相対的意味論に基づく変換主導型統計機械翻訳 ～未知語の出力～

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
[*1]s132057@ike.tottori-u.ac.jp　[*2]murakami@eecs.tottori-u.ac.jp

1 はじめに

機械翻訳には様々な手法がある．その中の一つに“相対的意味論に基づく変換主導型統計機械翻訳 [1]” [*1]（以下，TDSMT）がある．TDSMT では，対訳文と「*A* が *B* ならば *C* は *D*」で表現する変換テーブルを利用して翻訳を行う．対訳文を変換し，翻訳を行うため，文法を順守した翻訳が期待できる．しかし，入力文に対して翻訳可能な文の割合は低い．そのため，本論文では翻訳候補文中に未知語を出力する手法を提案することで，翻訳可能な入力文の数の向上を図る．

2 従来手法 “相対的意味論に基づく変換主導型統計機械翻訳 (TDSMT)”

従来の相対的意味論に基づく変換主導型統計機械翻訳 (以下，TDSMT) では，対訳文から IBM model1(単語翻訳確率) を利用して変換テーブルを作成する．次に対訳文と変換テーブルを利用して翻訳を行う．以下に手法の詳細を述べる．

2.1 変換テーブルの自動作成手法

TDSMT では，学習として，変換テーブルを自動作成する．変換テーブルの形式は「*A* が *B* ならば *C* は *D*」である．変換テーブルの自動作成手法は大きく分けて３つの手順から成る．日英翻訳による例を示す．

1 対訳単語の作成

IBM model1(単語翻訳確率) を用いて，対訳文から対訳単語を作成する．対訳単語の作成例を表 1 に示す．

表 1　対訳文から作られる対訳単語

対訳文	
私 の 医者 は 親切 だ	My doctor is gentle

対訳単語	
私	My
医者	doctor
親切	gentle
etc...	

2 文パターンの作成

対訳単語と対訳文（文パターン原文[*2]）を照合する．対訳文中で対訳単語にあたる箇所を変数化し，文パターンを作成する．文パターンの作成例を表 2 に示す．

3 変換テーブルの作成

文パターンと対訳文[*3]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テーブルとする．また，IBM model1 を用いて，変換テーブルの適用確率を付与する．この適用確率を利用して，変換テーブルはある閾値で枝刈りを行う．変換テーブルの作成例を表 3 に示す．

表 2　対訳単語と対訳文から作られる文パターン

対訳単語	
私	My
医者	doctor
親切	gentle

対訳文（文パターン原文）	
私 の 医者 は 親切 だ	My doctor is gentle

文パターン	
*X*1 の *X*2 は *X*3 だ	*X*1 *X*2 is *X*3

表 3　文パターンと対訳文から作られる変換テーブル

文パターン	
*X*1 の *X*2(医者) は *X*3 だ	*X*1 *X*2(doctor) is *X*3

対訳文（文パターン原文）	
私 の 医者 は 親切 だ	My doctor is gentle

対訳文	
彼 の 患者 は 冷静 だ	His patient is calm

*X*2 における変換テーブル	
日本語側	英語側
A:医者	*B*:doctor
C:患者	*D*:patient

図 1 に変換テーブルの自動作成手法のフローチャートを示す．

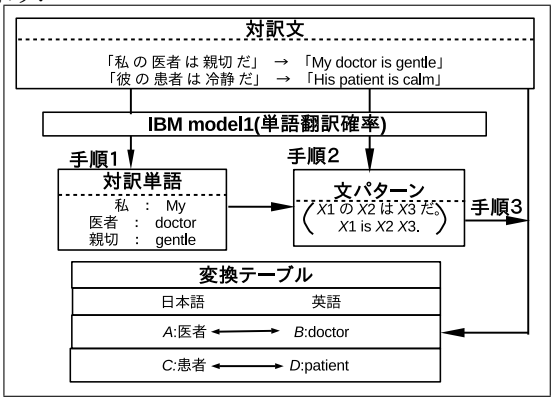

図 1　変換テーブルの自動作成手法のフローチャート

この例では「“医者”が“doctor”ならば“患者”は“patient”」が得られる．

2.2 翻訳手法

TDSMT における翻訳には，対訳文と 2.1 節で作成した変換テーブルを利用する．翻訳手法は大きく分けて３つ

[*1] この名称の“相対的意味論に基づく”は古瀬らの“経験的知識を活用する変換主導型機械翻訳 [2]”と差を明確にするための名称である．
[*2] 文パターンの作成に利用した対訳文を“文パターン原文”と呼ぶ
[*3] この対訳文は文パターン原文とは別の対訳文を利用する．

