訳システムの処理過程を図１に示す．まず，入力文として英文を入力する．翻訳部では翻訳ルール辞書中の翻訳ルールを用いて翻訳結果を生成する．生成された翻訳結果に誤りが含まれている場合は，人手で校正する．次にフィードバック部では，翻訳に使用された翻訳ルールに対する評価を行う．学習部では，入力文と正しい翻訳結果の組である翻訳例に対し，帰納的学習[7]と本手法を適用することにより，翻訳ルールと抽出ルールを自動獲得する．そして，それらを翻訳ルール辞書に登録する．その際，翻訳例自体も翻訳ルールとして翻訳ルール辞書に登録する．本システムで獲得される翻訳ルールは文翻訳ルールと部分翻訳ルールの２種類である．文翻訳ルールとは，文の構造をもつ翻訳ルールであり，翻訳例自体のような具象的なルールや，一般化された抽象的なルールがある．部分翻訳ルールとは翻訳例中の部分を表わしており，文翻訳ルールと同様に一般化されたものと，一般化されていないものがある．また，翻訳ルールは，翻訳例中の原言語文から得られる原言語部と目的言語文から得られる目的言語部の組からなる．一方，抽出ルールは，翻訳ルールとしても利用可能なため，翻訳ルール辞書に登録される．

２．１　翻訳部

翻訳部では，入力文に対し，既に獲得された翻訳ルールを用いて翻訳結果を生成する．はじめに，翻訳対象文に適用可能な文翻訳ルールと部分翻訳ルールを選択する．適用可能な文及び部分翻訳ルールとは，翻訳ルールの原言語部が翻訳対象文に表層レベルで一致するもの，及び翻訳ルールの原言語部の変数を除いた部分が入力文と同じ並びで全て含まれているものを指す．次に適用可能な文翻訳ルールと部分翻訳ルールを組み合わせることにより，翻訳結果を生成する．

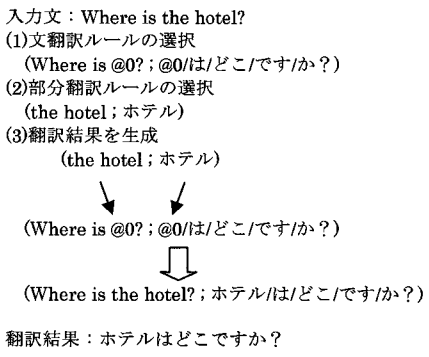

図２　翻訳結果生成の具体例

翻訳結果が複数生成された場合は，以下に示す具象度の高い文翻訳ルールが適用された翻訳結果を優先する．

$$\text{具象度（\%）}=\frac{\text{変数を除く単語数}}{\text{単語総数}}\times 100.0$$

更に，具象度が等しい文翻訳ルールが適用された翻訳結果が複数存在する場合は，以下に示す確実度の高い文翻訳ルールが適用された翻訳結果を優先する．

$$\text{確実度（\%）}=\frac{\text{正確実度数}}{\text{正確実度数}+\text{誤確実度数}}\times 100.0$$

文翻訳ルールの正確実度数及び誤確実度数については，２．２節で詳細を述べる．

図２に翻訳結果生成の具体例を示す．入力文"Where is the hotel?"に対する適用可能なルールとして，文翻訳ルール(Where is @0? ; @0/は/どこ/です/か？)，部分翻訳ルール(the hotel ; ホテル)が選択される．これらの翻訳ルールを組み合わせ，文翻訳ルール(Where is the hotel? ; ホテル/は/どこ/です/か？)が生成される．その結果，この文翻訳ルールの目的言語部である"ホテルはどこですか？"が翻訳結果となる．

２．２　フィードバック部

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う．まず生成された翻訳結果と正しい翻訳結果を表層レベルで比較する．比較の結果，一致した場合は適用された翻訳ルールの正確実度数を１増加させる．また，一致しない場合は適用された翻訳ルールの誤確実度数を１増加させる．

翻訳例
(How can I get to the hotel? ;
その/ホテル/へ/は/どう/やっ/て/行く/の/です/か？)
(This is the hotel where I'll be staying. ;
この/ホテル/に/滞在/し/ます．)

