

人称代名詞において1対多の対訳句を含む文の翻訳

森島葵 村上仁一

鳥取大学大学院 工学研究科 情報エレクトロニクス専攻

{s132056,murakami}@ike.tottori-u.ac.jp

1 はじめに

パターンに基づく統計翻訳 (PBSMT)[1] は対訳句と文パターンを用いて翻訳を行う手法である。対訳句は対訳学習文から抽出し、文パターンは対訳学習文から作成する。作成された文パターンと対訳句を用いて出力候補文を複数作成し、翻訳確率が最も高い出力候補文を選択して翻訳結果として出力する。

しかし現在のPBSMTでは、対訳学習文中に含まれる日本語句と英語句を1対1に変数化して文パターンを作成している。従って、入力文と参照文を比較したとき、日本語と英語が1対多に対応する対訳句を含む文において誤った翻訳結果を出力する[4]。

本論文では、PBSMTにおいて、人称代名詞に限定した1対多の対訳句を含む文から正しい出力候補文を生成する手法を提案する。

2 PBSMTの概要

PBSMTは、対訳句と文パターンを用いて翻訳を行う機械翻訳手法の一種である。PBSMTでは5つのステップを用いて翻訳を行う。PBSMTの流れを図1に示す。

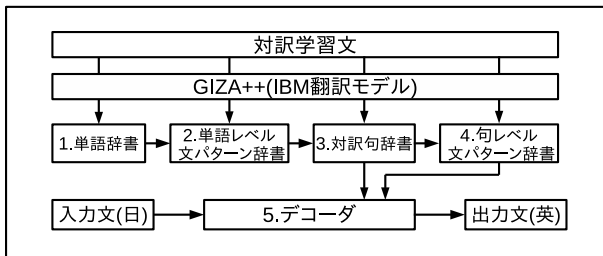


図1 PBSMTにおける従来手法の全体の流れ

2.1 単語辞書

PBSMTでは、まず対訳学習文から原単語と目的単語の単語アライメントをGIZA++[2]を用いて取得し、対訳単語の原単語と目的単語が双方向で対応する確率を示す単語確率を単語アライメントから得る。そして対訳単語に単語確率を付与した単語辞書を作成する。

2.2 単語レベル文パターン辞書

単語レベル文パターンは、対訳学習文中の対訳単語を単語辞書と照合し、照合に成功した場合は対訳単語を変数化して作られる。句レベル文パターンを多く作成するために、単語レベル文パターンは可能な限り複数作成される。

2.3 対訳句辞書

対訳句辞書は、対訳学習文から単語レベル文パターンを用いて対訳句を抽出し、抽出した対訳句に対訳句確率を付与して作成される。対訳句確率とは、対訳句の原言語句と目的言語句が双方向で対応する確率であり、対訳句に含まれる原単語と目的単語の単語確率から得る。

2.4 1対1の句レベル文パターン辞書

句レベル文パターンは、対訳学習文と対訳句を照合し、照合に成功した場合は日本語句と英語句を1対1に変数化して句レベル文パターンを可能な限り多く作成する。作成された句レベル文パターンは文パターン確率が付与される。

2.5 デコーダ

デコーダでは、作成された対訳句辞書と句レベル文パターン辞書を用いて、与えられた入力文から翻訳結果を出力する。PBSMTにおける日英翻訳の手順を以下に示す。

手順1 日文パターンの選択

入力文の字面と一致度が高い日文パターンを優先的に選択

手順2 日本語句の取得

日文パターンの変数部に対応する日本語句を取得

手順3 英文パターンの取得

日文パターンに対応する英語文パターンと文パターン確率を取得

手順4 英語句の取得

英文パターンの変数部に対応する英語句を取得

手順5 出力候補文の作成

英文パターンの変数部に英語句を代入して出力候補文を作成

手順6 翻訳確率の計算

言語生成確率 (*tri-gram*) と対訳句確率と文パターン確率の総積から翻訳確率を計算

手順7 出力候補文の選択

翻訳確率が最も高い出力候補文を選択して翻訳結果として出力

PBSMTにおける日英翻訳の流れを以下の図2に示す。なお、図2中の対訳句確率と文パターン確率と言語生成確率、翻訳確率はそれぞれ対数値であり、文パターン中の X_0 と X_1 は変数部である。