の手順から成る．日英翻訳における概要を示す．

1 変換テーブルの選択

変換テーブルの日本語側を利用して，入力文を対訳文の日本語側と一致させる変換テーブルを選択する．変換テーブルの選択例を表 4 に示す．

表 4　変換テーブルの選択例

対訳文	
彼 は 私 の 医者 だ	He is my doctor

変換テーブル	
日本語側	英語側
A:医者	*B*:doctor
C:患者	*D*:patient

入力文
彼 は 私 の 患者 だ

2 変換テーブルの適用

選択した変換テーブルの英語側に従い，対訳文の英語側を変換し，出力候補文を作成する．変換テーブルの適用例を表 5 に示す．

表 5　変換テーブルの適用例

対訳文	
彼 は 私 の 医者 だ	He is my doctor

抽出された変換テーブル	
日本語側	英語側
A:医者	*B*:doctor
C:患者	*D*:patient

出力候補文
He is my patient

3 出力文の決定

出力候補文の中から，言語モデルと変換テーブルの適用確率を用いて，出力文を決定する．

図 2 に従来手法のフローチャートを示す．

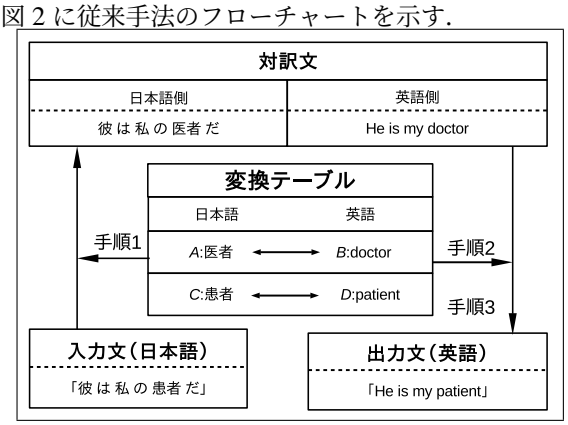

図 2　従来手法のフローチャート

表 6 に従来手法による実際の翻訳例を示す．入力文の「彼は大学で法律を修めた」に対して，「He completed the law at college」が出力される．

2.3　従来手法の問題点

従来手法では，対訳文と 2.1 節の方法で作成された変換テーブルの組み合わせだけでは入力文に対する翻訳可能な文の割合（以下，カバー率）は低い．また，対訳文に存在しない語が入力文中に出現した場合，翻訳ができない．

3　提案手法

本論文では，2.3 節の問題を解決するために，未知語を出力する手法を提案する．具体的には，入力文の単語を

表 6　従来手法の翻訳例

入力文	彼 は (大学) で (法律) (を) (修めた)	
参照文	He studied law at college	
対訳文 (日)	彼 は (テニス) で (相手) (を) (破った)	
対訳文 (英)	He (defeated) (his) (opponent) at (tennis)	
変換テーブル 1	*A*:テニス	*B*:tennis
	C:大学	*D*:college
変換テーブル 2	*A*:相手	*B*:opponent
	C:法律	*D*:law
変換テーブル 3	*A*:を	*B*:his
	C:を	*D*:the
変換テーブル 4	*A*:破った	*B*:defeated
	C:修めた	*D*:completed
出力文	He (completed) (the) (law) at (college)	

利用して未知語出力用変換テーブルを作成する．未知語出力用変換テーブルの形式は「*A* が *B* ならば *C* は *C*」である．次に，2.2 節と同様に翻訳を行う．

