フレーズベース SMT への対訳辞書の導入

大熊 英男 1)2)　　山本 博史 1)　　隅田 英一郎 1)2)

1）情報通信研究機構 音声言語グループ

2）ATR 音声言語コミュニケーション研究所

〒619-0288　京都府相楽郡精華町光台 2-2-2

1. はじめに

統計的機械翻訳（SMT）は構文や意味などの深い言語知識を必要とせずに、大量の対訳コーパスを用い機械翻訳を行うもので、近年の計算機の大容量、高速化に伴い現実化され、発展してきたものである。統計的機械翻訳はさらに、単語単位の翻訳とその位置の組み合わせでおこなってきた単純なものから、翻訳の単位をフレーズに広げてフレーズ単位の翻訳とその位置の組み合わせで翻訳をおこなうフレーズベース統計翻訳システム[1]へと発展してきた。この場合のフレーズとは言語学的なフレーズではなく数単語からなる単語列のことである。このフレーズを翻訳単位にすることで、少なくともそのフレーズ内では訳語選択と語順が正しく維持され、単語単位の統計翻訳システム持つ語順のモデル化の困難性という弱点に対応することができる。

しかし、対訳コーパスのみで翻訳システムが構築できる半面、逆にその構築に使用した対訳コーパス（訓練コーパスと呼ぶ）外の資源をシステムに取り込むことには困難を伴うことが多い。そのうちのひとつに対訳辞書がある。対訳辞書には医療、バイオなど特定の分野のものや、地名、人名やユーザ辞書があり、これらの情報をすべて含むような対訳コーパスの収集は事実上不可能である。従って、これらの辞書の翻訳システムへの導入はフレーズベース統計翻訳システムを実用化するためには欠かせないものである。

本稿では、この訓練コーパス外対訳辞書のうち地名や人名の固有名詞に対してフレーズベース統計翻訳システムに効率よく導入する方法を提案する。

2. 統計翻訳における対訳辞書の利用

訓練コーパス外対訳辞書をフレーズベース統計翻訳システムに導入するもっとも簡単でローコストなものは、対訳辞書の対訳ペアに適当な確率値を割り振り、訓練コーパスから作られた翻訳モデルであるフレーズテーブルに追加する方法である。これを従来手法と呼ぶ。しかし、この方法では該当単語は訳出されるものの、その単語の位置や全体の語順がおかしくなることが多い。以下の例は訓練コーパスにはない単語である「カーディフ」（イギリス、ウエールズの地名）とその訳「cardiff」をフレーズテーブルに登録して翻訳を実行したものである。

カーディフ 行き の 往復 切符 一 枚 ください

↓

cardiff for a round trip ticket please

この例では「カーディフ」という単語は正しく訳出されるものの、その位置は文の先頭になってしまい、おかしな文となっている。これはフレーズベースの統計翻訳における語順の制約はフレーズに大きく依存するにもかかわらず、「カーディフ」がフレーズテーブル内の、追加したエントリ以外の他のいかなるフレーズに含まれていないためである。また、語順制約のための言語モデルにも「cardiff」が含まれていないため、その適切な位置が決められないからである。

