

旅行会話文対訳コーパスからの 音声翻訳用言語モデルの自動抽出*

森	元		逞**
折	本	真	一***
福	永	良	一****
高	橋	伸	弥**

Automatic Extraction of Language Models for Speech Translation from a Travel Conversation Bilingual Corpus

Tsuyoshi MORIMOTO, Shin-ichi ORIMOTO,
Ryouichi FUKUNAGA and Shin-ya TAKAHASHI

We are developing a speech translation system from Japanese to English. In general, three kinds of language information (language model) are necessary for a speech translation; a source language model for speech recognition, translation rules from a source language to a target language and a target language model for sentence generation. It takes very much time and effort of developers to prepare such various kinds of language model. In this paper, a method of automatic extraction of such language models from a bilingual corpus is proposed. By comparing sentences, a similar part and different parts are extracted, and then translation rules and language models are automatically constructed. Some experiment results are also reported, which shows the effectiveness of the method.

Key Words: Speech Translation, Automatic Construction of Language Model, Learning of Translation Rules, Rule Extraction from a Bilingual Corpus

1. はじめに

旅行会話文を対象とした日英音声翻訳システムの開発を行なっている⁽¹⁾。旅行会話文はかなり定型的な文が多いことに着目し、翻訳規則を句単位に用意し、入力文とのパターンマッチング処理により効率的な翻訳を行う方式を実現している。一般に音声翻訳システムでは、原言

語で入力された音声を認識するための音声認識用言語モデル、原言語から目的言語への翻訳規則、および目的言語文生成のための言語モデルの3種類の言語情報が必要である(図1参照)。ここで、我々のシステムでは原言語は日本語、目的言語は英語である。(なお、以下では簡単のためこれらの3種の言語情報をあわせて「言語モデル」と呼ぶ)。これまでのシステムでは、これらの言語モデルをすべて人手で作成した。このシステムはまだ小規模(翻訳可能な文パターンは約300)であるが、それでもこれらの言語モデルの作成にはかなりの労力を要した。

本論文ではこの問題を解決するため、日本語と、対応

* 平成16年11月30日受付

** 電子情報工学科

*** 電子工学専攻

**** 電子工学専攻(現在、(株)東芝デジタルメディアエンジニアリング)

する英語の翻訳文が与えられたコーパス (対訳コーパス) を学習データとし、そこから上記3種類の言語モデルを自動的に抽出する方法を提案する。処理の中心となるのは対訳コーパスからの翻訳規則の抽出である。一旦翻訳規則が抽出されれば、音声認識用の日本語言語モデル、英語生成用の英語言語モデルの作成は比較的容易である。すなわち、各翻訳規則に現れる日本語句、英語句同士を、相互に接続すればネットワーク型の言語モデルが生成できる。

翻訳規則の自動抽出の問題については、これまでも多くの研究が行なわれている^{(2)~(5)}。Kajiら⁽²⁾は、日本語、英語の対訳文が与えられた場合、それらを各々構文解析し、両者の構文構造を比較することによって規則を抽出する方法を提案している。しかしこの方法では文パターンの種類が多くなった場合、それらを全てカバーする構文解析規則を用意するのが大変である。内山ら⁽³⁾は、まず日本語文と英語文に現われる単語のうち、和英辞書に記述されているものを単語の翻訳規則として抽出し、また元の日本語文、英語文においてこれらの単語を変数に置き換えたものを文パターンの翻訳規則として抽出する方法を提案している。しかしこの方法では、変数にどのような単語でも当てはめることができるため、構文・意味的に適切でない単語が選ばれ、その結果誤訳を生じる恐れがある。さらに本方法では文の翻訳規則しか抽出できず、文より小さな単位である句レベルの翻訳規則は抽出できない。またGuevenirら⁽⁴⁾は対訳コーパスの2文を比較し、一致した部分と相違する部分を求めてそれぞれを句レベルの翻訳規則とする手法を提案している。しかし対訳コーパスの規模が大きくなった時、具体的にどのような手順で比較すればよいかは述べられていない。なお、北村ら⁽⁵⁾は、多量の対訳コーパスを用いて単語の翻訳規則を抽出する方法を提案しているが、句レベルの翻訳規則は対象としていない。