3.1　未知語出力用変換テーブルの自動作成手法

未知語出力用変換テーブルは入力文と，2.1 節で作成した対訳単語から作成する．未知語出力用変換テーブルの自動作成手法は大きく分けて 2 つの手順から成る．日英翻訳による例を示す．

1 未知語出力用変換テーブルの作成

入力文と 2.1 節で作成された対訳単語から未知語出力用変換テーブルを作成する．未知語出力用変換テーブルの作成例を表 7 に示す．なお，今回の実験では翻訳における計算時間の短縮のため，「*C* ならば *C*」の部分にあたる単語は 2 単語以下に制限する．

表 7　未知語出力用変換テーブルの作成

対訳単語	
医者	doctor

入力文	
彼女 は 私 の 母 だ	

未知語出力用変換テーブル	
日本語側	英語側
A:医者 *C*:彼女	*B*:doctor *C*:彼女
A:医者 *C*:彼女 は	*B*:doctor *C*:彼女 は
etc...	

2 変換テーブルへの追加

未知語出力用変換テーブルを 2.1 節で作成した変換テーブルに追加する．この際，未知語出力用変換テーブルの適用確率に重みをつける．

図 3 に変換テーブルの自動作成手法のフローチャートを示す．

4　実験環境

4.1　実験データ

本実験では，電子辞書などの例文より抽出した単文コーパスを用いる．使用するデータの内訳を表 8 に示す．

表 8　実験データ

対訳文	160,000 文
入力文	1,000 文

4.2　評価方法

本研究では提案手法を 4 つの基準で評価する．

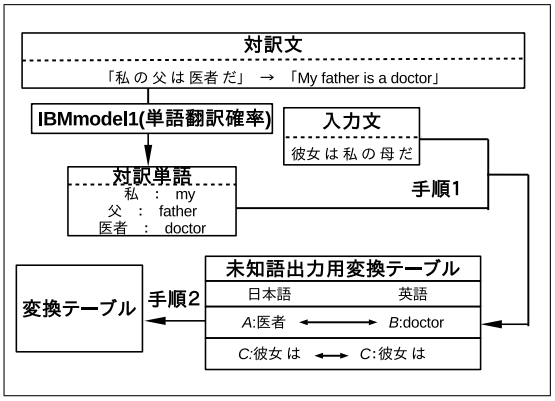

図 3　未知語出力用変換テーブルの自動作成手法のフローチャート

1) 作成した変換テーブルの数
   2.1 節の方法で作成した変換テーブルの数を調査する．
2) 作成した変換テーブルの精度
   2.1 節の方法で作成した変換テーブルの中からランダムで 100 個抜き出し，人手評価を行い精度を調査する．
3) 翻訳に利用する変換テーブルの数
   従来手法と提案手法で翻訳に利用する変換テーブルの数を比較調査する．なお，変換テーブルの数は入力文一文あたりの平均数で示す．
4) カバー率の調査
   TDSMT で翻訳実験を行う．1,000 文を翻訳し，カバー率を調査する．なお，この調査においては，翻訳の正誤は問わない．

5　実験結果

5.1　作成した変換テーブルの数

変換テーブルの自動作成手法で作成した変換テーブルの数を以下の表 9 に示す．

表 9　作成した変換テーブルの数

変換テーブルの数	5,744,175(個)

変換テーブルの例を表 10 に示す．

表 10　作成した変換テーブルの例

変換テーブル	*A*:反省	*B*:Reflection
	C:嵐	*D*:The storm
対訳文 1 (日)	(反省) は (知恵) を (増す)	
対訳文 1 (英)	(Reflection) (increases) (wisdom)	
対訳単語 (*X*00)	増す	increases
対訳単語 (*X*01)	知恵	wisdom
対訳単語 (*X*02)	反省	Reflection
文パターン (日)	*X*02 は *X*01 を *X*00	
文パターン (英)	*X*01 *X*00 *X*01	
対訳文 2 (日)	(嵐) は (勢い) を (増す)	
対訳文 2 (英)	(The storm) (grew in } {force)	