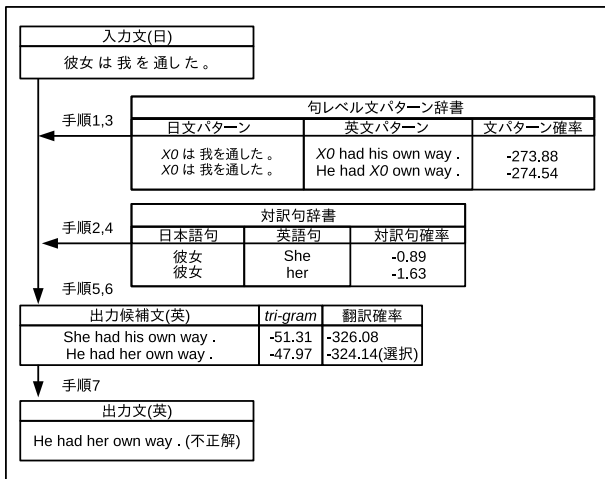


図2 PBSMTにおける日英翻訳の流れ

3 1 対多の対訳句を含む文の翻訳

3.1 1 対多の対訳句を含む文の概要

1 対多の対訳句を含む文とは、入力文と参照文を比較したときの日本語句と英語句が 1 対多に対応する対訳句を含む文である。1 対多の対訳句を含む文の例を表 4 に示す。表 4 の例 1 は、日本語句“彼女”と英語句“She”, 日本語句“彼女”と英語句“her”が対応する対訳句を含んでいる。

3.2 1 対多の対訳句を含む文の問題

現在の PBSMT では、対訳学習文から対訳句を変数化して文パターンを作成する。そして、変数化する対訳句の対応は 1 対 1 に限定している。従って、変数化する対訳句の対応が 1 対多である場合は複数の文パターンが作成される。また、1 対多の対訳句を含む文の翻訳では、どの文パターンを用いても誤った翻訳結果を出力する。従来手法における 1 対多の対訳句を含む文の翻訳例を以下の表 1 に示す。

表 1 では、入力文と参照文を比較したとき、日本語句“彼女”と英語句“She”, 日本語句“彼女”と英語句“her”が対応している。対訳学習文は日本語句“彼”と英語句“He”, 日本語句“彼”と英語句“his”が対応している。文パターン 1 は変数化する対訳句 1 を、文パターン 2 は変数化する対訳句 2 を変数化して作成する。文パターン 1 と代入する対訳句 1 を用いて出力候補文 1 を作成し、文パターン 2 と代入する対訳句 2 を用いて出力候補文 2 を作成する。しかし、いずれの出力候補文も入力文に対して誤った翻訳となる。

4 提案手法

4.1 1 対多の句レベル文パターン辞書

従来の PBSMT は、文パターンの作成時に変数化する対訳句が 1 対 1 に限定している。従って、誤った出力候補文しか出力できない問題がある。そこで、1 対多に変数化して文パターンを作成する。提案手法を用いた 1 対多の対訳句を含む文の翻訳例を以下の表 2 に示す。

表 2 では、表 1 の入力文と参照文、対訳学習文、変数

表 1 1 対多の対訳句を含む文の翻訳例 (従来手法)

入力文	彼女が我を通した。	
参照文	She had her own way .	
対訳学習文	日	<u>彼</u> が我を通した。
	英	<u>He</u> had <u>his</u> own way .
変数化する対訳句 1	日	<u>彼</u>
	英	<u>He</u>
文パターン 1	日	X0 が我を通した。
	英	X0 had <u>his</u> own way .
代入する対訳句 1	日	<u>彼女</u>
	英	<u>She</u>
出力候補文 1	She had <u>his</u> own way .	
変数化する対訳句 2	日	<u>彼</u>
	英	<u>his</u>
文パターン 2	日	X0 が我を通した。
	英	<u>He</u> had X0 own way .
代入する対訳句 2	日	<u>彼女</u>
	英	<u>her</u>
出力候補文 2	<u>He</u> had her own way .	

表 2 1 対多の対訳句を含む文の翻訳例 (提案手法)

文パターン 3	日	X0 が我を通した。
	英	X0 had X0' own way .
出力候補文 3	She had her own way .	

化する対訳句 1 と変数化する対訳句 2 と代入する対訳句 1 と代入する対訳句 2 を用いる。対訳学習文と変数化する対訳句 1 と変数化する対訳句 2 を用いて文パターン 3 を作成する。文パターン 3 と代入する対訳句 1 と代入する対訳句 2 を用いて出力候補文 3 を作成する。出力候補文 3 は入力文に対して正しい翻訳となる。

4.2 対訳句確率と文パターン確率の計算方法

対訳句確率と文パターン確率の計算式 1 を以下に示す。

$$P\left(\frac{J_0 \dots J_{N-1}}{E_0 \dots E_{M-1}}\right) = \prod_{m=0}^{M-1} \prod_{n=0}^{N-1} P(E_m | J_n) \prod_{n=0}^{N-1} \prod_{m=0}^{M-1} P(J_n | E_m) \quad (1)$$

