変換主導型統計機械翻訳の提案

安場裕人 [*1]　村上仁一 [*2]
[*1] 鳥取大学大学院 持続性社会創生科学研究科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s132057,murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳の手法として，パターン翻訳，統計翻訳等が研究されてきた．しかし，人手の翻訳には及ばない．そこで我々はあらたな手法として，変換主導型機械翻訳 (TDMT:Transfer Driven Machine Translation) を提案する．この手法では変換ルールを利用する．変換ルールを利用して学習文対を入力文と出力文に変換する．

変換ルールは“*A* が *B* ならば *C* は *D*”で表現する．日英翻訳の例では，*A* は学習文対内の日本語句，*B* は学習文対内の英語句，*C* は入力文内の日本語句，*D* は出力文内の英語句に当たる．変換ルールは人手で作成可能である．しかし，開発コストは高い．また，変換ルールの数とカバー率（翻訳可能な入力文の数）は比例していると考えられる．そこで，変換ルールを自動作成する手法を考案する．

我々は変換ルールの自動作成に成功した．また，精度の高い翻訳結果が得られた．

2 変換主導型機械翻訳 (TDMT:Transfer Driven Machine Translation)

TDMT では, 変換ルールを利用する．TDMT の概要を日英翻訳の例で示す．変換ルールは二つの対訳句（“*A*(学習文対内の日本語句) が *B*(学習文対内の英語句)”と“*C*(入力文内の日本語句) は *D*(出力文内の英語句)”）の組み合わせである．二つの対訳句を組み合わせることで，変換ルールは“*A* が *B* ならば *C* は *D*”という変換の知識を持つ．この変換ルールを用いて学習文対を変換し，入力文を翻訳する手法である．なお，この手法は Chomsky の生成文法に準じた翻訳手法である．以下に TDMT の手順を示す．

手順１　変換ルールの作成
変換ルールを人手で作成する．

表 1　変換ルールの人手作成

A が *B* ならば *C* は *D*			
A	医者	*B*	doctor
C	日本語教師	*D*	Japanese teacher

手順２　学習文対の日本語側への変換ルールの適用
変換ルールの *A* と *C* を利用して，学習文対の日本語側を入力文と一致させる．

手順３　学習文対の英語側への変換ルールの適用

表 2　学習文対の日本語側への変換ルールの適用

学習文対	日本語側	私の父は 医者 だ。
変換ルール	*A*	医者
	C	日本語教師
入力文		私の父は 日本語教師 だ。

手順２と同じ変換ルールの *B* と *D* を学習文対の英語側に適用し，出力文を作成する．

表 3　学習文対の英語側への変換ルールの適用

学習文対	英語側	My father is a doctor.
変換ルール	*B*	doctor
	C	Japanese teacher
出力文		My father is a Japanese teacher.

図 1 に TDMT の流れ図を示す．

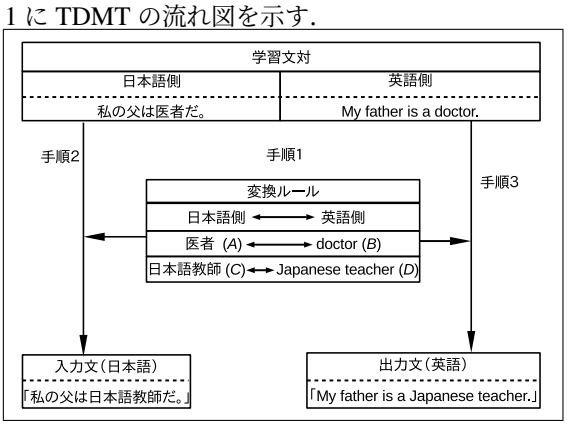

図 1　TDMT の流れ図

3 提案手法 (TDSMT:Transfer Driven Statistical Machine Translation

図 1 の TDMT では，人手で変換ルールを作成していた．しかし，高いカバー率を得るために，人手で変換ルールを大量作成するのは現実的に不可能である．そのため，我々は変換ルールを自動作成し翻訳を行う手法を提案する．

3.1 変換ルールの自動作成手法

変換ルールの自動作成手法を提案する．この方法は学習文対と対訳単語確率 (IBM model 1) を利用して，変換ルールを“*A* が *B* ならば *C* は *D*”という形式で抽出する．

最後に，変換ルールに確率を付与する．以下に日英翻訳における概要を示す．図 2 に自動作成手法の流れずを示す．

手順 1　対訳単語の作成

学習文対と対訳単語確率 (IBM model 1) を利用して対訳単語を作成する．

表 4　対訳単語の作成

学習文対	日本語側	私 の 父 は 医者 だ。
	英語側	My father is a doctor.
対訳単語 1	私	My
対訳単語 2	父	father
対訳単語 3	医者	doctor
etc...		

