統計的機械翻訳における地名の汎化の影響

関 拓也 , 山本 和英
長岡技術科学大学 電気系
E-mail:{seki,ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

現在、機械翻訳の分野では、翻訳知識を自動的に獲得する統計的機械翻訳が注目されている。統計的機械翻訳は、大量の対訳コーパスを統計的に解析することによって単語対応、単語の並びなどの情報を獲得し、それらの知識を翻訳知識として利用している。統計情報を利用して翻訳モデルを作成するため、構文や意味等の知識を必要としないことが特徴である。統計的機械翻訳には単語単位の翻訳とその並べ替えによって翻訳を行う単語に基づく翻訳モデルと、句単位の翻訳とその並べ替えによって翻訳を行う句に基づく翻訳モデル [1] がある。

統計翻訳において大量の対訳コーパスを学習に用いたとしても、全ての単語の対訳を網羅することは難しい。対訳コーパスにとっての未知語が存在する文を翻訳した時、翻訳器は未知語を翻訳することはできない。また、学習に存在しない語であるために、その単語を文中の適切な位置に配置することも難しい。

本研究では、学習データにとって未知となる可能性の高い語を含むカテゴリとして地名に注目した。そして、地名が未知であることによって翻訳精度が低下してしまうという問題に対応するための手法を提案した。

2　関連研究

大熊ら [3] は、学習コーパスに含まれない知識として、対訳辞書等の外部の知識を、既に構築されている翻訳システムの中に導入する手法を提案している。大熊らの手法を以下、既存手法と呼ぶ。既存手法は、入力文に地名を示す未知語が含まれる場合に、未知の地名の単語を学習コーパスに頻出する地名の単語に置き換えて翻訳を行う。その後に辞書を用いて、地名の単語を置き換えて翻訳した部分を、目的の地名の単語に翻訳するという方法をとり、未知語を含む文の翻訳の精度の改善を行っている。既存手法を例 1 に示す。

例 1
- 条件
  - 「**長岡**」は**未知の地名**。
  - 学習コーパスで**最も頻出する地名**は「**東京**」。
- 入力：**長岡** へ 行く 。
  ↓　未知の地名の単語の置き換える。
- 置換後：**東京** へ 行く 。
  ↓　翻訳を行う。
- 翻訳後：I go to **tokyo** .
  ↓　対訳辞書を用いて目的の地名の単語へ変換する。
- 最終出力：I go to **nagaoka** .

しかし、未知語を頻出する単語で順番に置き換えていくだけであるので、学習された知識が十分に利用されない場合も出てくる。例 2 に示す場合が考えられる。

例 2
- 入力：**長岡** へ 行く 。
- 条件
  - 「**長岡**」は**未知の地名**。
  - 学習コーパスで**最も頻出する地名**は「**東京**」。
  - 「**東京 へ 行く 。**」という句は学習コーパスに**存在しない**。
  - 「**大阪 へ 行く 。**」という句は学習コーパスに**存在する**。

例 2 の場合、既存手法では「長岡」を「東京」と変換して翻訳を行う。しかし、句内での語順が保たれていることを考慮すると、可能な限り長い句を用いて翻訳を行った方が、翻訳精度は改善すると考えられる。Koehn ら [1] は、翻訳に用いる句の単語数の上限と翻訳精度の関係から、長い句を用いることにより、翻訳精度が向上するとしている。そうした場合、「長岡」を「東京」と変換するよりも「大阪」と変換した方が、より長い句を用いて翻訳を行えるため、良い翻訳精度を得られるのではないかと考えた。既存手法では、例 2 のような場合に、翻訳知識を十分に利用できない。

本手法では、この問題に対応するための方法として地名の汎化を行った。本手法では、例 1 の場合には、翻訳に使用する「大阪へ行く。」という句を「PLACE へ行く。」と汎化を行う。そして、入力文も「長岡へ行く。」から「PLACE へ行く。」と汎化し、翻訳を行うことによって、より長い句を用いた翻訳が可能となり、翻訳精度が改善するのではないかと考えた。

3　提案手法

3.1　手法概要

本稿では、通常の翻訳知識を用いた翻訳モデルを通常翻訳モデルと呼び、提案手法によって得た翻訳知識を用いた翻訳モデルを汎化翻訳モデルと呼ぶ。図 1 に本手法の汎化翻訳モデルの構築手順を示す。