我々はこれらの問題を解決するため、次のような方法を実現することとした。まず、文献(3)と同様な方法で、単語の翻訳規則ならびに一時的な文の翻訳規則を抽出する。ただし、元の日本語文、英語文においてこれらの単語を統語・意味制約付きの変数に置き換えておく。そのためにシソーラスを利用する。これにより、得られた文の翻訳規則における変数には統語・意味の制約が付与されることになる。次に、得られた日本語、英語それぞれの文の翻訳規則同士を比較することにより句の翻訳規則を抽出する。一般に日本語会話文は文末に類似した表現が使用されやすく、またそれらに対応する英語表現は文頭に現れる傾向にある。このため、日本語の比較は文末から、英語は文頭から比較し、日英それぞれに一致する表現部分および異なる表現部分を求め、それらの日英の

部分同士を句の翻訳規則とする。以上のように句の翻訳規則が抽出されたら、日本語言語モデル、英語言語モデルの作成は前述したように行う。

本論文では、まず2章で我々の音声翻訳システムの概要およびそこで用いられる言語モデルについて説明する。次に、我々が実現した翻訳規則の抽出、言語モデルの作成の具体的な方法について述べる。また実際に言語モデルの抽出実験を行ったので、その結果を報告する。

2. 音声翻訳システムの概要

図1に全体のシステム構成をしめす。まず入力された日本語音声の音声認識が行われる。音声認識器としてはHVite⁽⁶⁾を使用し、またHMMはJulius⁽⁷⁾のHMMを使用している。言語モデルはネットワーク文法として定義する。

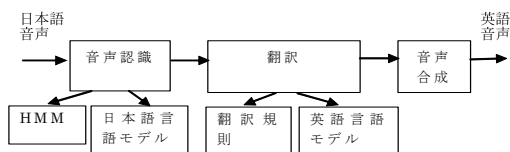


図1 音声翻訳システムの構成

翻訳規則として、単語翻訳規則と句翻訳規則を用意する。それぞれ以下のように、単語単位および句単位に日本語と対訳の英語を定義したものである。

・単語翻訳規則

駅：station：#place

郵便局：post-office：#place

ワイン：wine：#drink

・句翻訳規則

どこ-です-か：where-is：#where_is

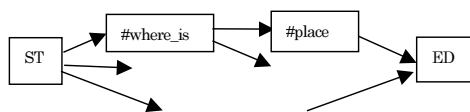
欲しい-のですが：I-would-like-to-have：

#would_like_to_have

%1(#place_name)-から-%2(#place_name)へ-行く：go-from-%1-for-%2：#go

ここで、各規則の最後のカラム（#の付いたもの）は各規則（単語翻訳規則の場合は対訳の日本語、英語単語の対）に付与した名称であり、これをクラス名と呼ぶ。このクラス名は次に述べる変数の定義および英語言語モデルで使用される。まず変数であるが、これは規則内の“%数字”で定義されているものであり、同じ番号の変数が日本語と英語で対応していることを表している。また日本語内の変数はその後ろにカッコでその変数として代入できる単語のクラス名が指定されている。よって単語のクラス名としてシソーラスなどから得られる意味コードを付与することにより、統語的・意味的な制約を与えることができる。なお句翻訳規則のクラス名は、以下に

述べるように英語言語モデルと対応をとるためのものであり、命名法には特に制約はない。最後に英語言語モデルであるが、これは図2のようにクラスをノードとし、それらをリンクにより接続したネットワーク文法として定義されている。すなわち、クラスの接続方法を定義したもとなっている。