ここで、 J_n は対訳句中の日本語単語、 E_m は対訳句中の英語単語、 N は日本語単語数、 M は英語単語数、 $P(J_n | E_m)$ は英語単語 E_m が日本語単語 J_n に翻訳される確率 (GIZA++ の値) を表す。

5 実験

5.1 実験目的

本実験では、1 対多の対訳句を含む文の翻訳において、1 対多に変数化して作成した文パターンを用いた出力候補文の翻訳確率が、1 対 1 に変数化して作成した文パターンを用いた出力候補文の翻訳確率よりも高くなるか

を調査する。

5.2 実験方法

本研究では、以下の手順に従って実験を行う。

- 手順 1 対訳学習文に含まれる 1 対多の対訳句を 1 対 1 に変数化して文パターン 1 と文パターン 2 を作成
- 手順 2 文パターン 1 と代入する対訳句 1 を用いて出力候補文 1 を作成
- 手順 3 文パターン 2 と代入する対訳句 2 を用いて出力候補文 2 を作成
- 手順 4 対訳学習文に含まれる 1 対多の対訳句を 1 対多に変数化して文パターン 3 を作成
- 手順 5 文パターン 3 と代入する対訳句 1 と対訳句 2 を用いて出力候補文 3 を作成
- 手順 6 対訳学習文の人称代名詞を置換して入力文を作成
- 手順 7 出力候補文の翻訳確率を式 1 を用いて計算

5.3 実験データ

入力文は、電子辞書などの例文から 1 対多の対訳句を含む文を抽出した単文コーパス [3] を用いる。対訳学習文は、日本語句“彼”と英語句“He”，日本語句“彼”と英語句“his”の対訳句を含む文を用いる。入力文は、対訳学習文中の日本語句“彼”と英語句“He”，日本語句“彼”と英語句“his”の対訳句を日本語句“彼女”と英語句“She”，日本語句“彼女”と英語句“her”の対訳句に置換した文を用いる。使用するデータの内訳を表 3 に示す。また、入力文の例を以下の表 4 に示す。

表 3 実験データ

対訳学習文	160,000 文
入力文	100 文

表 4 入力文の例

入力文 1(日)	彼女 は我を通した。
参照文 1(英)	She had her own way .
入力文 2(日)	彼女 は派手に演じすぎた。
参照文 2(英)	She overdid her part .
入力文 3(日)	彼女 は手を開いた。
参照文 3(英)	She opened her hand .

5.4 評価方法

本実験では、対訳学習文を 1 対 1 に変数化した文パターン 1 を用いて出力候補文 1 を作成し、出力候補文 1 の翻訳確率 1 を求める。文パターン 2 を用いて出力候補文 2 を作成し、出力候補文 2 の翻訳確率 2 を求める。1 対多に変数化した文パターン 3 を用いて出力候補文 3 を作成し、出力候補文 3 の翻訳確率 3 を求める。翻訳確率 3 が翻訳確率 1 と翻訳確率 2 よりも高く、出力候補文 3 が翻訳結果として出力する文の数を調査する。

5.5 実験結果

翻訳結果の例を以下の表 5 に示す。表 5 の出力候補文 3 の翻訳確率が出力候補文 1 と出力候補文 2 よりも高い。従って、出力候補文 3 が翻訳結果として出力される。また、節 5.4 に基づいて入力文を評価した結果を以下の表 6 に示す。表 6 の は出力候補文 3 の翻訳確率が出力候補文 1 と出力候補文 2 よりも高い文の数、×は出力候補文 3 の翻訳確率が出力候補文 1 と出力候補文 2 よりも低い文の数である。

表 5 翻訳確率の計算式 1 を用いた翻訳結果

入力文	彼女 は役を派手に演じすぎた。	
参照文	She overdid her part .	
対訳学習文	日	彼 は役を派手に演じすぎた。
	英	He overdid his part .
1 対 1 に変数化した文パターンを用いる翻訳 1		
変数化する対訳句 1	日	彼
	英	He
文パターン 1	日	X0 は役を派手に演じすぎた。
	英	X0 overdid his part .
代入する対訳句 1	日	彼女
	英	She
言語生成確率 1	-37.818663	
対訳句確率 1	-1.057094	
文パターン確率 1	-654.503993	
出力候補文 1	She overdid his part .	
翻訳確率 1	-693.379750	
1 対 1 に変数化した文パターンを用いる翻訳 2		
変数化する対訳句 2	日	彼
	英	his
文パターン 2	日	X0 は役を派手に演じすぎた。
	英	He overdid X0 part .
代入する対訳句 2	日	彼女
	英	her
言語生成確率 2	-65.287694	
対訳句確率 2	-2.159697	
文パターン確率 2	-659.626271	
出力候補文 2	He overdid her part .	
翻訳確率 2	-727.073.662	
1 対多に変数化した文パターンを用いる翻訳		
文パターン 3	日	X0 は役を派手に演じすぎた。
	英	X0 overdid X0' part .
言語生成確率 3	-68.496735	
対訳句確率 3	-3.216791	
文パターン確率 3	-450.378065	
出力候補文 3	She overdid her part .	
翻訳確率 3	-522.091591(選択)	