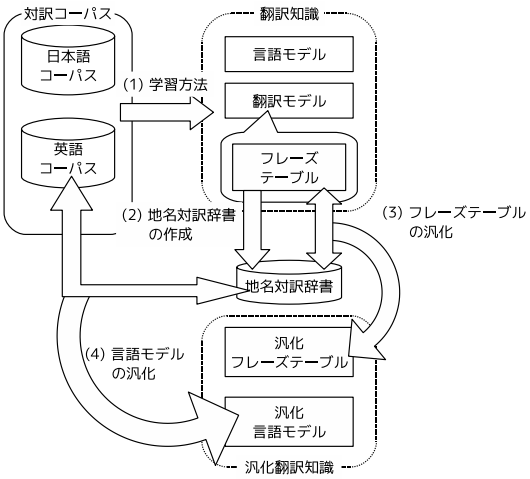

図 1: 汎化翻訳モデルの構築

本手法では、統計的機械翻訳のデコーダとして Moses[1] を使用する。Moses は翻訳知識として言語モデル及び、フレーズテーブルに登録された情報を利用して翻訳を行う。言語モデルとしては N-gram モデルを使用する。フレーズテーブルは翻訳モデルの学習によって得られる。フレーズテーブルには、対訳となる句、句中の単語同士のアライメント情報、翻訳確率、語彙の重み、句のペナルティの情報が登録されている。本稿では、フレーズテーブル、言語モデルの 2 つの翻訳知識を汎化したものを汎化翻訳知識と呼ぶ。提案手法では、この汎化翻訳知識を用いて翻訳を行う。

本手法では、対訳となる地名を、統計情報と形態素情報を用いて取得する。その後、得られた地名の対訳を用いて翻訳知識の汎化を行う。3.2 節以降に、処理の詳細を示す。

3.2 学習方法

本稿では、単語アライメントツールとして GIZA++[2]、N-gram 言語モデルツールとして IRST LM[3] を使用する。N-gram 言語モデルは、英語の原形と品詞、それぞれについて作成する。猪澤ら [5] の評価実験の結果を参考として単語 5-gram の言語モデルを作成した。

翻訳モデルには Moses の factored 翻訳モデルを用いる。factored 翻訳モデルでは、英語の表層形だけではなく、原形、品詞、活用等の要素も考慮した翻訳が可能となる。このため、表層形のみの学習と比べて、ゼロ頻度問題が生じ難くなることがこの翻訳モデルの利点である。本手法では、地名の対訳を得るために日本語の原形と品詞、英語の原形と品詞情報を利用している。そこで、「日本語の原形と品詞」-「英語の原形と品詞」の対訳文を用いて学習を行う。日本語の形態素解析器として ChaSen[4]、英語の形態素解析器として TreeTagger[5] を用いる。

3.3 地名対訳辞書の作成

この節では、翻訳知識の地名を汎化するために利用する、地名対訳辞書の作成方法を説明する。ここでは、対訳となる地名を得るために、フレーズテーブル中の対訳となる句と、翻訳確率の情報を用いる。まず、日本語の句の中から、「名詞-固有名詞-地域-一般」または、「名詞-固有名詞- 地域-国」の品詞タグのついた単語、1 語からなる句を選択する。さらに、その日本語の句と対訳となる英語の句が「N(名詞)」または「J(形容詞)」の 1 単語であるものを選択する。選択された対訳句を地名の対訳として辞書に登録する。例 3 に地名対訳辞書に登録される対訳の例と、登録されない単語を示す。

例 3
地名対訳辞書に登録される対訳:
日本 | 名詞-固有名詞-地域-国 | | | japan|NN
アジア | 名詞-固有名詞-地域-一般 | | | asian|JJ
地名対訳辞書に登録されない対訳:
アメリカ | 名詞-固有名詞-地域-国 | | | the|DT u.|JJ s.|NN