⇩　共通部分を抽出

共通部分
(the hotel ; ホテル) ⇒ 部分翻訳ルールとして獲得

抽出部分を変数におきかえたもの
(How can I get to @0? ;
その/@0/へ/は/どう/やっ/て/行く/の/です/か？)
(This is @0 where I'll be staying. ;
この/@0/に/滞在/し/ます．)
⇒ 文翻訳ルールとして獲得

図３　帰納的学習による翻訳ルール獲得の具体例

２．３　学習部

学習部では，帰納的学習を用いることで翻訳例対から翻訳ルールを獲得する．更に，本手法により抽出ルールを獲得すると共に，それを用いることで，個々の翻訳例から効率よく翻訳ルールを獲得する．

２．３．１　帰納的学習を用いた翻訳ルールの獲得

帰納的学習では，翻訳例対において，共通部分と差異部分を多段階に抽出することにより，翻訳ルールを獲得する．処理過程を以下に示す．

（１）２つの翻訳例対の原言語文間と目的言語文間それぞれにおいて共通部分と差異部分を決定する．本稿における共通部分とは，表層的に一致する１つ以上の単語から構成される単語列である．共通部分は，共通部分の位置と長さから決定する[8]．そして共通部分以外の単語列を差異部分とする．

（２）決定された共通部分と差異部分において，原言語文と目的言語文間の共通部分の組み合わせ，及び差異部分の組み合わせを考え，組み合わせの数の少ない方を部分翻訳ルールとして獲得する．

（３）抽出された部分を変数に置き換えることにより，文翻訳ルールを獲得する．このようにして獲得された翻訳ルール間からも多段階に共通部分と差異部分を抽出することで，より一般化された翻訳ルールを獲得する．

帰納的学習による翻訳ルール獲得の具体例を図３に示す．２つの翻訳例の共通部分として，原言語部分においては"the hotel"，目的言語部分においては"ホテル"と決定する．翻訳例(How can I get to the hotel? ; その/ホテル/へ/は/どう/やっ/て/行く/の/です/か？)における，共通部分として原言語文では，"the hotel"，目的言語文では"ホテル"が該当するので組み合わせは１：１になる．一方，差異部分として原言語文では"How can I get to"，目的言語文では"その"，"へ/は/どう/やっ/て/行く/の/です/か"が該当するので組み合わせは１：２となる．そこで組み合わせの少ない共通部分(the hotel ; ホテル)を抽出し，部分翻訳ルールとして獲得する．次に，抽出された部分を変数に置き換えることにより(How can I get to @0? ; その/@0/へ/は/どう/やっ/て/行く/の/です/か？)を文翻訳ルールとして獲得する．もう一方の翻訳例(This is the hotel where I'll be staying. ; この/ホテル/に/滞在/し/ます．)に対しても同様に(the hotel ; ホテル)を抽出し，抽出された部分を変数に置き換えることにより (This is @0 where I'll be staying. ; この/@0/に/滞在/し/ます．)を文翻訳ルールとして獲得する．

２．３．２　省略可能性による翻訳ルールの獲得

本手法により獲得される抽出ルールは２種類である．１つは，(a @0 room ; @0/部屋)，(a @0 ; @0)のように，それ自体も抽出部分に含まれる抽出ルールである．本稿ではこれらを第１抽出ルールと呼ぶ．もうひとつは(with @0 ; @0/付き/の)のように変数に対

翻訳例
(I'd like a quiet room with a bath.;
バス/付き/の/静か/な/部屋/を/お願い/し/ます.)
部分翻訳ルール
(a room;部屋)
(1)原言語文からの抽出
a quiet room
(2)目的言語文からの抽出
①左側探索
バス/付き/の/静か/な/部屋/を/お願い/し/ます.
P(部屋|静か)=0