手順 2　単語レベル文パターンの作成

学習文対内で手順 1 で作成した対訳単語にあたる部分を変数化し，単語レベル文パターンを作成する．

表 5　単語レベル文パターンの作成

学習文対	日本語側	私 の 父 は 医者 だ。
	英語側	My father is a doctor.
対訳単語 1	私	My
対訳単語 2	父	father
対訳単語 3	医者	doctor
単語レベル	日本語側	*X1* の *X2* は *X3* だ。
文パターン	英語側	*X1 X2* is a *X3*.

手順 3　変換ルールの作成

学習文対に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を変換ルールとする．

表 6　変換ルールの作成

文パターン	日本語側	私 の 父 は 医者 だ。	
原文	英語側	My father is a doctor.	
単語レベル	日本語側	*X1* の *X2* は *X3* だ。	
文パターン	英語側	*X1 X2* is a *X3*.	
学習文対	日本語側	私 の 母 は 日本語教師 だ。	
	英語側	My mother is a Japanese teacher.	
X3 の変換	*A* が *B*	*A*:医者	*B*:doctor
ルール	*C* が *D*	*C*:日本語教師	*D*:Japanese teacher

手順 4　変換ルールへの確率の付与

変換ルールの A，B，C，D の学習文対中に出現する頻度を利用し，確率を計算する．計算式を以下に示す．

$$\text{変換ルールの確率} = \left(\frac{P(A,B)}{P(A)} * \frac{P(A,B)}{P(B)}\right) * \left(\frac{P(C,D)}{P(C)} * \frac{P(C,D)}{P(D)}\right)$$

ここで，$P(A,B)$ は A と B の学習文対中の共起頻度を，$P(A)$ は A の学習文対中の頻度を表す．

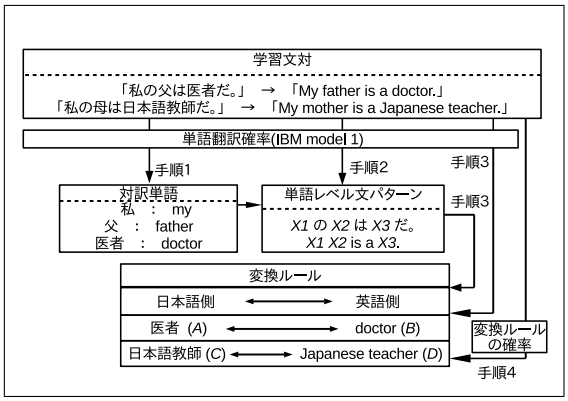

図 2　変換ルールの自動作成手法の流れ図

3.2　変換主導型統計機械翻訳 (TDSMT) の手順

TDSMT では，自動作成した変換ルールと学習文対を利用して，入力文を変換し出力候補文を作成する．そして，統計を用いて，出力文を決定する．日英翻訳における，TDSMT の手順を以下に示す．

手順 1　3.1 節の手法による変換ルールの作成

変換ルールを 3.1 節の手法で作成する．

表 7　3.1 節の手法による変換ルールの作成

A が *B* ならば *C* は *D*			
A	医者	*B*	doctor
C	日本語教師	*D*	Japanese teacher

手順 2　入力文への変換ルールの適用

変換ルールの A と C を利用して，入力文を学習文対の日本語側と一致させる．

表 8　入力文への変換ルールの適用

入力文		私の父は 日本語教師 だ。
変換ルール	*C*	日本語教師
	A	医者
学習文対	日本語側	私の父は 医者 だ。

手順 3　学習文対の英語側への変換ルールの適用

手順 2 と同じ変換ルールの B と D を学習文対の英語側に適用し，出力候補文を作成する．

表 9　学習文対の英語側への変換ルールの適用

学習文対	英語側	My father is a doctor.
変換ルール	*B*	doctor
	D	Japanese teacher
出力文		My father is a Japanese teacher.