しかし、この状態では、対訳とならない対も多く含んでいるため、対訳とならない対を除外する。まず、日-英翻訳確率と英-日翻訳確率の積が 0.01 未満の対訳句を除外する。次に、日英対訳辞書を用いて地名ではない単語を除外する。対訳の日本語が日英対訳辞書の地名に含まれない対訳句を除外する。日英対訳辞書として、英辞郎 [7] を用いた。英辞郎の単語の内、地名、地名-1、地名-2、地名- 3、国名、国名-1、国名-2、州名のタグが付与された対訳であり、かつ、ChaSen の形態素解析によって、「名詞-固有名詞-地域-一般」または「名詞-固有名詞-地域-国」の品詞タグが付与された対訳を用いた。以降、英辞郎から作成した地名の対訳辞書を英辞郎地名対訳辞書と呼ぶ。以上の処理を行い得られた地名の対訳辞書を地名対訳辞書と呼び、この辞書を用いて翻訳知識の汎化を行う。

3.4 フレーズテーブルの汎化

3.3 節で作成した地名対訳辞書を用いてフレーズテーブルの汎化を行う。

フレーズテーブル中の対訳句に、地名対訳辞書の対訳が含まれている場合、その地名の単語の部分を汎化する。日本語は、原形と品詞が一致したものを汎化する。原形を「PLACE」、品詞を「PLACE」と汎化する。英語は、地名対訳辞書の原形の情報のみを用いて汎化を行う。原形を「PLACE」と汎化する。英語の汎化先の品詞は、汎化前の単語に付与されている品詞を基にして決定する。すなわち、汎化前の品詞が、「N(名詞)」の時は「NN-PLACE」、「J(形容詞)」または、それ以外の品詞の時は「JJ-PLACE」と汎化する。日本語の地名を翻訳した場合、英語中では、「N」か「J」に分類されることが多いと考え、この 2 種類の品詞に汎化することとした。

汎化の例を示す。

例 4:
日本語の原文:
パリ | **名詞-固有名詞-地域-一般**　へ | 助詞-格助詞-一般
行く | 動詞-自立　たい | 助動詞　。| 記号-句点
英語の原文:
i|NP　wish|VVP　to|TO　go|VV　to|TO
paris|**NNS**　.|SENT

日本語の汎化:
PLACE|**PLACE**　へ | 助詞-格助詞-一般
行く | 動詞-自立　たい | 助動詞　。| 記号-句点
英語の汎化:
i|NP　wish|VVP　to|TO　go|VV　to|TO
PLACE|**NN-PLACE**　.|SENT

3.5 言語モデルの汎化

3.3 節で作成した地名対訳辞書を用いて英語の言語モデルの汎化を行う。

地名対訳辞書に登録された単語が学習コーパス中に存在するかどうかを検索する。存在した場合、その単語の部分を汎化する。原形が一致する部分を汎化する。汎化方法は 3.4 節と同様に「PLACE|NN-PLACE」または「PLACE|JJ-PLACE」とする。

3.6 チューニング

通常翻訳モデルと汎化翻訳モデル、それぞれでチューニングを行う。チューニングに用いる対訳文は、学習コーパスに含まれない対訳文から作成する。100 文対の対訳文を用いてチューニングを行う。対訳文中の全ての地名対訳が、英辞郎地名対訳辞書に登録されている対訳文を、チューニング用の対訳文として採用した。得られたチューニング用の対訳文を加工せず、そのままの状態とした対訳文を通常チューニングデータと呼ぶ。日本語の地名の単語部分を「PLACE|PLACE」と変換し、英語の地名の単語部分を「PLACE」と変換したチューニング用の対訳文を汎化チューニングデータと呼ぶ。通常翻訳モデルは、通常チューニングデータを用いてチューニングを行う。汎化翻訳モデルは、汎化チューニングデータを用いてチューニングを行う。

チューニングを行った翻訳モデルを用いて評価実験を行った。

4 評価実験

本稿では、対訳コーパスとして、CREST[6] コーパス 374,085 文対を用いた。そのうち、372,985 文対を翻訳モデル及び言語モデルの学習に用いた。残りの 1,100 文対の内、100 文対をチューニングに用い、1,000 文対を open の評価データとした。評価、及びチューニングに用いる文は、日本語文の単語数が 5 単語以上のものを採用した。

評価は、通常翻訳と未知語翻訳、既存手法、提案手法の 4 つの翻訳結果の比較によって行った。各翻訳手法の詳細は 4.1 節で述べる。翻訳結果の評価は、BLEU[2] による自動評価と人間による主観評価を行った。

4.1 評価データの作成

評価データは closed と open の対訳文から、それぞれ 1,000 文対ずつ作成した。3.6 節の通常チューニングデータの作成方法を用いて作成された評価データを通常評価データと呼ぶ。汎化チューニングデータの作成方法を用いて、作成された評価データを汎化評価データと呼ぶ。

