統計機械翻訳の未知語処理における NMT の利用

林 和輝　　村上 仁一
鳥取大学 工学部　電気情報系学科
b16t2089h@edu.tottori-u.ac.jp
murakami@tottori-u.ac.jp

1 はじめに

“相対的意味論に基づく変換主導型統計機械翻訳 (以下, TDSMT)[1] ” は学習文対と変換テーブルを用いて翻訳を行う手法である. TDSMT は学習文対を変換し, 翻訳を行うため文法性の高い翻訳が期待できる. しかし, TDSMT は変換テーブルの学習文対と入力文が完全一致しなければ変換の適用が不可能なため, 入力文の数に対する出力文の数 (以下カバー率) が少ない.

出力可能な入力を増やすために, 安場らは未知語出力用変換テーブルを自動作成する手法を提案した. しかしながら, この手法では出力文中に未知語が日本語のまま出力される. そのため, 相当する文意が読み取れないという問題がある. そこで本研究では, TDSMT における未知語の翻訳をニューラル機械翻訳 (以下, NMT) によって行い, 翻訳精度の向上を目指す.

2 相対的意味論に基づく変換主導型統計機械翻訳 (TDSMT: Transfer Driven Statistical Machine Translation)[1] の概要

TDSMT の概要を日英翻訳の例で示す. TDSMT は学習文対と変換テーブルを利用して翻訳を行う手法である. 変換テーブルは, 「“ *A(*学習文対内の日本語句*)* が *B(*学習文対内の英語句*)* ならば *C(*入力文内の日本語句*)* は *D(*出力文内の英語句*)* ”」という形式をもつ.

2.1 変換テーブルの作成手順

具体例として, 「私 の 父 は 医者 だ 。」「私 の 母 は 日本語教師 だ 。」という 2 つの学習文対から変換テーブルを作成する例を示す.

手順 1 対訳単語の作成
学習文対と対訳単語確率 *(*IBM model 1[2]) を利用して対訳単語を作成する.

表 1 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

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 2 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
単語レベル文パターン	日本語側	*X1* の *X2* は *X3* だ 。
	英語側	*X1 X2* is a *X3* .

手順 3 変換テーブルの作成
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テーブルとする.

表 3 変換テーブルの作成

文パターン 原文	日本語側	私 の 父 は 医者 だ 。	
	英語側	My father is a doctor .	
単語レベル 文パターン	日本語側	*X1* の *X2* は *X3* だ 。	
	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 の変換テーブル	*A・B*	*A:* 医者	*B: doctor*
	C・D	*C:* 日本語教師	*D: Japanese teacher*

手順 4 変換テーブルへの確率の付与
変換テーブルの *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.

2.2 翻訳の手順

以下に具体例として, 「私 の 父 は 医者 だ 。」という入力文を翻訳する流れを示す.

手順 1 学習文対の日本語側への変換テーブルの適用
変換テーブルの *A* と *C* を利用して, 学習文対の日本語側と入力文を一致させる.

手順 2 学習文対の英語側への変換テーブルの適用
手順 1 と同様に変換テーブルの *B* と *D* を学習文対の英語側に適用し, 出力文を作成する.

手順 3 最終的な出力文の決定
複数の出力文候補が得られた場合, 変換テーブルの確率と言語モデルにより出力文を決定する.

図 1 に TDSMT の翻訳の流れを示す.

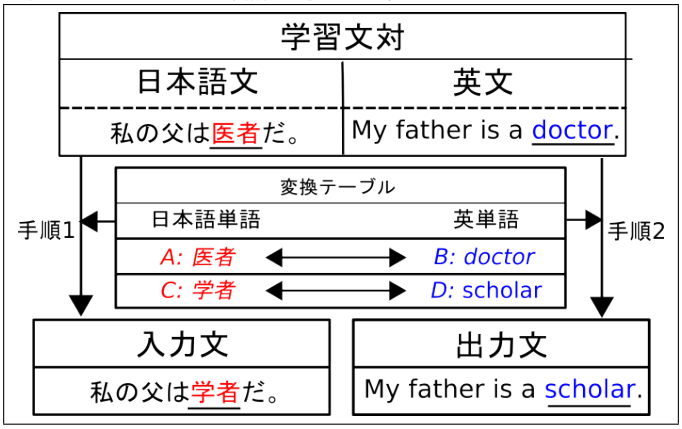

図 1 従来手法の翻訳の流れ

2.3 未知語処理

TDSMT の翻訳において A に対応する C が存在する変換テーブルが見つからない場合, 未出力となる. そこで従来手法 [1] では未知語が出現した際には「A が B なら C は C」と表現する未知語出力用変換テーブルを作成し, 変換テーブルに追加する. これにより入力文中の未知語を出力文中に原言語のまま出力することでカバー率が向上できる. 表 *4* に従来手法における未知語出力用変換テーブルの例を示す.