表6 式1を用いた翻訳結果の評価(100文)

	×
100	0

表6より,すべての入力文に対して正しい出力候補文が選択されていることが分かる.

6 考察

6.1 式2を用いた翻訳確率の計算

対訳句確率と文パターン確率の計算方法を以下の式2に示す.ここで, J_n は対訳句中の日本語単語, E_m は対訳句中の英語単語, N は日本語単語数, M は英語単語数, $P(J_n|E_m)$ は英語単語 E_m が日本語単語 J_n に翻訳される確率(GIZA++の値)を表す.

$$\log_2 P\left(\frac{J_0 \dots J_{N-1}}{E_0 \dots E_{M-1}}\right) = \prod_{m=0}^{M-1} \sum_{n=0}^{N-1} \log_2 P(E_m|J_n) + \prod_{n=0}^{N-1} \sum_{m=0}^{M-1} \log_2 P(J_n|E_m) \quad (2)$$

1対多の対訳句を含む文の翻訳において,別の翻訳確率の計算式2を用いた翻訳結果の評価結果を以下の表7に示す.

表7 式2を用いた翻訳結果の評価(100文)

	×
6	94

表6と表7を比較すると,式2は式1よりも翻訳精度が大幅に減少することが分かる.

なお,翻訳確率の計算方法は,式1と式2の他にも存在する.今後,翻訳確率の計算方法を調査していく.

6.2 1対3の対訳句を含む文

本研究で扱っていない1対3の対訳句を含む文の例を表8に示す.1対3の対訳句を含む文の翻訳は,1対2の対訳句を含む文の翻訳よりも難しい.

表8 1対3の対訳句を含む文の例

日	彼女 は親類の厄介になった。
英	She inflicted herself on her relatives .

6.3 複数の1対多の対訳句を含む文

複数の1対多の対訳句を含む文の例を以下の表9に示す.表9では,日本語句“彼女”と英語句“She”,日本語句“彼女”と英語句“her”は1対多の対訳句である.同様に,日本語句“娘”と英語句“daughter”,日本語句“娘”と英語句“her”も1対多の対訳句である.

しかし,“daughter”に係る英語句“her”は日本語句“彼女”を指し,“attitude”に係る英語句“her”は日本語句“娘”を指す.本研究で提案した文パターンの作成方法では,日本語句と英語句が正しく変数化できない.

表9 2つの1対多の対訳句を含む文の例

日	彼女 は 娘 の態度を叱った。
英	She scolded her daughter for her attitude .

6.4 人称代名詞以外の1対多の対訳句を含む文

人称代名詞以外の1対多の対訳句を含む文の例を以下の表10に示す.表10では,日本語句“私の父”と英語句“My father”,日本語句“私の父”と英語句“his”は1対多の対訳句である.

日本語句“私の父”と英語句“My father”は一般名詞であり,英語句“his”は人称代名詞である.人称代名詞以外の1対多の対訳句を含む文の翻訳は,今後の課題である.

表10 人称代名詞以外の1対多の対訳句を含む文の例

日	私の父 は失業した。
英	My father lost his job .

7 おわりに

本実験では,1対多の対訳句を含む文の問題について説明し,1対多の対訳句を含む文の翻訳において正しい翻訳結果を出力できる文パターンの作成手法を提案した.実験では,1対多の対訳句を含む文の翻訳において,翻訳確率の計算式1は正しい出力候補文を選択した.

参考文献

- [1] 江木孝史“句に基づく対訳句パターンの自動作成と統計的手法を用いた英日パターン翻訳”,言語処理学会第20回年次大会, A6-2, pp.951-954, 2014.
- [2] Franz Josef Och, Hermann Ney:“ A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, pp.19-51, 2003.
- [3] 村上仁一, 藤波進“日本語と英語の対訳文対の収集と著作権の考察”, 第一回コーパス日本語学ワークショップ, pp.119-130. 2012.
- [4] 村上仁一, 森島葵“省略された人称代名詞を含む文のパターンに基づく統計翻訳 2012.