4.1.1 節から 4.1.4 節で、評価を行う各翻訳方法について説明を行う。

4.1.1 通常翻訳

通常翻訳では、通常翻訳モデルを用いて翻訳を行う。評価データは通常評価データを用いる。翻訳後に英辞郎地名対訳辞書及び、

地名対訳辞書を用いて、翻訳結果の地名の単語部分を「PLACE」と置き換えた。

4.1.2 未知語翻訳

未知語翻訳では、通常翻訳モデルを用いて翻訳を行う。評価データは汎化評価データを用いる。汎化評価データでは地名部分が「PLACE|PLACE」となっている。「PLACE|PLACE」は学習コーパスに含まれない単語であるから、通常翻訳モデルにとっては未知語となる。そこで、この汎化評価データを通常翻訳モデルに翻訳を行わせる方法を、未知語翻訳とした。

4.1.3 既存手法

既存手法として、大熊らの手法を行った。既存手法では、通常翻訳モデルを用いて翻訳を行う。既存手法では、通常評価データ中の地名を学習コーパス内で頻出する地名で置き換えた文を評価データとする。複数個の地名が入力文に存在する場合には、同じ地名が一文中に重複して出現しないように、学習コーパスにおける出現頻度の高い地名で、順番に未知の地名の置き換えを行う。そして、翻訳後に英辞郎地名対訳辞書を使用して地名を目的の地名に置き換える。今回は翻訳後に地名の部分を「PLACE」と置き換えた。既存手法では、置き換えを行う地名を、日本語の形態素解析によって「名詞-固有名詞-地域-一般」の品詞タグが付与された単語としていたが、本稿では「名詞-固有名詞-地域-一般」及び「名詞-固有名詞-地域- 国」とした。

4.1.4 提案手法

提案手法では、汎化翻訳モデルを用いて翻訳を行う。評価データは汎化評価データを用いる。

4.2 正解データとの比較方法

通常翻訳を除く翻訳方法では、入力された文中の地名が全て未知語であった場合を想定している。全ての翻訳方法で、地名部分は全て「PLACE」と翻訳される。複数の地名が入力文に含まれる場合も考慮し、日本語中の地名と対応する番号を翻訳結果の地名に付与した。全ての翻訳結果に対して、この処理を行った。正解データの英語文の地名も、「PLACE」と汎化し、対訳となる日本語の番号を付与した。そして、翻訳の結果と正解データとの比較を行った。

4.3 地名対訳辞書の登録語数

日本語の学習コーパス中には「名詞-固有名詞-地域-一般」または「名詞-固有名詞-地域-国」と品詞を付与された単語が 1,098 存在した。それらの単語を含む対訳句はフレーズテーブルに 17,539 存在した。提案手法によって、98 の地名対訳を得ることが出来た。これにより、フレーズテーブルに登録された句の内、7,899 の対訳句を汎化することが出来た。形態素解析によって付与された品詞が正しいとすると、以下のような結果となる。

・学習コーパスに存在する地名の内、8.9 %が地名対訳辞書に登録された。

・フレーズテーブル内の地名を含む対訳句の内、45.0 %を汎化することが出来た。

4.4 翻訳結果

表 1 に各翻訳手法の BLEU による自動評価結果を示す。自動評価は closed と open の評価データでそれぞれ 1,000 文ずつ行った。

表 1: 翻訳結果の BLEU 値

翻訳方法	closed	open
通常翻訳	59.32	13.29
未知語翻訳	38.53	10.29
既存手法	41.80	13.07
提案手法	54.75	13.86

主観評価は open の評価データの中から無作為に抽出した 100 文を 1 人の評価者で評価を行った。

主観評価の基準を示す。

評価 A：入力文と同じ意味に理解できる文であり、語順も正しく、必要のない語を含まない文。

評価 B：入力文と同じ意味に理解できる文であるが、語順に違和感がある、または、必要のない単語を含む文。

評価 C：入力文と同じ意味に理解できない文である。または、全く意味を理解できない文。

主観評価の結果を表 2 に示す。

表 2: open データの主観評価結果

翻訳方法	評価 A	評価 B	評価 C
通常翻訳	7	29	64
未知語翻訳	2	9	89
既存手法	5	29	66
提案手法	2	33	65

5 考察

5.1 本手法の有効性について

表 1 及び、表 2 に示す、通常翻訳の翻訳精度と未知語翻訳の翻訳精度の比較から、文中の既知の地名が未知の地名となることによって、翻訳精度が大きく低下することがわかる。この原因としては、1 節で述べたように、語の並べ替えの問題が大きいと考えられる。