5.2　作成した変換テーブルの精度

表 9 の変換テーブルの数から，ランダムに 100 個を抜き出し，その精度を人手評価した．なお，「*A* が *B* ならば *C* は *D*」の「*A* が *B*」の部分と「*C* は *D*」の部分を分割し，それぞれの対応を評価した．評価基準を表 11 に示す．

評価結果を表 12 に示す．表 12 より変換テーブルの精度は高いことがわかる．

表 11　変換テーブルの評価基準

○	日本語側と英語側が正しく対応している
△	日英のどちらかに余分な単語を含む
×	日本語側と英語側が間違った対応をしている

表 12　作成した変換テーブルの精度

	○	△	×
A が *B*	95	2	3
C は *D*	75	16	9

評価例を表 13 に示す．

表 13　作成した変換ルールの評価例

評価	変換テーブル		
○	*A* が *B*	こすった	rubbed
○	*C* は *D*	を 務めた	acted as
○	*A* が *B*	魚	fish
△	*C* は *D*	彼 は 助け を	for help
○	*A* が *B*	稲	rice
×	*C* は *D*	彼女 の 指輪	diamond
×	*A* が *B*	した	They
△	*C* は *D*	を 食らった	I got a

5.3　翻訳に利用する変換テーブルの数

従来手法と提案手法で翻訳に利用する変換テーブルの数を表 14 に示す．なお，変換テーブルの数は入力文一文あたりの平均数で示す．表 14 より提案手法では従来手法と比較し，翻訳に利用する変換テーブルの数は大幅に増加している．

表 14　翻訳に利用する変換テーブルの平均数

従来手法	13,503(個)
提案手法	173,004(個)

未知語出力用変換テーブルの例を表 15 に示す．

表 15　未知語出力用変換テーブルの例

未知語出力用	*A*:戦った	*B*:fought
変換テーブル	*C*:持ち込ま れた	*C*:持ち込ま れた
対訳文 1 (日)	彼ら は 勇敢 に 戦った	
対訳文 1 (英)	They fought bravely	
入力文	その 事件 は 調停 に 持ち込ま れた	

5.4　カバー率の調査

従来手法と提案手法において入力文 1,000 文での翻訳実験を行った．各手法のカバー率を表 16 に示す．表 16 より，従来手法ではカバー率は低かった．しかし，未知語出力用変換テーブルを利用することでカバー率は大きく上昇した．

表 16　各手法のカバー率

従来手法	19.6 % (196 文/1,000 文)
提案手法	60.6 % (606 文/1,000 文)

翻訳例を表 17 に示す．

6　考察

6.1　従来手法と Moses[4] との翻訳精度の比較

従来手法で出力可能であった入力文 196 文からランダムに抽出した 100 文に対して，Moses との対比較評価を行った．評価基準を表 18 に示す．

結果を表 19 に示す．表 19 より，TDSMT は未知語を出力しない場合において，Moses とほぼ同等の翻訳精度を

表 17　提案手法の翻訳例

入力文	(国内 事情) が (よく) なった	
参照文	The domestic conditions have improved	
対訳文 (日)	(外) が (暗く) なった	
対訳文 (英)	It has gotten (dark) (outside)	
変換テーブル 1	*A*:外	*B*:outside
	C:国内 事情	*D*:国内 事情
変換テーブル 2	*A*:暗く	*B*:dark
	C よく:	*D*:better
出力文	It has gotten (better) (国内 事情)	

表 18　対比較評価の評価基準

従来手法〇	TDSMT の出力文の方が優れている
Moses 〇	Moses の出力文の方が優れている
差なし	二つの出力文の優劣がつけにくい
同一出力	二つの出力文が一致している

示した．

表 19　従来手法と Moses との対比較評価結果

従来手法〇	Moses 〇	差なし	同一出力
16	17	45	22

評価例を表 20 に示す．

表 20　従来手法と Moses との人手評価例

従来手法〇	
入力文	彼女 は ろうそく を 吹き 消した
参照文	She puffed the candle out
TDSMT	**She blew out the candle**
Moses	She is a candle

Moses 〇	
入力文	日程 を よく 調べ て ください
参照文	Please check the dates carefully
TDSMT	Please investigating a good schedule
Moses	**Please check carefully the schedule**