表 4　従来手法における未知語出力用変換テーブル

A	医者	*B*	doctor
C	漁師	*C*	漁師

2.4 問題点

従来手法では出力文中の未知語の翻訳は行われておらず, 入力文における未知語に相当する部分の意味が読み取れないという問題が存在する. よって本研究では NMT を用いて未知語の翻訳を行い, 翻訳精度の向上を目指す.

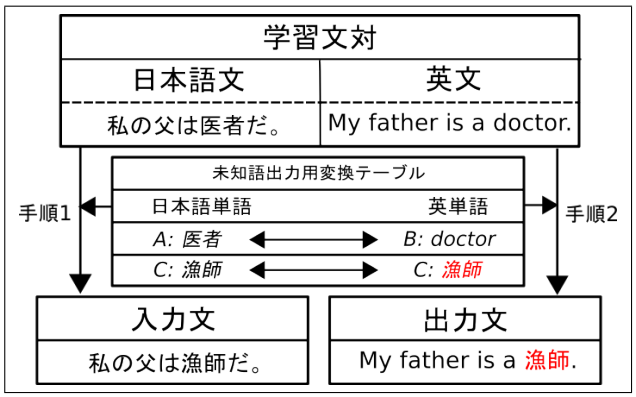

図 2　従来手法の翻訳の流れ

3 提案手法

本研究では TDSMT において, 未知語が出現した際に NMT を利用して変換テーブルを作成する手法を提案する. 具体的には, 未知語が出現した際にその単語に NMT を利用して翻訳を行い, その結果を変換テーブルの D(出力文内の英語句) に格納する. 翻訳時にこの未知語出力用変換テーブルを利用することで未知語が英語で出力されることが期待できる. 提案手法における翻訳の流れを図 3 に示す.

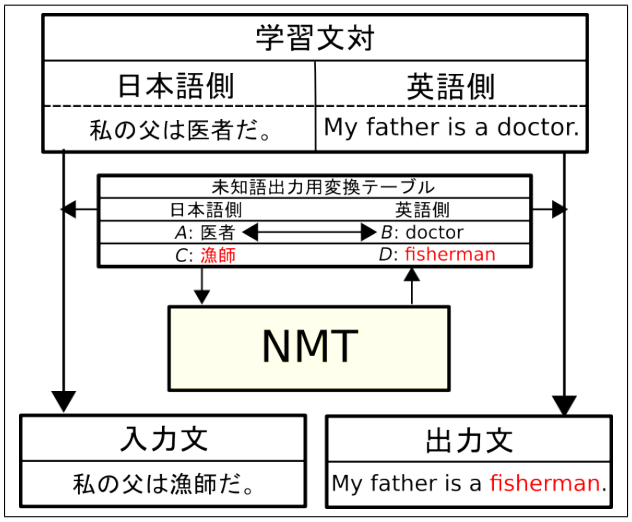

図 3　提案手法の翻訳の流れ

3.1 NMT による翻訳

手順 1　従来手法と同様に学習文対から対訳単語を作成する.

手順 2　NMT に手順 1 で作成した対訳単語を学習させ翻訳モデルを作成する.

手順 3　入力文を連続未知単語として分割する. 具体的には入力文を 1 単語, 2 単語連続, 3 単語連続ごとに分割する. 連続未知単語の例を表 5 に示す.

手順 4　手順 3 で作成した入力文の各連続未知単語を NMT を用いて翻訳を行う. 入力文の各連続未知単語の翻訳の例を表 5 に示す.

表 5　連続未知単語での分割の例

入力文	
私 の 父 は 漁師 だ 。	
連続未知単語	NMT の翻訳結果
私	I
私 の	my
私 の 父	My father
の	of
の 父	father
の 父 は	father is

3.2 翻訳テーブルの作成

手順 5　NMT の翻訳結果を変換テーブルの D にあたる部分に格納し未知語出力用変換テーブルを作成する. 作成したテーブルを変換時に使用するテーブルの候補に追加する. 翻訳テーブル作成の例を表 6 に示す.

表 6　未知語出力用変換テーブル

対訳単語	
医者	doctor
入力文	
私 の 父 は 漁師 だ 。	
未知語出力用変換テーブル	
A: 医者	*B*: doctor
C: 私	*D*: I
A: 医者	*B*: doctor
C: 私 の	*D*: My
A: 医者	*B*: doctor
C: 私 の 父	*D*: My father
etc...	

手順 6　変換テーブルを使用して翻訳を行う.

図 4 に未知語処理を行い翻訳を行う流れを示す　.