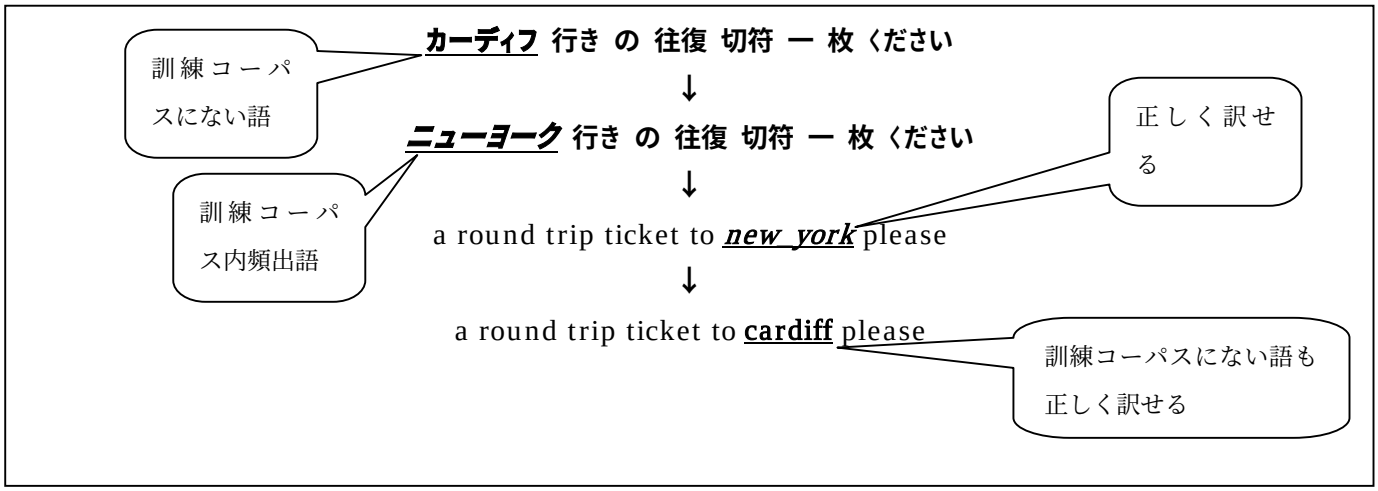

図 1 提案手法の翻訳過程

3. 対訳辞書とフレーズの融合

3.1. 基本アイディア

提案手法の翻訳過程を図１に示す。

提案手法は従来手法の問題を解決するために訓練コーパス内の頻出語を利用する。入力内にある訓練コーパス外対訳辞書の単語を訓練コーパス内の同カテゴリの頻出語と置き換え、デコーダがうまく訳せるようにして翻訳を進める。ここで頻出単語を利用するのは、同じ文脈でよりよく訓練されている、つまりフレーズテーブル内のたくさんのエントリに含まれていて、十分な統計量が取られている可能性が高いからである。図１の例では、訓練コーパスにない語「カーディフ」を訓練コーパス内の同カテゴリの頻出語「ニューヨーク」に置き換え翻訳し、翻訳結果内から「ニューヨーク」の訳語である「new_york」を探して、それを「カーディフ」の訳語である「cardiff」に置き換えている。

3.2. 実装方法

実現にはまず、訓練コーパス外辞書を場所､名前､組織のようなカテゴリに分けておく。次に訓練コーパスからそのカテゴリごとに頻出単語をいくつか取得し、そのおのおのに対訳を付与した表を作成しておく。これを置換単語表と呼ぶ。翻訳システムへの入力内の単語を置き換えるアルゴリズム

```
ReplaceInput(input)
foreach cat in categories
  i := 0
  foreach word in input
   if searchword(word, dic[cat]) then
     replaceword.pair := GetReplaceWord(cat,i,input)
     word := replaceword.pair.src
     put (word, replaceword.pair) to replaced.table
    end
  end
end
return (input, replaced.table)
```

図 2 アルゴリズム

を図２に示す。GetReplaceWord はカテゴリごとの頻出単語テーブルから i 番目の置換単語を取り出す関数である。置換単語が入力文字列内にないことを確かめる必要がある。入力文字列内に置換単語候補があった場合には、その単語を飛ばして、次の候補を見に行き､入力文字列内に候補がなくなるまでテーブル内で i を進める。

また、replace.table は置換結果を保存しておくテーブルである。このテーブルを利用して翻訳後の結果から置換した単語の再置き換えをおこ

	日本語	英語
訓練セット	680701	680701
テストセット	10150	10150

表 1　コーパスサイズ（文数）

なう。

4.　実験

4.1.　実験条件

実験には ATR の旅行会話基本表現集を用いた。このコーパスは旅行会話でよく使われる表現を集めたものである。コーパスのサイズを表１に示す。

日本語コーパスは茶筅で、英語コーパスは ATR 内で使用しているタガーでそれぞれ形態素解析をおこない、翻訳システムのモデルを作成した。実験はもともとのテストセットから次の２種類のテストセットを作成し、おこなった。テストセットから茶筅のタギングによる地域-一般カテゴリを持つ語を３つまで含む文を取り出し、それらの語を順に訓練コーパス外単語である「カーディフ」「ボースマス」「ベルファスト」に置き換えたもの。テストセットから人名-名、人名-姓カテゴリを持つ語をそれぞれ３つまで含む文を取り出し、それらの語をそれぞれ順に訓練コーパス外単語である「ナオミチ」「ミツアキ」「ノブキチ」、「コイズミ」「オザワ」「ナカソネ」に置き換えたもの。これらの新たに作られたテストセットをそれぞれ、場所テストセット、人名テストセットと呼ぶ。これらの日本語テストセットに対し、同様に置き換えをおこなった英語テストセットを用意し、日英翻訳実験の自動評価の参照訳とした。また、同じテストセットを逆方向に使用して英日翻訳実験もおこなった。