差なし	
入力文	一時 の 誘惑 に 負けた
参照文	He succumbed to a momentary temptation
TDSMT	He lost temptation one time
Moses	I gave in to temptation of one o'clock

6.2　提案手法と Moses との翻訳精度の比較

提案手法で出力可能であった入力文 606 文からランダムに抽出した入力文 100 文に対して，Moses との対比較評価を行った．評価基準は表 18 と同様である．結果を表 21 に示す．なお，未知語が出現した場合はその未知語が正しく翻訳されると想定して評価を行った．表 21 より，提案手法の場合では Moses に勝る翻訳精度を示した．未知語を正しく翻訳することで，TDSMT は Moses よりも精度の高い翻訳結果を得ることが期待できる．

表 21　提案手法と Moses との対比較評価結果

提案手法〇	Moses 〇	差なし	同一出力
25	16	50	9

評価例を表 22 に示す．

6.3　相対的意味論 [3]

相対的意味論は「北」の意味を「南の逆の方角」であると意味付ける．このように，ある言葉の意味を別の言葉の意味を利用して定義付ける理論である．翻訳においては，日本語の言葉の意味を英語の言葉とすることによって拡張し，利用する．つまり「日本語句 *C* の意味は英語句 *D* である.」として利用する．本論文では「*A* が *B* ならば *C* は *D*」という変換テーブルの作成の際に利用する．

表 22　提案手法と Moses との人手評価例

提案手法〇	
入力文	彼 は 国立 劇場 に 出演 した
参照文	He appeared on the stage of the National Theater
TDSMT	**He appeared on 国立 劇場**
Moses	He played a part in the National

Moses 〇	
入力文	彼 の 名前 が 思い出せ なかった
参照文	His name escaped me
TDSMT	His name 思い出せ なかった
Moses	**I couldn't think of his name**

差なし	
入力文	家族 は 社会 の 基本 単位 で ある
参照文	The family is the basic unit of society
TDSMT	The family is a 基本 単位 of society
Moses	The family is a unit of society

表 10 の変換テーブルでは，一見，*A* と *C* は非関連に感じられる．しかし，「犬が dog ならば猫は cat」の対偶を考える．この命題の対偶である「猫が cat でないならば犬は dog でない」が成立する．この対偶は表 10 の「*A* が *B* ならば *C* は *D*」が成立する証明となる．

6.4　TDSMT の利点

TDSMT は翻訳の手順を詳しく解析することが可能である．このため，改善点の探索が非常に容易である．一方，現在主流となっているニューラル機械翻訳や Moses などの句に基づく統計翻訳では，このような詳しい解析は困難である．

6.5　未知語出力用変換テーブルの制限

現在，未知語出力用変換テーブルの「*C* は *C*」の部分に単語数の制限 (2 単語以内) をつけている．この制限を外すことによりさらにカバー率を向上させることが可能となる．ただし，未知語出力用変換テーブルの数が増加するため，計算時間も増加する．

6.6　手法の改善

TDSMT による翻訳では，変換テーブルの誤りが出力文に出力される．人手評価で Moses よりも翻訳精度が低い出力文ではこの問題が多く観測された．今後は，変換テーブルにおいて，表 12 に見られる，評価△や評価 × を削減する方法を考案する．

7　おわりに

従来の"相対的意味論に基づく変換主導型統計機械翻訳"では翻訳可能な入力文は少なかった．提案手法により未知語を出力することで，翻訳可能な入力文の数は増加した．さらに，未知語が正しく翻訳できた場合，翻訳精度において，TDSMT は Moses を上回る可能性が示された．

参考文献

[1] 安場裕人，村上仁一．"変換主導型統計機械翻訳の提案 "，自然言語処理学会第 24 回年次大会，ポスター (3)，P7-9，Mar.2018.

[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] ナシーム・ニコラス・タレブ："ブラック・スワン [上]―不確実性とリスクの本質"，ハードカバー，Jun.2009.

[4] Moses: "Open Source Toolkit for Statistical Machine Translation", Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180. 2007.