TDSMT で作成された自動対訳句の前後環境を用いた精度向上

中島 綜太○　　村上 仁一
鳥取大学 自然言語処理研究室
b16t2075h@edu.tottori-u.ac.jp
murakami@tottori-u.ac.jp

1 はじめに

機械翻訳において, 相対的意味論に基づく変換主導型統計機械翻訳 TDSMT[1](Transfer Driven Statistical Machine Translation : 以下 TDSMT と記述する) が提案されている. TDSMT は変換テーブルを用いて, 入力文を変換し, 出力文を作成する. 変換テーブルは “「*A* が *B*」ならば「*C* は *D*」” で表現できる. 変換テーブルは学習文対から自動作成する. しかし, 誤った変換テーブルを作成することがある. そこで, 本研究は, 前後環境を利用して誤った変換テーブルを削除する方法を提案する.

2 変換テーブル作成方法

変換テーブルは学習文対から自動作成する. 以下に変換テーブルの作成手順を示す. また, 表 1 に変換テーブルの作成過程の例を示す.

手順 1　対訳単語の作成
学習文対 (1), (2) と対訳単語確率 (IBM model 1[2]) を利用して対訳単語を作成する. 対訳単語は変換テーブルの *A* と *B* の部分に相当する.

手順 2　単語レベル文パターンの作成
手順 1 で作成した対訳単語に相当する部分を変数化し, 単語レベル文パターンを作成する.

手順 3　変換テーブルの作成
学習文対 (2) と単語レベル文パターンを照合する. 変数化した対訳単語と, 変数に当たる対訳句は変換テーブルの *C* と *D* の部分に相当する.

表 1 変換テーブルの作成過程

対訳単語 (手順 1)	日本語	医者		
	英語	doctor		
学習文対 (1)	日本語	私 の 父 は 医者 だ 。		
	英語	My father is a doctor .		
単語レベル文パターン (手順 2)	日本語	*X1* の *X2* は *X3* だ 。		
	英語	*X1 X2* is a *X3* .		
学習文対 (2)	日本語	私 の 母 は 教師 だ 。		
	英語	My mother is a teacher .		
X3 の 変換テーブル (手順 3)	*A*	医者	*B*	doctor
	C	教師	*D*	teacher

3 問題点

TDSMT は誤った変換テーブルが存在する場合, 誤った翻訳結果を出力する. 表 2 に誤った変換テーブルの作成過程を示す. 表 2 において, *X2* の変換テーブルの「*C*」は「髪」である. しかし,「*D*」は「dyed」である.

表 2 誤った変換テーブルの作成過程

対訳単語 (手順 1)	日本語	英語		
	英語	English		
学習文対 (1)	日本語	私 は 英語 を 勉強し た 。		
	英語	I studied English .		
単語レベル文パターン (手順 2)	日本語	*X1* は *X2* を *X3* た 。		
	英語	*X1 X3 X2* .		
学習文対 (2)	日本語	彼女 は 髪 を 染め た 。		
	英語	She had her hair dyed .		
X3 の 変換テーブル (手順 3)	*A*	英語	*B*	English
	C	髪	*D*	dyed

4 提案手法 (変換テーブル選択)

本研究は誤った変換テーブルの削除を目的とする. 変換テーブルは句の置き換えを利用している. 本手法は, 分布仮説 [3] を利用し, 前後環境が等しい句は置き換えが可能という考えに基づく. 前後環境は句の前後の単語を利用する. 具体的には, 前後の単語の比較を学習文対内にて行う. そして, 一致する前後の単語の組み合わせが存在しなければ変換テーブルを削除する. 変換テーブル選択は日本語と英語を利用する. 以下は提案手法の手順である.

手順 1　変換テーブルの句 (*A*,*B*,*C*,*D* に相当) の前後の単語を学習文対から取り出す

手順 2　手順 1 で取り出した前後の単語を比較する
日本語選択 : *A* と *C* の前後の単語を比較する
英語選択 : *B* と *D* の前後の単語を比較する

手順 3　合致する前後の単語が無ければ変換テーブルを削除する

なお, 提案手法はモノリンガルコーパスのみを利用して行うことが可能である.