(注) ST: 開始ノード ED: 終了ノード

図2 英語言語モデル

処理は以下のように行う。まず音声認識により日本語単語列が得られると、すべての単語および単語列について翻訳規則と一致するかどうかを調べる。一致すれば、対応する英語単語ないしは句をとりあえず部分翻訳とする。同時に対応する英語言語モデルのクラス・ノードを活性化する。そして英語言語モデルにおいて開始ノードから活性化されたノードを順次たどり、最終ノードまで到達するノード列を得る。最後にその順番に部分翻訳を並べ替えて英文を生成する。もし開始ノードから終了ノードまで完全につながらない場合は、最大長の部分ノード列を得て、それに対応する部分英文を翻訳結果とする。

3. 翻訳規則の抽出

我々が対象とした旅行会話文の一部を図3に示す。

スカートを探しています
 お手洗いはどこですか？
 市役所にはどうやっていったらいいですか？
 クリーム入りコーヒーを1杯もらえますか？
 今夜中にスーツにプレスしてもらえますか？
 医者を呼んでいただけませんか？
 ここで写真を撮るとってもいいですか？

図3 会話文の例

これらは市販されている4つの旅行会話例文集から、一般的に用いられると思われるもの約320文を選んだものであり、文献(1)のシステムで対象とした例文と同じものである。なお図3では省略しているが、各文には対応する英語の文がある。これらの対訳文を学習データとして規則の抽出を行う。

図4に規則抽出の全体の流れをしめす。まず学習データ内の日本語部分を形態素に分割し、次に各日本語形態

素で和英辞書の辞書引きを行う。このとき活用のある形態素(動詞、形容詞など)は標準形(終止形)を用いる。

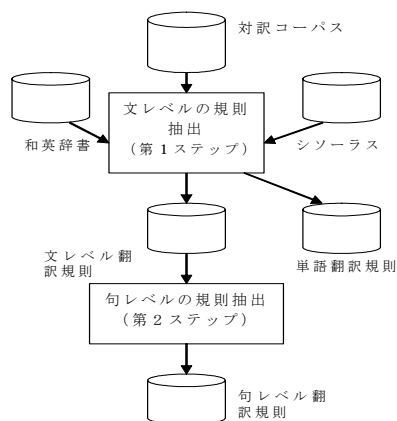


図4 処理の流れ

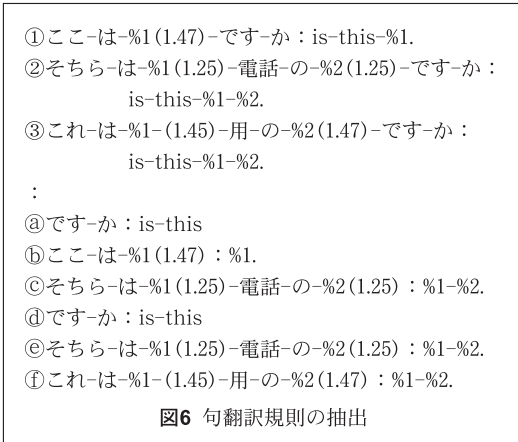
辞書引きしてエントリが見つかったものは辞書から対応する英単語を取り出し、同じ英単語が対訳英文にあるかどうかを調べる。もし一致するものがあれば、それを単語翻訳規則として取りだす。これにより、辞書には複数の訳語があっても、旅行会話というドメインに特化した単語翻訳規則を抽出することができる。一方、元の日本語文、および英語文で該当する単語は変数に置き換え、これを文翻訳規則とする。上記の処理において、和英単語辞書としては一般に公開されている EDICT⁽⁶⁾を用いた。また前述したように単語翻訳規則にはシソーラスから求めた分類コードをクラス名として付与するが、シソーラスとしては国立国語研究所で作成された分類語彙表⁽⁷⁾を用いた。なお、同表では分類コードを図5のように付与している。そこで、今回は上位3桁(X_1, Y_1, Y_2)を用いることにした。

X_1, Y_1, Y_2, Y_3, Y_4
 X_1 : 品詞 (ただし、4分類)
 Y_1 : 意味分類
 Y_2, Y_3, Y_4 : 意味細分類

図5 分類語彙表の分類コード

第2ステップでは、まず第1ステップで得られた文の翻訳規則全体を、日本語文の文末からの逆順でソートする。次に隣接する2つの文規則について、日本語は文末から逆方向に一致する部分を求め、また同時に英文は文頭から順方向に一致する部分を求める。そしてまず日本語の一致部分と英語の一致部分を1つの句翻訳規則とす

る. また日本語, 英語それぞれで一致しなかった部分が2つ残るが, これら各々を句の翻訳規則とする. 例を図6にしめす.