バス/付き/の/静か/な/部屋/を/お願い/し/ます.
P(部屋|の)=0.06>0⇒"静か/な"は省略可能
抽出:静か/な/部屋,の/静か/な/部屋

バス/付き/の/静か/な/部屋/を/お願い/し/ます.
P(部屋|付き)=0⇒左側探索を終了

②右側探索
バス/付き/の/静か/な/部屋/を/お願い/し/ます.
P(お願い|部屋)=0.01>0⇒"を"は省略可能
抽出:/部屋/を,部屋/を/お願い

バス/付き/の/静か/な/部屋/を/お願い/し/ます.
P(し|部屋)=0⇒右側探索を終了
(3)類似度を用いて組み合わせを選択

組み合わせ	類似度
(a quiet room;静か/な/部屋)	**0.66**
(a quiet room;の/静か/な/部屋)	0.63
(a quiet room;部屋/を)	0.42
(a quiet room;部屋/を/お願い)	0.37

⇩
選択された組み合わせ
(a quiet room;静か/な/部屋)
(4)抽出ルールの獲得
quiet,静か/な⇒変数化
⇩
(a @0 room;@0/部屋)
⇒第1抽出ルールとして獲得

図4　第1抽出ルール獲得の具体例

応する部分のみを抽出対象とする抽出ルールである.これらを第2抽出ルールと呼ぶ.

このような抽出ルールを用いることで翻訳例中の原言語文と目的言語文間における句レベルの対応関係を決定する.その結果,より汎用性の高い翻訳ルールを効率よく獲得できると考えられる.

(1)第1抽出ルールの獲得

第1抽出ルールの獲得の処理過程を以下に示す.

i)部分翻訳ルールの単語が全て含まれているような翻訳例を選択する.このとき翻訳例において,原言語文もしくは目的言語文いずれかの共通部分は2つでなければならない.

ii)翻訳例の原言語文と部分翻訳ルールの原言語部との間,かつ翻訳例の目的言語文と部分翻訳ルールの目的言語部との間それぞれにおいて次の処理を行う.

iii)共通部分が2つの場合,2つの共通部分で挟まれた部分を共通部分と共に抽出する.

iv)共通部分が1つの場合,共通部分の両側方向に省略可能な部分を探索するために次の処理を行う.

a) 左側探索では,共通部分と共通部分の左側から2つ目に位置する単語とのbigram確率を求める.その場合,過去に入力された全翻訳例を用いてbigram確率を求める.bigram確率が0の場合は,共通部分の左側から3つ目,4つ目と対象単語を1単語ずつずらしbigram確率が0より大きくなる対象単語が出現するまで続ける.対象単語が決定されると,共通部分と対象単語で挟まれた部分を共通部分と共に抽出する.更に,共通部分と対象単語に挟まれた部分を,共通部分と対象単語と共に抽出する.この処理を,bigram確率が0となる対象単語が出現するまで続ける.また,bigram確率が0の単語のみしか存在しない場合は,文頭までを抽出する.

b) 右側探索では,左側探索と同様の処理を,共通部分の右側方向に対して行う.

v)抽出された原言語部分と目的言語部分との間で単語分割方式による類似度を求め,閾値以上の類似度を持つ組み合わせを選択する.閾値以上の組み合わせが複数存在した場合,類似度が最大のもののみを選択する.単語分割方式による類似度は2.3.2節の(5)で詳細を述べる.

vi)選択された組み合わせにおいて,共通部分以外の部分を省略可能な部分として,変数に置き換えることで第1抽出ルールを獲得する.

vii)獲得された第1抽出ルール間において,帰納的学習を適用することで,より汎用的な第1抽出ルールを獲得する.例えば(a @0 room;@0/部屋)と(a @0 hotel;@0/ホテル)より(a @0;@0)を獲得する.