4.1 学習文対の日本語で選択

表 3 の変換テーブルを表 4 の学習文対を利用して選択する場合を例に説明する.

表 3 変換テーブルの例

A	嫌いだっ	*B*	disliked
C	追いかけ	*D*	made

表 4 学習文対 (日本語側)

日本語文 (1)	文頭 私 は 彼 が 嫌いだっ た 。
日本語文 (2)	文頭 私 は 彼女 を 追いかけ た 。

表 3 の句 A , C の前後の単語を表 4 の学習文対内 (日本語側) で比較する.

- 学習文対 (1) において “**嫌いだっ**” の前は「が」, 後ろは「た」となっている
- 学習文対 (2) において “**追いかけ**” の前は「を」, 後ろは「た」となっている

従って, 表 3 のテーブルを削除する.

4.2 学習文対の英語で選択

表 3 の句 B , D の前後の単語を表 5 の学習文対内 (英語側で比較する.

表 5 学習文対 (英語側)

英語文 (1)	文頭 I disliked him .
英語文 (2)	文頭 I made after her .

4.3 学習文対の英語と日本語で選択

日本語, 英語の両者で選択を行う.

5 実験

実験は TDSMT で変換テーブルを作成した後に, 4.1, 4.2, 4.3 節の提案手法を用いて行う. 実験の評価は選択前の変換テーブルと選択後の変換テーブルの総数と精度を比較して行う. 変換テーブルの精度評価は C と D の訳の正誤によって判断する. 表 6 に変換テーブル作成に用いた学習文対の総数を示す. なお. 表 6 の学習文対内で提案手法の比較を行なっている.

表 6 テーブル作成に使用した学習文対の総数

学習文対	159,998 対

6 実験結果

6.1 テーブル数

選択前と提案手法を行った選択後のテーブル総数を表 7 に示す.

表 7 テーブル総数

	テーブル数
選択前	4,308,398
日本語選択	2,851,461
英語選択	1,285,035
日本語 & 英語選択	986,031

6.2 テーブル精度

選択前と選択後の変換テーブルからランダムに 100 個抜き出し, 精度を調査した. 精度評価は変換テーブルの C と D の訳の正誤によって判断した. 精度調査の結果を表 8 に示す. 評価基準は以下のようにした.

○: C, D の訳が正しい
△: 不足している部分がある, または不必要な部分がある
×: C, D の訳が間違っている

表 8 変換テーブル精度調査

	○	△	×
選択前	67	30	3
日本語選択	76	20	3
英語選択	85	13	2
日本語 & 英語選択	86	13	1

6.2.1 変換テーブル具体例 : 日本語選択

○と評価したテーブルの例を表 9 に示す. △と評価したテーブルの例を表 10 に示す. × と評価したテーブルの例を表 11 に示す.

表 9 ○としたテーブル (日本語選択)

変換テーブル			
A	見捨て	B	disowned
C	否認 し	D	deserted
日本語文 (1)	西川 君 の ドイツ 語 も 見捨て た もの で は ない 。		
日本語文 (2)	その 交渉 へ の いかなる 関与 も 否認 し た 。		

表 10 △としたテーブル (日本語選択)

変換テーブル			
A	患者	B	patient
C	株	D	value of her stock
日本語文 (1)	患者 の 容体 は 急 に 悪化 し た 。		
日本語文 (2)	彼女 の 株 の 価値 が かなり 損なわ れ た 。		

表 11 × としたテーブル (日本語選択)

変換テーブル			
A	やっ	B	did
C	得 なかっ	D	derived
日本語文 (1)	彼 は 人殺し も やっ た 。		
日本語文 (2)	宗教 から は 何 の 慰め も 得 なかっ た 。		

6.2.2 変換テーブル具体例 : 英語選択

○と評価したテーブルの例を表 12 に示す. △と評価したテーブルの例を表 13 に示す. × と評価したテーブルの例を表 14 に示す.