まずステップ1により①, ②…のような文翻訳規則が得られているとする. そこで, ①と②から, 一致した部分として㉔のような規則が, また残りの部分から, ㉕, ㉖のような規則が得られる. 同様に②と③から, ㉗, ㉘, ㉙の3つの規則が得られる.

4. 日本語, 英語言語モデルの作成

前で述べたように日本語言語モデル (音声認識用), 英語言語モデル (英文生成用) はいずれもネットワークとして定義される. ここではその作成法について述べる. 入力の前段の処理で得られた句翻訳規則である. 図7に説明をしめしているが, 例は図6と同じものである. まず①と②の比較により, 一致した部分㉔と, 一致しなかった部分㉕, ㉖が得られる. そこで, ㉔, ㉕, ㉖それぞれをノードとし, 日本語, 英語について図にしめすように設定する. 次に, ②と③の比較で得られた㉗, ㉘, ㉙についても同様にネットワークに追加する. ただし, ㉔と㉗, ㉔と㉘は同じものであるためノードの追加は行わず, 必要なリンクだけを設定する. 以下同様な処理を繰り返して, ネットワークを作成する.

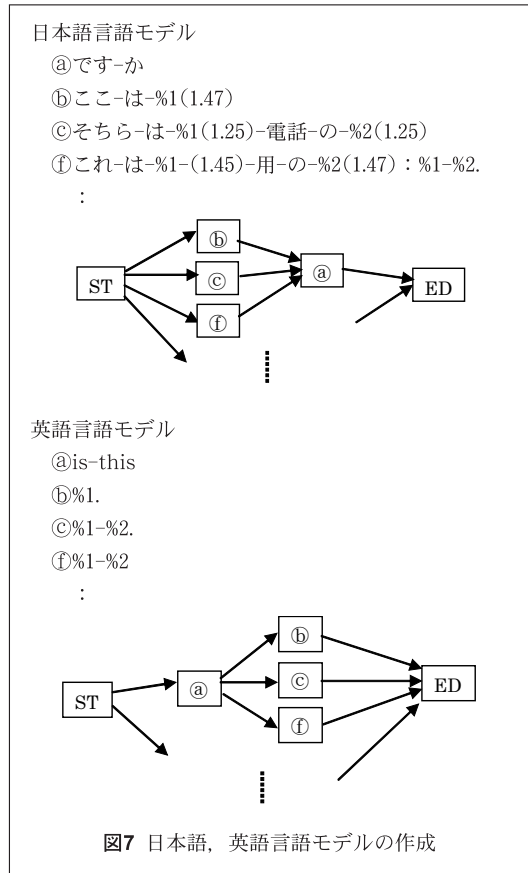
5. 実験結果と評価

(1) 対訳コーパス

日英対訳コーパスから規則の抽出実験を行なった. 対訳コーパスに含まれている文数および語彙数を表1にしめす.

表1 対話コーパス

語彙数	日本語: 670語 英語: 617語
文数	321文



(2) 言語情報の抽出と翻訳能力

抽出, 生成された言語モデルの各種の値を表2にしめす. また比較のために, 文献(1)で使用している人手で作成した言語モデル (以下, これをベース・モデルと呼ぶ) の値をカッコ内にしめしている.

表2 抽出された言語モデル

日本語言語モデル	ノード数	4450 (537)
	リンク数	6434 (2271)
	平均分岐数	2.1 (4.6)
翻訳規則	単語翻訳規則	139 (241)
	句翻訳規則	535 (306)
英語生成言語モデル	ノード数	587 (170)
	リンク数	981 (294)

日本語, 英語言語モデルのノード数は大幅に増加し, またこれに伴ってリンク数も増大している. ただし音声認識率に大きく関係する日本語言語モデルの平均分岐数は半分以下に減少している.