部分翻訳ルールを用いた第1抽出ルール獲得の具体例を図4に示す.翻訳例と部分翻訳ルールの共通部分は原言語文において"a"と"room",目的言語文においては"部屋"となる.原言語文においては,共通部分"a"から"room"までの"a quiet room"を抽出する.目的言語文においては共通部分"部屋"の両側方向に省略可能な部分を探索する.まず,左側探索を行う."部屋"の左側2つ目に位置する"静か"と"部屋"のbigram確率を求め,"静か"と"部屋"の間"な"が省略可能かどうかを判定する.この場合はP(部屋|静か)は0となるため省略不可とみなし対象を次の単語へずらす."の"と"部屋"のbigram確率は0より大きいため,"静か/な"を省略可能と考え,"静か/な/部屋"と"の/静か/な/部屋"を抽出する.次の単語へ対象をずらし,"付き"と"部屋"のbigram確率は0であるため,左側探索を終了する.続いて右側探索を行う."部屋"と"お願い"のbigram確率は0より大きいため,"を"を省略可能と考え,"部屋/を"と"部屋/を/お願い"を抽出する.次の単語へ対象をずらし,"部屋"と"し"

の bigram 確率は0であるため，右側探索を終了する．抽出された原言語部分と目的言語部分との間で単語分割方式による類似度を求める．その結果，閾値以上の類似度をもつ組み合わせである(a quiet room ; 静か/な/部屋)が選択され，"quiet"と"静か/な"を変数に置き換えることで，(a @0 room ; @0/部屋)を第1抽出ルールとして獲得する．今回は選択する組み合わせの類似度の閾値として 0. 65 を用いている．

vii）このようにして獲得される(a @0 room ; @0/部屋)，(a @0 hotel ; @0/ホテル)，(the @0 hotel ; @0/ホテル)などの第1抽出ルール間に帰納的学習を適用し，より汎用的な(a @0 ; @0)，(the @0 ; @0)といった第1抽出ルールを新たに獲得する．

（2）第1抽出ルールを用いた翻訳ルールの獲得

第1抽出ルールを用いた翻訳ルールの獲得の処理過程を以下に示す．

i）第1抽出ルールの変数以外の単語が全て含まれているような翻訳例を選択する．

ii）翻訳例の中で，第1抽出ルールの変数に対応する部分を決定し，抽出する．変数に対応する部分を決定するために，翻訳例の原言語文と第1抽出ルールの原言語部との間，もしくは，翻訳例の目的言語文と第1抽出ルールの目的言語部との間それぞれにおいて，以下の処理のいずれかを行う．

a) 第1抽出ルールの原言語部もしくは目的言語部のいずれかにおいて，変数の両側が共通部分である場合，翻訳例において2つの共通部分に挟まれた部分と共通部分を共に抽出する．

b) 第1抽出ルールの原言語部もしくは目的言語部のいずれかにおいて，変数の片側のみに，共通部分がある場合，翻訳例の共通部分の片側方向を探索範囲とし，類似度を用いて単語の対応関係を決定し，対応部分と共通部分を共に抽出する．

c) 第1抽出ルールの原言語部もしくは目的言語部のいずれかにおいて，変数のみの第1抽出ルールの場合は，翻訳例の全てを探索範囲とし，類似度を用いて単語の対応関係を決定し，対応部分と共通部分を共に抽出する．

図5に上記の b) と c) の組み合わせである第1抽出ルールを用いた翻訳ルール獲得の具体例を示す．第1抽出ルール(a @0 ; @0)を適用する場合，原言語文においては処理過程の b) により，翻訳例の原言語文中の"a"の右側方向が探索範囲となる．目的言語文においては処理過程の c) により，全てが探索範囲となる．まず，"a" の右側に隣接する単語"reservation"と目的言語文の単語全てとの類似度を求め，類似度が一番高い単語である"予約"を対応させる．そして（a reservation ; 予約）を抽出することで，部分翻訳ルールを獲得し，更に(I made @0 in Tokyo. ; 東京/で/@0/し/ました.)を文翻訳ルールとして獲得する．次に"reservation"の右側の"in"と目的言語文の単語全てとの類似度を求めるが，類

翻訳例

(I made a reservation in Tokyo. ;
東京/で/予約/し/ました.)

第1抽出ルール

(a @0 ; @0)

(1) "reservation" に対応する単語を決定する．

	東京	で	予約	し	まし	た
reservation	0.19	0.12	**0.73**	0.12	0.17	0.15

(a reservation ; 予約)

⇒ 部分翻訳ルールとして獲得

(I made @0 in Tokyo. ; 東京/で/@0/し/ました.)