表 12 ○としたテーブル (英語選択)

変換テーブル			
A	変え	B	changed
C	上 の 方 へ 引き上げ	D	pulled up the top part of
英語文 (1)	I changed my plans .		
英語文 (2)	I pulled up the top part of my trousers .		

表 13 △としたテーブル (英語選択)

変換テーブル			
A	間違え	B	mistook
C	思いきっ て 言葉 を かけ	D	hazarded some remarks to break
日本語文 (1)	I mistook the hospital for a hotel .		
日本語文 (2)	I hazarded some remarks to break the monotony of the journey .		

表 14 × としたテーブル (英語選択)

変換テーブル			
A	取り下げ	B	withdrawn
C	障害 物 が 何一つ	D	fair for our advance
英語文 (1)	The request was withdrawn .		
英語文 (2)	The way was fair for our advance .		

6.2.3 変換テーブル具体例：日本語 & 英語選択

○と評価したテーブルの例を表 15 に示す. △と評価したテーブルの例を表 16 に示す. × と評価したテーブルの例を表 17 に示す.

表 15 ○としたテーブル (日本語 & 英語選択)

変換テーブル			
A	本	*B*	book
C	快い 驚き の ショック	*D*	pleasant shock of surprise
日本文 (1)	本 を 買っ た 。		
日本語文 (2)	快い 驚き の ショック を 受け た 。		
英語文 (1)	文頭 I bought a book .		
英語文 (2)	文頭 I felt a pleasant shock of surprise .		

表 16 △としたテーブル (日本語 & 英語選択)

変換テーブル			
A	ウイスキー	*B*	whiskey
C	彼 は 次 の 歌手	*D*	next singer surprise
日本語文 (1)	文頭 ウイスキー を 1 杯 もらおう 。		
日本語文 (2)	文頭 彼 は 次 の 歌手 を 紹介 し た 。		
英語文 (1)	文頭 He drank a little of the whiskey .		
英語文 (2)	文頭 He announced the next singer .		

表 17 × としたテーブル (日本語 & 英語選択)

変換テーブル			
A	任務	*B*	duty
C	道順	*D*	town hall
日本語文 (1)	文頭 故意 に 自分 の 任務 を 回避 し た 。		
日本語文 (2)	文頭 市役所 へ の 道順 を 尋ね た 。		
英語文 (1)	文頭 They are eager to be released from the duty .		
英語文 (2)	文頭 He asked directions to the town hall .		

7 考察

7.1 テーブル数

表 7 より, テーブル数は日本語選択の手法が英語選択の手法よりも多かった. よって, 選択に使用する言語に依ってテーブル数が異なる原因を, 表 18 の変換テーブルの選択を例として考察する.

表 18 変換テーブル例

A	サッカー	*B*	soccer
C	野球	*D*	baseball

7.1.1 英語選択

英語においては, 動詞が三単元, 過去形のような活用の変化によりスペルが変化する. また, 名詞は単数形と複数形をとりスペルが変化する.

表 18 の変換テーブルを表 19 の英語コーパス文を利用して英語選択を行なう例を考える.

表 19 英語コーパス文の例

英語文 (1)	I play soccer .
英語文 (2)	He plays baseball .

B である「soccer」の前の単語は表 19 の英語文 (1) において「play」, 後の単語は「.」である. *D* である「baseball」の前の単語は英語文 (2) により「plays」, 後の単語は「.」である. よって, 英語選択では表 18 の変換テーブルは *B* と *D* の前後の単語が合致しないことにより削除される. 従って, 英語選択においては, 単語の活用によりスペルが変化することで, 前後の単語が合致しない場合があると考える.

7.1.2 日本語選択

日本語文では句の前後に「の」,「は」,「を」のような助詞が存在する. 日本語選択においても, 前後の単語が合致した際の前後の組み合わせは助詞の組み合わせが多くみられた.