翻訳可能な日本語, すなわち受理可能な日本語の文数

は以下のように推定できる。まず日本語言語モデルを用いて実際に文をランダムに1,000文生成し、それらが正しい日本語文であるかどうかを目視でチェックした。その結果、正解の日本語文は64.0%であった。また生成された日本語文の1文あたりの平均単語数は9.6語であった。これらの数値と、言語モデルの平均分岐数が2.1であることから、

$$2.1^{9.6} \times 0.65 = 793$$

文を受理することができると言える。すなわち、学習データ321文に対し、

$$793 / 321 = 2.47 \text{ 倍}$$

の文を受理し翻訳することができるようになったことになる。

(3) 各コンポーネントの性能評価実験

自動抽出された言語モデルを用いて、まず音声認識部、翻訳部それぞれを単独で使用した場合の性能評価実験を行った。対象とした文(または発話)数は156文であり、また音声認識実験では発話者3名がそれぞれ異なる文を発話したデータを用いた。結果を表3に示す。表に示したように、音声認識、翻訳いずれもベース・モデルに対し10%以上正解率が向上している。

表3 各コンポーネントの性能評価

音声認識	単語認識率	99.1% (97.6%)
	文認識率	97.4% (85.9%)
翻訳	文翻訳率	99.5% (89.7%)

(カッコ内はベース・モデル)

(4) システム全体の性能評価

上記の音声認識の実験で得られた認識結果を翻訳部に加え、音声翻訳システム全体としての翻訳正解率を求めた。結果を表4に示す。コンポーネントの性能評価と同様に、ベース・モデルに対し大幅に性能が向上していることがわかる。

表4 システム全体の性能評価

		翻訳		
		正	誤	計
音声認識	正	96.9% (75.6%)	0.5% (10.3%)	97.4% (85.9%)
	誤	0% (6.4%)	2.6% (7.7%)	2.6% (14.1%)
	計	96.9% (82.0%)	3.1% (18.0%)	100% (100.0%)

(カッコ内はベース・モデル)

6. まとめ

旅行会話文を対象に、対訳コーパスから音声翻訳に必要な音声認識用言語モデル、翻訳規則、英語生成用言語モデルの3種の言語モデルを自動的に抽出する方法を提案した。またその性能評価実験を行い、人手で作成したモデルに比べ、音声認識率、翻訳率とも大幅に向上することを確認した。

本手法では対話コーパスの日本語部分を文末から比較し、また同時に英語部分を文頭から比較し、一致した部分および異なる部分をそれぞれ句翻訳規則として抽出している。またそのような句を単位として日本語言語モデル、英語言語モデルを構築している。これは、おおまかには日本語および英語をそれぞれ前後に2つの句に分割する方法といえる。従って、これよりさらに小さな単位で句を抽出することはできない。今後は、さらに小さな句単位を抽出する方法について検討する予定である。

参考文献

- 森元, 高橋: 定型文音声翻訳システム, 福岡大学工学集報, No.69 (2002)
- H. Kaji, Y. Kida, Y. Morimoto: "Learning Translation Templates from Bilingual Text", Proc. of COLING-92, pp.672-678 (1992)
- 内山, 荒木, 宮永, 柄内: "機能学習による機械翻訳手法の評価実験," 情処研資, NL93-4, pp.23-30 (1993)
- H.A. Guevenir, I. Cicekli: Learning Translation Templates from Examples, Information Systems, Vol.23, No.6 (1998)
- 北村, 松本: "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)
- http://www.csse.monash.edu.au/~jwb/j_edict.html
- 国立国語研究所: 「分類語彙表」形式による語彙分類表 (増補版) (1996) HTK
- S. Young et al.: The HTK Book (for Ver. 3.0), 1999 [<http://htk.eng.cam.ac.uk/>]
- http://www.csse.monash.edu.au/~jwb/j_edict.html