評価は従来手法として置き換えに使用した語とその対訳をフレーズテーブルに登録し翻訳を実行したものと、今回の提案手法である頻出語での置き換えをおこなって翻訳を実行した場合の比較でおこなった。フレーズテーブルは本稿で扱う Pharaoh[2]互換のデコーダの場合、対応する入力言語のフレーズと出力言語のフレーズとそれらの出現確率等を並べたテーブルであり、このテーブルに同じフォーマットで追加するだけである。また置換単語表にはそれぞれのカテゴリで頻度の高い順に３単語ずつ用意した。

		場所	名前
	文数	106	60
日英	従来手法	39.87	36.87
	提案手法	**44.14**	**40.18**
英日	従来手法	**42.09**	21.87
	提案手法	41.91	**25.53**

表 2　自動評価結果

なおデコードには ATR で作成した Pharaoh 互換の Cleopatra を使用した。

4.2.　実験結果

自動評価結果を表２に示す。値は BLEU のスコアである。表が示すとおり、英日の場所テストセットを除き提案手法が従来手法よりよい値となった。ただし、この BLEU 値は参照訳が一文であり、複数の分を使う通常の評価に比べて、信頼性が高下がるので、より詳しく調べるためさらに、翻訳結果に差があった文のみ主観評価をおこなった。その結果を表３に示す。評価は A:完全訳、B:部分訳、C:理解可能訳、D:理解不可能訳として A、A+B、A+B+C の割合を比較した。この主観評価では提案手法の優位性がはっきりと示された。この主観評価の中から提案手法が従来手法よりよい結果となった具体例を図３に、逆に従来手法が提案手法よりよりよい結果となった具体例を図４に示す。括弧内は実際の評価値である。

なお場所テストセットの英日翻訳の自動評価のみで提案手法が従来手法に劣るあるいは従来手法が好成績となった原因として考えられるのは、従

		場所			名前		
日英	文数	51			39		
		A	A+B	A+B+C	A	A+B	A+B+C
	従来手法	47.06%	72.55%	92.16%	38.46%	61.54%	82.05%
	提案手法	**70.59%**	**84.31%**	**94.12%**	**64.10%**	**82.05%**	**89.74%**
英日	文数	57			34		
		A	A+B	A+B+C	A	A+B	A+B+C
	従来手法	33.33%	75.44%	80.70%	50.00%	67.65%	91.18%
	提案手法	**73.68%**	**80.70%**	**89.47%**	**58.82%**	**76.47%**	91.18%

表 3 主観評価結果

来手法では該当単語が文頭に訳出されることが多く、また、参照訳も文頭に該当単語が多く現れているからと考えられる。

5. まとめと今後の予定

本稿では、フレーズベース統計翻訳システムに対して訓練コーパス外の対訳辞書を適切に追加する方法を提案し、その有効性を確認した。

今回の発表は固有名詞の特定のカテゴリに限ったものであるが、名詞、動詞、形容詞などのより一般的な品詞を持つ語に対しても、そのカテゴリ分けや置換テーブルの作成などを自動化し、導入できるようにすることがフレーズベース SMT システムの実用化のためには不可欠である。

[1] P. Koehn, F. Joseph Och, and D. Marcu.
Statistical Phrase-Based Translation.
In *Proc. of HLT*, 2003

[2] P. Koehn
Pharaoh: a Beam Search Decoder for
Phrase-Based SMT.
In *Proc. of the 6th AMTA*, pages 115–124, 2004

入力： 私 は 今晩 カーディフ を 発ち ます
従来手法(B)： i cardiff leaving tonight
提案手法(A)： i'm leaving cardiff tonight

入力： カーディフ は ファッション の 中心地 です
従来手法(C)： the center of cardiff fashion
提案手法(A)： cardiff is the center for fashion

入力： カーディフ 行き の 最終 は 何時 です か
従来手法(B)： what time is the last cardiff go
提案手法(A)： what time is the last train to cardiff

図 3 主観評価例（成功）

入力： カーディフ 行き の 列車 は 何 番 ホーム です か
従来手法(A)： what platform does the train for cardiff
提案手法(D)： what number is home the train for cardiff

図 4 主観評価例（失敗）