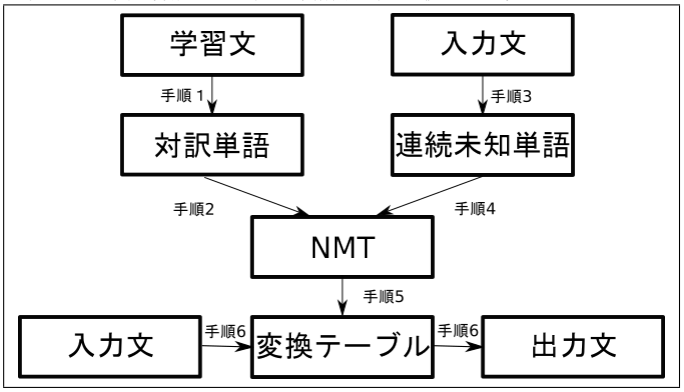

図 4　出力の流れ

4 実験設定

4.1 実験目的

従来手法と提案手法の比較実験により, 提案手法の有効性を検証する.

4.2 提案手法の翻訳性能評価

従来手法と提案手法の精度評価として, 以下の項目を調査し比較する.

- 出力を得ることのできた入力文の数, 及びその内の原言語の単語を含まないものの数
- 翻訳精度 (自動評価)
- 翻訳精度 (人手評価)

4.3 実験データ

使用するデータの内訳を表 7 に示す. また, NMT の翻訳モデルには Encoder-Decoder モデルを使用する [3].

表 7　実験データ

	TDSMT の学習	NMT の学習
学習文対	159,998 文	対訳単語 1,065,615 対
ディベロップ文	/	対訳単語 10,871 対
テスト文	100 文	

5 実験結果

5.1 カバー率

表 8 に出力文を得ることのできた入力文の数と, その内の日本語単語を含む文の数を示す.

表 8　カバー率

	入力文	出力文	日本語を含む出力文
従来手法	100	84	44
提案手法	100	84	11

5.2 翻訳精度 (自動評価)

従来手法と提案手法で翻訳実験を行う.　そして, 翻訳結果に対して自動評価により結果の比較を行う. 表 9 に自動評価の結果を示す. この結果より, 未知語処理に NMT を利用した提案手法が, 従来手法より優れた結果になった.

表 9　実験結果

	BLEU	RIBES	TER
従来手法	0.059	0.637	0.818
提案手法	0.076	0.677	0.756

5.3 翻訳精度 (人手評価)

従来手法と提案手法で翻訳実験を行う. そして, 翻訳結果に対して人手により結果の比較を行う.

○　提案手法の方が優れていた出力
△　提案手法と従来手法のどちらが優れているか一概に判断のできない出力
×　従来手法の方が優れている出力

評価結果を表 10 に示す.

表 10　人手評価の結果

○	△	×
14	57	12

6 考察

6.1 翻訳精度の自動評価結果

表 9 より, 提案手法では従来手法よりも高い精度となった.　特に BLEU は上昇幅が大きくなった. これは従来手法では日本語単語だった部分が提案手法によって英単語に翻訳されたことで単語の対応が取れるようになったためだと考えられる.

6.2 翻訳精度の人手評価

人手評価は, 提案手法では従来手法との精度の違いは無かった.(表 10)

6.2.1 提案手法の方が優れていた出力

提案手法の方が優れていた出力の例を表 11 に示す. この出力の流れを解析し, 考察する.

表 11　○の出力例

入力文	この 機会 に 徹底的 な 論議 を 期待 し たい 。
参照文	It is to be hoped that the panel members will take this opportunity to thoroughly debate the issue .
従来手法	This want hope 徹底的 な 論議 the in my opportunity .
提案手法	This want hope thorough discussion in my opportunity .

適応された文パターン及び変換に使用された学習文対を表 12 に示す. 従来手法では表 13 の変換テーブルによって未知語「徹底的な 議論」が日本語のまま出力されている. 提案手法では表 14 の変換テーブルによって英語で出力されている. したがって, 提案手法で追加されたテーブルによって出力の精度が向上した.

表 12　適用された文パターンと変換に使用された学習文対

学習文対	日本語側	私 は 胸 に 鋭い 痛み を 感じ た 。
	英語側	I felt a sharp pain in my chest .
文パターン	日本語側	*X1 X2* に *X3 X4 X5 X6* 。
	英語側	*X1 X6 X5 X3 X4* in my *X2* .

表 13　従来手法において適応された変換テーブル

X3			
A	鋭い	*B*	sharp
C	徹底的な 議論	*C*	徹底的な 議論
X4			
A	痛み	*B*	pain
C	を	*D*	the

表 14　提案手法において適応された変換テーブル

X3			
A	鋭い	*B*	sharp
C	徹底的な	*D*	thorough
X4			
A	痛み	*B*	pain
C	議論 を	*D*	discussion

6.2.2 提案手法と従来手法のどちらが優れているか一概に判断のできない出力

提案手法と従来手法のどちらが優れているか一概に判断のできない出力の例を表 15 に示す. 従来手法で未知語だった「あり 父親 だっ」は提案手法で「was father」へと翻訳されている. しかし,「献身 的 な 父」は従来手法で「dedication husbnd」という正しい出力が為されているが提案手法では,「unfortunate husband husband」という誤った出力をしている. よって△と評価した.