表 18 の変換テーブルを表 20 のコーパスを利用して英語選択を行なう例を考える. *A* である「サッカー」の前の単語は表 19 の日本語文 (1) より「は」, 後の単語は「を」である. *C* である「野球」の前の単語は日本語文 (2) においては前の単語は「は」, 後の単語は「を」である. よって, 日本語選択では表 18 の変換テーブルは *A* と *C* の前後の単語が合致することにより削除されない.

表 20 日本語コーパス文の例

日本語文 (1)	私 は サッカー を する 。
日本語文 (2)	彼 は 野球 を する 。

つまり, 日本語選択では, 日本語文中の助詞によって前後の単語が合致し, 英語選択を行った場合削除されるテーブルが削除されない場合がある. よって, 日本語文の句の前後に助詞が存在することが日本語選択において最も多くテーブル数が残った原因だと考える.

7.1.3 モノリンガルコーパスの利用

本手法はモノリンガルコーパスを追加利用することが可能である. モノリンガルコーパスによって, 選択に利用する文数を増やし, 英語選択後のテーブル数を増加させる. そして, 選択に利用する文数が増えることにより, 活用の変化に対応することが可能になる. 具体的には, 表 21 の追加文によって, 表 18 の変換テーブルは英語選択後に削除されなくなる.

表 21 英語モノリンガルコーパス文の追加例

追加文	He plays soccer .
英語文 (1)	I play soccer .
英語文 (2)	He plays baseball .

7.2 選択前に × とした変換テーブルの選択結果

表 8 の変換テーブルの精度調査結果より, × と評価した選択前のテーブルは 3 個存在した. × と評価した 4 個のテーブルを表 22. 23, 24 に示す.

表 22 × としたテーブル (1) (選択前)

変換テーブル			
A	馬	*B*	horse
C	すり に ご	*D*	against pickpockets

表 23 × としたテーブル (2) (選択前)

変換テーブル			
A	すっかり	*B*	completely
C	猛烈 な 駆け足 の あと 湯気 を	*D*	after a hard

表 24 × としたテーブル (3) (選択前)

変換テーブル			
A	建物	*B*	building
C	の 動議	*D*	to adjourn

選択前の 4 個の × とした変換テーブルの選択結果を調査した. 調査結果を表 25 に示す.

表 25 より, 選択前の 3 個の × とした変換テーブルは日本語選択, 英語選択の両手法によって全て削除された. ま

表 25 選択前に × とした 4 つの変換テーブルの選択結果

	削除された数	残った数
日本語選択	3	0
英語選択	3	0
日本語 & 英語選択	3	0

た, これら 3 つの変換テーブルは, 両選択で全て削除されたことによって, 日本語 & 英語選択 でも全て削除された.

7.3 提案手法の改善

ここでは提案手法の改善策について述べる. 提案手法の改善策として, △と評価したテーブルを減らすことに着目する.

7.3.1 日本語選択

表 7 より, 日本語選択後のテーブル数は英語選択後よりも多い. しかし, 表 8 より, △と評価したテーブルの数は英語手法よりも多い. 日本語選択で△と評価した変換テーブルの例を表 26, 27 に示す.

表 26 △としたテーブル (1) (日本語選択)

変換テーブル			
A	乗り遅れ	*B*	missed
C	ニス を 薄く 塗っ	*D*	gave the desk a thin coat
日本語文 (1)	彼 は 机 に ニス を 薄く 塗っ た 。		
日本語文 (2)	2 分 違い で 電車 に 乗り遅れ た 。		

表 27 △としたテーブル (2) (日本語選択)

変換テーブル			
A	集まっ	*B*	gathered
C	軒 を 連ね て い	*D*	lined with shops
日本語文 (1)	通り に は 店 が 軒 を 連ね て い た 。		
日本語文 (2)	有志 の 者 が 集まっ た 。		