⇒ 文翻訳ルールとして獲得

(2) "in" に対応する単語を決定する．

	東京	で	予約	し	まし	た
in	0.16	0.14	0.14	0.13	0.12	0.11

"in" と類似度が高い単語がないため，探索を終了．

図5　第1抽出ルールを用いた翻訳ルール獲得の具体例

文翻訳ルール

(I'd like a room with @0. ;
@0/付き/の/部屋/に/し/たい.)

(One with @0. ; @0/付き/の/を/お願い/し/ます.)

⇩ 変数と，変数に隣接する共通部分を抽出

共通部分

(with @0 ; @0/付き/の) ⇒ 第2抽出ルールとして獲得

図6　第2抽出ルール獲得の具体例

似度の高い単語がないため探索を終了する．今回は，対応関係を決定するための類似度の閾値として 0. 35 を用いている．

（3）第2抽出ルールの獲得

第2抽出ルールの獲得の処理過程を以下に示す．

i）（2）で得られた文翻訳ルールの中から，原言語文，目的言語文の両方において，変数に隣接する部分が共通部分である文翻訳ルール対を選択する．

ii）変数と変数に隣接する共通部分を共に抽出し第2抽出ルールとして獲得する．

図6に第2抽出ルールの獲得の例を示す．第1抽出ルールを適用することにより獲得された文翻訳ルールの中から，変数に隣接する共通部分を持つ文翻訳ルール対を選択する．次に，変数と変数に隣接する共通部分(with @0 ; @0/付き/の)を抽出し，第2抽出ルールを獲得する．

（4）第2抽出ルールを用いた翻訳ルールの獲得

第2抽出ルールを用いた翻訳ルールの獲得の処理過程を以下に示す．

ｉ）第２抽出ルールの変数以外の単語が全て含まれているような翻訳例を選択する．
ii）翻訳例の中で，第２抽出ルールの変数に対応する部分を決定し，抽出する．変数に対応する部分を決定するために，以下の処理のいずれかを行う．

a) 第２抽出ルールの原言語部もしくは目的言語部のいずれかにおいて，変数の両側が共通部分である場合，翻訳例において２つの共通部分に挟まれた部分を抽出する．

b) 第２抽出ルールの原言語部もしくは目的言語部のいずれかにおいて，変数の片側のみに，共通部分がある場合，翻訳例の共通部分の片側方向を探索範囲とし，類似度を用いて単語の対応関係を決定し，対応部分を抽出する．

図７に第２抽出ルールの原言語部，目的言語部が共に b）に該当する場合の翻訳ルールの獲得の具体例について示す．第２抽出ルール(with @0；@0/付き/の)を適用する場合，原言語文においては処理過程の b）により，翻訳例の原言語文中の"with"の右側方向が探索範囲となる．目的言語文においては処理過程の b）により，翻訳例の原言語文中の"付き/の"の左側方向が探索範囲となる．"with"の右側に隣接する単語"shower"と"付き/の"の左側に隣接する"シャワー"の類似度を求める．類似度が閾値以上であるため(shower；シャワー)を抽出し，部分翻訳ルールとして獲得し，(I'd like a room with @0. ；@0/付き/の/部屋/に/し/たい．)を文翻訳ルールとして獲得する．ここでも，（２）と同様に対応関係を決定するための類似度の閾値として 0．35 を用いている．

（５）単語分割方式による類似度

本稿では，出現頻度の低い単語列間の類似度を求めるために，単語列を単語に分割して類似度を求める．この方法はデータスパースネスなコーパスに対して有効と考えられる．個々の単語の類似度は，式(1)の Dice 係数によって求められる．

$$sim(W_s,W_t)=\frac{2\times f_{st}}{f_s+f_t} \qquad (1)$$

f_sは単語 W_sが原言語文に独立で出現する頻度，f_tは単語 W_tが目的言語文に独立で出現する頻度，さらに f_{st}は単語 W_s,W_tが対訳文中に共出現する頻度を示している．単語列同士の類似度は，以下の手順で求める．