例 5:

原文：イングランド を バス で 旅行 する

汎化文：PLACE0 を バス で 旅行 する

未知語翻訳：PLACE0 make a bus tour of

提案手法：make a bus tour of PLACE0

例 5 のように未知語翻訳では地名である「PLACE」が未知語であるために、正しい翻訳が行えない文が確認できた。翻訳に必要な句が全て得られているにもかかわらず、未知の単語を正しい位置に配置できないという問題が生じることを確認した。これに対して、本手法では、未知の地名を地名が配置されるべき位置に配置することができていた。

しかし、closed の評価データの翻訳では、通常翻訳と比較して翻訳精度が低下してしまうという問題もある。この原因の 1 つとして、翻訳時に使用される句の長さが関係していると考える。句の長さと翻訳精度の関係について、5.2 節で考察を行う。

5.2 句の長さと翻訳精度

図 2 及び、図 3 に各翻訳手法で、評価データを翻訳した際に、最終的に出力された文に使用された日本語の句の平均の単語数と、BLEU 値の関係を示す。使用された全ての日本語の句 (以降、全ての句) 及び、地名を含む日本語の句 (以降、地名を含む句) の平均単語数と BLEU 値の関係を示す。表 3 に、各翻訳方法で使用された句の平均単語数の詳細を示す。表 4 に翻訳結果の文に正解の単語が含まれる割合を示す。

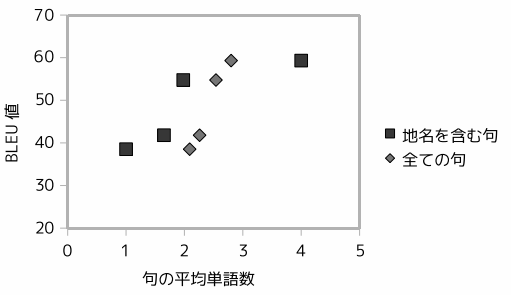

図 2: 使用される句の長さと翻訳精度の関係 (closed)

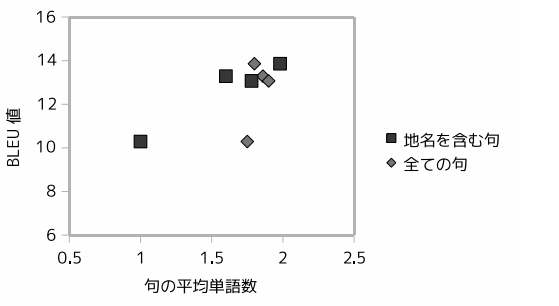

図 3: 使用される句の長さと翻訳精度の関係 (open)

表 3: 翻訳方法と句の平均単語数

	closed		open	
翻訳方法	全て	地名を含む	全て	地名を含む
通常翻訳	2.80	4.00	1.86	1.60
未知語翻訳	2.09	1.00	1.75	1.00
既存手法	2.26	1.65	1.90	1.78
提案手法	2.54	1.98	1.80	1.98

表 4: 正解の単語が含まれる割合

翻訳方法	closed	open
通常翻訳	74.9	54.1
未知語翻訳	65.3	53.2
既存手法	65.3	53.8
提案手法	71.9	53.8

この節では、提案手法によって、既存手法よりも長い句を用いた翻訳が可能となったかについてと、長い句を選ぶことによって翻訳精度が向上する傾向にあるということについて述べる。表 3 の結果から、open の全ての句の平均単語数の平均を除き、提案手法によって既存手法よりも長い句を用いた翻訳が可能となっていることがわかる。図 2 及び図 3 の結果から、使用される句の単語数の増加によって翻訳精度が向上する傾向にあることが確認できる。また、表 4 の結果から、長い句を用いることによって訳語の選択精度も向上していることを確認できる。このため、提案手法では、長い句を用いた翻訳が可能となったことによって、既存手法よりも翻訳精度が高くなったのであると考える。