表 15　△の出力例

入力文	献身 的 な 夫 で あり 父親 だっ た 。
参照文	He was a devoted husband and father .
従来手法	He あり 父親 だっ a dedication husband .
提案手法	He was father unfortunate husband her husband .

6.2.3 従来手法の方が優れていた出力

従来手法の方が優れていた出力の例を表 16 に示す. この出力の流れを解析し, 考察する.

表 16 × の出力例

入力文	良心 が 彼女 を 苦しめ た 。
参照文	Her conscience stung her .
従来手法	She suffered from a conscience .
提案手法	She suffered from a worried her .

適用された文パターン及び変換に使用された学習文対を表 17 に示す. 従来手法では「良心」を表 18 の変換テーブルにより, 「conscience」という正しい翻訳が成されていた. しかし, 提案手法では, 表 19 の提案手法による未知語出力用変換テーブルを使用して翻訳を行なっている. その結果「worried her」という誤った出力がされている. つまり, 提案手法で追加したテーブルにより精度が低下している.

表 17 適用された文パターンと変換に使用された学習文対

学習文対	日本語側	頭痛 が 彼女 を 苦しめ た 。
	英語側	She suffered from a headache .
文パターン	日本語側	*X1* が 彼女 を 苦しめ た 。
	英語側	She suffered from *X1* .

表 18 従来手法において *X1* に適応された変換テーブル

A	頭痛	*B*	headache
C	良心	*D*	conscience

表 19 提案手法において *X1* に適応された変換テーブル

A	頭痛	*B*	headache
C	良心	*D*	worried her

6.3 未知語の出力

表 8 より, 日本語の単語を含む出力は従来手法では, 44 文であった. 提案手法では 11 文となった. このため出力文中での日本語単語の数は削減されたことが確認できる.

提案手法で日本語単語が出力された原因を考察する. 今回の実験では従来手法でも変換テーブルを作成し, それを追加している. そのため提案手法で作成した未知語出力用変換テーブルよりも従来手法で作成した未知語出力用変換テーブルの方が適用確率が高い場合, 出力文中に日本語単語が出力される. 従来手法でのテーブル作成を行わなければ, 出力文中に原言語は出現されなくなるがカバー率が低下すると考えられる.

6.4 従来手法の出力において, 未知語を含んでいた文章における提案手法の評価

入力文において, 従来手法で翻訳を行った際に出力文に日本語が含まれていた文に着目し, 本手法の未知語処理における有用性を検証する. 人手評価の結果を表 20 に示す. 表 21 に○と判断した出力の具体例を示す.

表 20 人手評価の結果

○	△	×
10	32	2

表 21 ○の出力例

入力文	彼 は 麻薬 所持 の 疑い で 検挙 さ れ た 。
参照文	He was arrested for having drugs .
従来手法	He was arrested on suspicion of 麻薬 所持 .
提案手法	He was arrested on suspicion of accepting drugs .

表 10 と比較すると, 評価 × の出力文は 2 文まで大きく減少している. したがって, 未知語処理における本手法の有用性が証明された. また, 本手法で正しい出力がされない文章は従来手法の出力で日本語を含まない文章に多いことがわかる. つまり, 未知語が存在しない文章の翻訳の際に本手法は精度が低くなる.

以上より, 事前に未知語出力用変換テーブルを作成するのではなく, 予め従来手法で翻訳を行った後, 翻訳できなかった部分のみを取り出して NMT によって翻訳を行うという手法ならば精度が向上すると考えられる.

7 おわりに

本研究では, TDSMT の未知語処理に NMT を使用した場合の精度について調査した. 実験により, 精度を向上させることはできなかったが, 出力中の日本語単語の出現を減らすことができた. 本研究の精度を向上させる手法として, 事前に未知語処理を行うのではなく, 一度従来手法で翻訳を行った後に, 出力文に出現した日本語の部分を NMT で翻訳する手法が考えられる.

参考文献

[1] 安場 裕人, 村上 仁一, “ 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 ” 言語処理学会第 25 回年次大会, A5-4, 2019.

[2] Peter F.Brown. Stephen A.Della Pietra, Vincent J.Della Pietra. Robert L.Mercer . The mathe-matics of statistical machinetranslation:Parameter Estimation. *Computational Linguistics. 1993* 言語処理学会第 25 回年次大会, P1-5, 2019.

[3] Sequence to Sequence Learning with Neural Networks Ilya Sutskever Google Oriol Vinyals Google Quoc V. Le Google NIPS 2014