ｉ）原言語部と，目的言語部の各単語間の類似度を Dice 係数により求める．

ii）原言語部の各単語に対し，類似度の高い目的言語部の単語を対応させる．対応させた際の類似度の平均値を目的言語部に対する原言語部の類似度とする．

iii）同様に目的言語部の各単語に対し，類似度の高い原言語部の単語を対応させる．対応させた際の類似度の平均値を原言語部に対する目的言語部の類似度とする．

iv）ii）iii）で求めた２つの値の平均値を原言語部と目的言語部の単語分割方式による類似度とする．

翻訳例
(I'd like a room with shower.；
シャワー/付き/の/部屋/に/し/たい．)
第２抽出ルール
(with @0；@0/付き/の)
(1) "shower" に対応する単語を決定する．

	シャワー
shower	**1.00**

(shower ；シャワー)
⇒ 部分翻訳ルールとして獲得
(I'd like a room with @0.；
@0/付き/の/部屋/に/し/たい．)
⇒ 文翻訳ルールとして獲得

図７　第２抽出ルールを用いた
翻訳ルール獲得の具体例

(1)各単語間の類似度を Dice 係数により求める．

	静か	な	部屋		
a	0.02	0.04	0.24	→	部屋：0.24
quiet	1	0.3	0.07	→	静か：1.00
room	0.05	0.11	0.72	→	部屋：0.72

↓ ↓ ↓ ⇩

quiet：1.00　quiet：0.3　room：0.72　平均 0.65

⇩

平均 0.67

(2)求められた２つの値の平均値を単語分割方式による類似度とする．

$$sim(a\ quiet\ room, 静か/な/部屋)=\frac{0.67+0.65}{2}=0.66$$

図８　単語分割方式による類似度の具体例

図８に"a quiet room"と"静か/な/部屋"の単語分割方式による類似度の例を示す．まず，各単語間の類似度を Dice 係数により求める．次に，原言語部の各単語に対し，類似度が高い目的言語部の単語を対応させる．"a"には，"部屋"が対応し，"quiet"，"room"にはそれぞれ"静か"，"部屋"が対応する．このように対応させた単語間の類似度の平均値である 0．65 を，目的言語部に対する原言語部の類似度とする．同様に目的言語部の各単語に対し，類似度が高い原言語部の単語を対応させる．"静か"には，"quiet"が対応し，"な"，"部屋"にはそれぞれ"quiet"，"room"が対応する．このように対応させた単語間の類似度の平均値である 0．67 を，目的言語部に対する原言語部の類似度とする．このようにして求めた２つの値の

平均値である 0．66 を原言語部と目的言語部の単語分割方式による類似度とする．

3．性能評価実験

3．1　実験方法

実験データには旅行用英会話テキスト 10 冊（翻訳例 1,710 文）を使用した．そのうち 948 文を学習データとし，残りの 762 文を評価データとして用いた．実験は 2．3．2 で述べた省略可能性による翻訳ルールの獲得処理を適用したシステムと，適用しないシステムを用い，比較を行った．手順としては，まず辞書が空の状態にし，1 文ずつ翻訳と学習を行った．

3．2　評価基準

システムが生成した翻訳結果は有効な翻訳結果と無効な翻訳結果に分類される．有効な翻訳結果とは以下のいずれかを指している．

（1）未登録語を含まない正しい翻訳結果
（2）未登録語を含む正しい翻訳結果

未登録語とは，名詞，形容詞，及び名詞句に相当するもので，これらが未登録語の場合は，訳語を与えることにより未登録語を含まない正しい翻訳結果を容易に得られる．システムが翻訳結果を複数生成した場合は，2．1 節の順位付けにより 1 位になった翻訳結果のみを評価の対象とした．

3．3　実験結果

表 1 に省略可能性による翻訳ルールの獲得処理を適用した場合，表 2 に適用しなかった場合の実験の結果を示す．省略可能性による翻訳ルールの獲得処理を適用することで有効な翻訳率は 5．5 ポイント増加した．また，それぞれの有効な翻訳率の推移を図 9 に示す．