5.3 地名の網羅性の評価

4.3 節の結果から、地名対訳辞書の網羅性がわかる。提案手法による地名対訳辞書の作成では、学習データ中の地名の 8.9 %の対訳が得られた。これにより、フレーズテーブル中の地名を含む対訳句の 45.0 %を汎化することが出来た。

本手法では、より正確な地名の対訳を得るために、日本語 1 単語-英語 1 単語対応の対訳のみを採用した。そのため、「アメリカ | 名詞-固有名詞-地域-国 | | | the|DT u.|JJ s.|NN」や「ニューヨーク | 名詞-固有名詞-地域-一般 | | | new|JJ york|VVP」などの 1 対多対訳の地名の汎化を行うことが出来なかった。アメリカやニューヨークなどの単語は学習コーパス中にも頻出する語であったが、汎化が行えなかった。そのために、翻訳知識の汎化を十分に行うことが出来なかった。

6 今後の課題

汎化を行う地名を統計情報と形態素情報、日英対訳辞書の知識を用いて獲得した。その結果、正確な地名の対訳を得ることが出来た。地名対訳辞書に登録された地名は、学習コーパスに含まれる地名の単語の 8.9 %であった。そして、フレーズテーブル内の地名を含む対訳句の内、45.0 %を汎化した。しかし、残りの 55.0 %の対訳句は翻訳に利用されることはなく、学習によって得られた翻訳知識を十分には利用できていない。本手法では、より正確な地名の対訳を得るために、汎化する地名の対象を日本語と英語の単語が、1 単語-1 単語対応している対訳句のみとしていた。しかし、1 対多、多対多の対訳の地名も存在する。これらの対訳の獲得も翻訳知識を十分に利用するためには必要である。

文中の地名が既知である場合には、通常翻訳と比べて提案手法の翻訳精度が低下してしまうという問題があることも判った。汎化手法の見直し、または、入力文が未知の語を含む場合と、既知の語を含む場合とで使用する翻訳モデルを変更するなどの、翻訳手法の改善が必要である。

7 おわりに

本手法では、地名に限定して翻訳知識の汎化を行った。日本語文中で、地名のカテゴリに属する単語は、英語文中でも地名のカテゴリに属するという考えのもとで地名の汎化を行った。そして、未知の地名を含む文を、そのままの状態で翻訳する未知語翻訳と、提案手法による翻訳の翻訳精度を比較した結果、open の評価データの翻訳で、BLEU 値が 3.57 ポイント向上した。主観評価の結果においても、翻訳精度の改善が得られた。これらの結果から、未知の地名を含む文の翻訳精度向上に本手法が有効であることを示した。

使用したツール及び言語資源

1) デコーダ Moses,
http://www.statmt.org/moses/
2) アライメント GIZA++,
http://www.fjoch.com/GIZA++.html
3) 言語モデル IRST LM,
http://sourceforge.net/projects/irstlm
4) 日本語形態素解析器 ChaSen, Ver.2.4.2,
奈良先端科学技術大学院大学 松本研究室,
http://chasen-legacy.sourceforge.jp/.
5) 英語形態素解析器 TreeTagger,
the University of Stuttgart,
http://www.ims.uni-stuttgart.de
/projekte/corplex/TreeTagger/.
6) 日英対訳コーパス CREST,
「セマンティックタイポロジーによる言語の等価変換と生成技術」プロジェクト.
7) 日英対訳辞書 英辞郎, Ver.54,

参考文献

1] Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical Phrase-Based Translation. Proceedings of HLT-NAACL 2003.

2] Papineni, K., Roukos, S., Ward, T., and Zhu., W.-J. BLEU: aMethod for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL),pp.311-318,2002.

3] 大熊英男, 山本博史, 隅田英一郎. フレーズベース SMT への対訳辞書の導入. 言語処理学会第 13 回年次大会,pp.380-383.

4] 荒牧英治, 黒橋禎夫, 柏岡秀紀, 加藤直人. 確率的用例ベース翻訳の実現. 言語処理学会第 11 回年次大会,pp.843-846.

5] 猪澤雅史, 村上仁一, 徳久雅人, 池原悟. 統計翻訳における単文・重文複文の翻訳精度の評価. 言語処理学会第 14 回年次大会, pp.869-872.

6] 今村賢治, 隅田英一郎, 松本裕治. 直訳性を利用した機械翻訳知識の自動構築. 自然言語処理, Vol.11 No.2, pp.85-100.