手順 4　最終的な出力文の決定

複数の出力候補文が得られた場合，変換ルールの確率と言語モデル (trigram) により出力文を選択する．

日英翻訳における TDSMT の流れ図を図 3 に示す．

変換主導型機械翻訳 (TDMT) と変換主導型統計機械翻訳 (TDSMT) の違いを以下に示す．

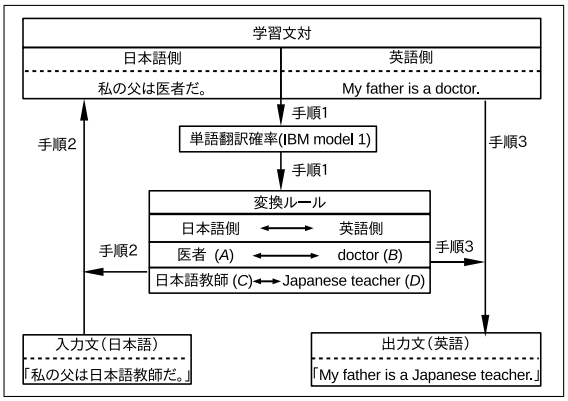

図 3　TDSMT の流れ図

表 12　作成された変換ルールの例

パターン原文	日	この 時計 は 遅れる。
	英	This clock loses.
単語レベル文パターン	日	*X00 X01* は *X02*。
	英	*X00 X01 X02*.
学習文対	日	生命 保険 の 契約 は そう は いか ない。
	英	But this is impossible for life insurance contract.
日本語句 (*A*)		時計
英語句 (*B*)		clock
日本語句 (*C*)		保険の契約
英語句 (*D*)		insurance

a　変換ルールの自動作成
b　統計を利用した出力文の決定
c　TDMT では学習文対の日本語側を変換，一方 TDSMT では入力文を変換

4　実験設定

4.1　実験目的

TDSMT の翻訳実験により，TDSMT の有効性を検証する．

4.1.1　自動作成手法の評価

変換ルールの自動作成手法の評価として，以下の項目を調査する．

- 作成された変換ルールの数
- 作成された変換ルールの精度

4.1.2　TDSMT の翻訳性能評価

TDSMT の翻訳性能の評価として，以下の項目を調査する．

- 出力を得ることのできた入力文の数
- 翻訳精度（人手評価）

4.2　実験データ

実験には電子辞書などの例文より抽出した単文コーパス [1] を用いる．使用するデータの内訳を表 10 に示す．

表 10　実験データ

学習文対	160,000 文対
入力文	800 文

5　実験結果

5.1　作成された変換ルールの数

作成された変換ルールの数の調査を行った．調査結果を表 11 に，表 12 に作成された変換ルールの例を示す．

表 11　作成された変換ルールの数

変換ルールの数	1,568,339 個

表 11 より，変換ルールの大量作成に成功した．

5.2　変換ルールの精度

作成された変換ルールの精度の調査を行った．作成された変換ルールから，100 個をランダムで抜き出し，その精度を三段階で評価した．変換ルール中の“*A* が *B*”の部分と，“*C* は *D*”の部分を別々に評価を行った．調査結果を表 13 に，評価例を表 14 に示す．表中における，○は日本語側と英語側で正しい対応をしているものである．△は日英のどちらかに余分な単語を含むものを意味する．×は日本語側と英語側で間違った対応をしているものである．

表 13　変換ルールの精度

	○	△	×
A が *B*	100	0	0
C は *D*	37	37	26

表 14　変換ルールの例

評価	変換ルール		
○	*A* が *B*	大阪	Osaka
○	*C* は *D*	100 メートル競争	100-meter race
○	*A* が *B*	学校	school
△	*C* は *D*	角を左	the corner
○	*A* が *B*	とても	very
×	*C* は *D*	宝石	of

表 13 より，一つの変換ルールで“*A* が *B*”と“*C* が *D*”のどちらも間違っているものは存在しなかった．間違った対応を含む変換ルールは全体の 25 %と少ない．

5.3　翻訳可能な入力文の数 (カバー率)

TDSMT において翻訳実験を行う．入力文 800 文における出力文を得られた文数を表 15 に，表 16 に翻訳例を示す．

表 15　翻訳成功文数 (800 文中)

出力文を得られた文数	74 文

表 15 より，TDSMT のカバー率は約 9% と低い．

表 16　翻訳例

入力文		私 の 疑惑 は 大きく なっ た。	
出力文		My suspicion increased.	
参照文		My suspicion grew.	
対訳文	日	この 時計 は 遅れる。	
	英	This clock loses.	
変換ルール 1	*A* が *B*	この	This
	C は *D*	私の	My
変換ルール 2	*A* が *B*	時計	clock
	C は *D*	疑惑	suspicion
変換ルール 3	*A* が *B*	遅れる	loses
	C は *D*	大きくなった	increased