3．4　考察

評価データ 762 文について，省略可能性による翻訳ルールの獲得処理を適用することで，有効な翻訳結果になったのは 48 文，逆に無効な翻訳になったのは 6 文であった．それぞれの翻訳結果の具体例を表 3 に示す．結果としては，省略可能性による翻訳ルールの獲得処理を適用することにより，有効な翻訳結果は 42 文増加したこととなる．これは，省略可能性による翻訳ルールの獲得処理を適用することで，良質な翻訳ルールが増大したためである．この結果より，本手法が有効であることが明らかとなった．省略可能性による翻訳ルールの獲得処理を適用することで，無効な翻訳になった 6 文については，文翻訳ルールは正しいが，誤った部分翻訳ルールが適用されていたものが 5 文あり，誤った文翻訳ルールが適用されていたものが 1 文であった．

4．おわりに

本稿では，学習型機械翻訳手法において，翻訳例中の省略可能な部分に着目することにより，翻訳ルールを効率よく獲得するための手法を提案し，その有効性について述べた．性能評価実験の結果，省略可能性による翻訳ルールの獲得処理を適用することで有効な翻訳率が 5．5 ポイント増加した．この結果は，本手法の有効性を示すものである．

今後は，精度向上のための更なる改良を検討する．そして，より学習能力の高い学習型機械翻訳システムの構築に向けて研究を進める予定である．

表 1　省略可能性による翻訳ルールの獲得処理を適用した場合の翻訳結果

有効な翻訳率	内訳	
	未登録語なし	未登録語あり
55．4%(422)	71．3%(301)	28．7%(121)

表 2　省略可能性による翻訳ルールの獲得処理を適用しなかった場合の翻訳結果

有効な翻訳率	内訳	
	未登録語なし	未登録語あり
49．9%(380)	73．2%(278)	26．8%(102)

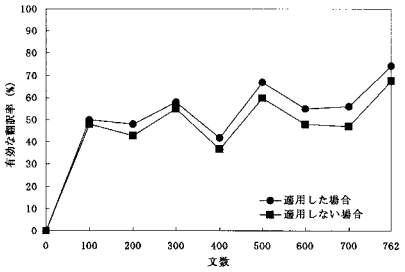

図 9　有効な翻訳率の推移

参考文献

[1] 野村浩郷(編)，言語処理と機械翻訳，講談社,東京，1991.

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Jhon D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation", Computational Linguistics, Vol. 16, no. 2, pp. 79-85, 1990.

[3] 佐藤理史(著)，日本認知科学会(編)，アナロジーによる機械翻訳，共立出版，東京，1997

[4] 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.

表３　翻訳結果の具体例

省略可能性による翻訳ルールの獲得処理の適用による有効性の具体例		
英文	省略可能性による翻訳ルールの獲得処理を適用した場合の翻訳結果	省略可能性による翻訳ルールの獲得処理を適用しなかった場合の翻訳結果
Where can I catch the bus?	バス乗り場はどこですか？	@0 バスはどこですか？
I'd like a room with shower.	シャワー付きの部屋をお願いします.	付きのシャワー部屋をお願いします.
省略可能性による翻訳ルールの獲得処理の適用による誤りの具体例		
英文	省略可能性による翻訳ルールの獲得処理を適用した場合の翻訳結果	省略可能性による翻訳ルールの獲得処理を適用しなかった場合の翻訳結果
I'm a tourist.	私は案内です.	私は@0 です.

[5] 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, pp1565-1579, 1996.

[6] 越前谷博, 荒木健治, 桃内佳雄, 栃内香次, “翻訳例に基づく再帰チェーンリンク型学習による機械翻訳手法”, 電子情報通信学会論文誌 DII, Vol. J85-D-II, No. 12, pp. 1840-1852, 2002.

[7] 荒木健治(著), 自然言語処理ことはじめ−言葉を覚え会話のできるコンピューター, 森北出版, 東京, 2004

[8] Hiroshi Echizen-ya, Kenji Araki, “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum”, Proceedings of the Eleventh Machine Translation Summit (MT SUMMIT XI), pp. 151-158, Copenhagen, Denmark, 2007.