表 26 において合致した前後の単語は 「に」 ,「た」である. また, 表 27 において合致した前後の単語は「が」 ,「た」である. 表 26 と表 27 の合致した前後の単語は共通して助詞を含む,

そこで, 前後 2 単語を比較する改善策を提案する. 前後 2 単語を比較することで, より正確に前後環境の利用が可能になると考えられる. 例として, 表 26 の変換テーブルを前後 2 単語を比較して選択を行なう. *A* である「乗り遅れ」の前の 2 単語は,「電車」「に」, 後の 2 単語は,「た」「。」 である. *C* である「ニス を 薄く 塗っ」の前の 2 単語は,「机」「に」, 後の 2 単語は「た」「。」 である. よって, 表 26 は前後 2 単語を比較することで選択において削除される. 表 27 も同様に削除される. したがって, 前後 2 単語を比較することで△のテーブルが多く残る問題を改善できると考えられる.

7.3.2 英語選択

英語選択後の変換テーブルの中で△と評価したテーブルの合致した前後の単語を調査する. 英語選択で△と評価した変換テーブルの例を表 28, 29 に示す.

表 28 △としたテーブル (1) (英語選択)

変換テーブル			
A	魚	*B*	fish
C	すばやく ズボン	*D*	pair of pants
日本語文 (1)	He swims like a fish .		
日本語文 (2)	I quickly pulled on a pair of pants .		

表 29 △としたテーブル (2) (英語選択)

変換テーブル			
A	銀行	*B*	bank
C	5 月 6 日 の	*D*	morning of May 6
日本語文 (1)	She put the money in the bank .		
日本語文 (2)	The vessel arrived at the port the morning of May 6 .		

表 28 において合致した前後の単語は 「a」 ,「.」である. また, 表 29 において合致した前後の単語は「the」 ,「.」である. 表 28 と表 29 の合致した前後の単語は共通して冠詞を含む,

日本語選択の改善策と同様に, 英語選択でも前後 2 単語を比較する改善策が有用だと考えられる. 例として, 表 28 の変換テーブルを前後 2 単語を比較して選択を行なう. *B* である「bank」の前の 2 単語は,「in」「the」 , 後の単語は,「.」 である. 後の単語が「.」のみであり, 後の単語が 2 つに満たない場合は,「.」の 1 単語のみを後の環境として利用する. *D* である「moring of May 6」の前の 2 単語は,「port」「the」 , 後の単語は 「.」 である. よって, 表 28 は前 2 単語, 後ろ 1 単語を比較することで選択において削除される. 表 29 も同様に削除される.

7.4 テーブルしきい値

TDSMT はテーブルに確立を付与し, 順位付けを行う [4]. そして, 順位が低いテーブルはしきい値を用いた枝刈りにより削除する. 確率付与には IBM Model を利用している.

本実験では枝刈り後のテーブルに対して提案手法を行った. 枝狩り前のテーブルを用いて提案手法を行った場合, 選択後のテーブル数は枝刈り後を選択した場合よりも多くなる. しかし, 本来枝刈りにおいて削除される誤ったテーブルが残る可能性がある.

8 おわりに

本研究は, TDSMT において作成された自動対訳句の精度向上を目的とした. そして, 前後環境を利用して, 誤った変換テーブルを削除する方法を提案した. 提案手法によって, テーブルの精度が向上することを確認できた.

参考文献

[1] 安場裕人, 村上仁一 (2018). 変換主導型統計機械翻訳の提案, 言語処理学会 第 24 回年次大会 発表論文集.

[2] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer (1993). The mathematics of statistical machine translation:Parameter Estimation. *Computational Linguistics*.

[3] Zellig S.Harris (1954). Distributional structure, *Word*, Vol. 10, No.23, pp. 146-162

[4] 三井 盛彰 (2020). global の変数確率を利用した変換テーブルの精度向上, 2019 年度鳥取大学工学部電気情報系学科卒業論文.