5.4　翻訳精度（人手評価）

TDSMT において翻訳実験を行う．そして，翻訳結果に対して，人手による評価を行う． 表中における，○は入力文の意味と出力文の意味が同じもの，△は出力文に単語の過不足がある，または構文が間違っているもの，×は入力文の意味が出力文で読み取れないものである．

評価結果を表 17 に，評価例を表 18 に示す．

表 17　人手評価結果 (50 文)

○	△	×
40	6	4

表 18　TDSMT の人手評価例

評価		
○	入力文	彼 は 進歩 的 な 考え を 持っ て いる 。
	出力文	He has progressive ideas.
△	入力文	彼ら は すぐ に 財政 困難 に 陥っ た 。
	出力文	They feel to soon to financial difficulties.
×	入力文	彼女 は 家 に は い ない でしょ う。
	出力文	She generations will yet.

表 17 より，TDSMT は約 80% と高い翻訳精度を持つ．

6　考察

6.1　経験的知識を活用する変換主導型機械翻訳 [2] と本研究の違い

この研究の動機となった先行研究として，古瀬らの行った経験的知識を活用する変換主導型機械翻訳 [2] が挙げられる．古瀬らの翻訳手法と今回提案する手法との違いを考察する．

経験的知識を活用する変換主導型機械翻訳の変換知識と本研究の変換ルールは対応している．変換知識は経験的知識（実際に使用されている言語表現をもとにした知識）を利用して，人手で作成されている．また，4 つのレベル（ストリングレベル，パタンレベル，文法レベル，解析レベル）にわけ，より詳しい解析のもとに作成されている．

一方で我々は“*A* が *B* ならば *C* は *D*”という推定の知識を利用している．そのため，我々の変換ルールは二つの対訳句の対という非常に簡素な形で作成されている．また，変換ルールは自動で作成されている．さらに，自動作成の手法も非常に簡単な手法をとっている．

本研究の変換ルールは古瀬らの変換知識と比べ，非常に簡単な原理，作成方法を用いている．

6.2　カバー率について

TDSMT の問題点として，カバー率が低いことが挙げられる．このため，自動作成する変換ルールの総数を増加させる手法を考案する必要がある．

なお，自動作成された変換ルールでは，表 14 のように，人手では作成困難な対訳句の組み合わせを抽出した．よって，人手で作成した変換ルールで高いカバー率を得るのは困難であると考えられる．

6.3　翻訳精度について

TDSMT では表 16 の変換ルール 1〜3 が同じ単語レベル文パターンから作成される場合，精度の高い翻訳を出力した．同じ単語レベル文パターンから作成される変換ルールを利用する割合を出力文の選択に利用することで翻訳精度の向上を期待できる．

なお，一般的な統計翻訳である Moses との比較では，TDSMT より Moses の方が翻訳精度が高かった．これは，変換ルール中の“*C* は *D*”の部分に誤り (表 14 の△や×) を含むことが原因である．しかし，*TDSMT* では動作の記録が残る．そのため，誤りの考察や，翻訳精度向上に向けた改善が行いやすい．

6.4　パターンベース統計翻訳 [3] への応用

パターンベース統計翻訳は自動で文パターンと対訳句を作成し，翻訳を行う．しかし，対訳句は“*C* は *D*”の形式である．“*A* が *B* ならば *C* は *D*”の形式である変換ルールを利用するパターン翻訳手法を開発することが今後の研究として挙げられる．

なお，パターンベース統計翻訳において，翻訳に利用する各対訳句を作成する単語レベル文パターンの原文が一致する場合が TDSMT の手法となる．

7　おわりに

本研究では，変換主導型統計機械翻訳と，そこで利用する変換ルールの自動作成手法を提案した．結果として，変換主導型統計機械翻訳では精度の高い出力文の作成に成功した．また，変換ルールの自動作成手法では変換ルールの大量作成に成功した．しかし，変換主導型統計機械翻訳はカバー率が低い．従って，変換ルールの増加や手法の改良よるカバー率の向上が必要である．

参考文献

[1] 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.

[2] 古瀬蔵，隅田英一郎，飯田仁．“経験的知識を活用する変換主導型機械翻訳 ”，情報処理学会論文誌，Vol.35 No.3 pp.414-425．Mar.1994.

[3] 江木孝史．“句に基づく文パターンを用いた英日翻訳 ”， 修士論文， 2013.