

相対的意味論に基づく変換主導型パターンベース統計機械翻訳 (TDPBSMT) の提案

中村 勇太¹ 村上仁一²
鳥取大学 工学科 電気情報系学科
s152072@ike.tottori-u.ac.jp¹
murakami@eecs.tottori-u.ac.jp²

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 (以下, TDSMT)” が提案されている [1]. TDSMT は, 学習文対と変換テーブルを用いて, 原言語文を入力とし, 目的言語文を出力する手法である. 変換テーブルは “ A が B ならば C は D ” で表現する.

しかしこの手法で出力文を得るためには, 変換テーブルを適用した, 入力文が学習文対に完全に一致する必要がある. 従って, 入力文数に対して, 得られる出力文数が少ないという問題がある.

そこで本稿では, “相対的意味論に基づく変換主導型統計機械翻訳 (以下, TDPBSMT)” を提案する. この手法は, 変換テーブルを “ A が B ” と “ C が D ” の 2 つに分割する. 次に, “ A が B ” を利用して文パターンを作成する. そして, “ C が D ” を文パターンに適用する. 提案手法によって, 従来手法と比較して出力文数が向上する.

2 相対的意味論に基づく変換主導型統計機械翻訳 (TDSMT) [1]

“相対的意味論に基づく変換主導型統計機械翻訳” では, 相対的意味論に基づいて変換テーブルを作成する. そして, 変換テーブルと学習文対を利用して翻訳を行う.

2.1 TDSMT の手順

TDSMT の手順を示す. 手順は「学習」と「翻訳」の二部からなる.

2.1.1 学習の手順

TDSMT における学習は「変換テーブルの作成」のみである. 本節で作成手順を示す.

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. 例として, 表 1 に示す学習文対を使用して, 表 2 に示す対訳単語を作成する.

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.

表 2 作成される対訳単語

対訳単語 1	彼	His
対訳単語 2	弟	brother
対訳単語 3	学生	student

手順 2 単語レベル文パターンの作成

学習文対内で対訳単語に当たる部分を変数化し, 単語レベル文パターンを作成する. 例を表 3 に示す.

表 3 単語レベル文パターンの作成例

学習文対 (日本語側)	彼の兄は医者だ。
学習文対 (英語側)	His brother is a doctor.
単語レベル文パターン (日本語側)	$N0$ の $N1$ は $N2$ だ
単語レベル文パターン (英語側)	$N0$ $N1$ is a $N2$

手順 3 変換テーブルの作成

学習文対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を変換テーブルとする. 表 4 では変数 $N2$ の部分から変換テーブル “学生” が “student” ならば “教師” は “teacher” が得られる.

表 4 変換テーブルの作成例

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.
単語レベル文パターン (日本語側)	$N0$ の $N1$ は $N2$ だ。
単語レベル文パターン (英語側)	$N0$ $N1$ is a $N2$.
照合する学習文対 (日本語側)	私の母は教師だ。
照合する学習文対 (英語側)	My mother is a teacher.
変換テーブル ($N2$)	A: 学生 B: student C: 教師 D: teacher

手順 4 変換テーブルに確率を付与

変換テーブルの各単語が学習文対に出現する頻度を利用し, 確率を計算する.

2.1.2 翻訳の手順

本節で TDSMT における翻訳の手順を示す. 入力文を “私の姉は教師だ。” とする.

手順 1 入力文に日本語側の変換テーブルを適用

変換テーブルの C と A 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 5 では入力文中の “教師” を “生徒” に変換する.

表 5 日本語側変換テーブルの適用例

入力文	私の姉は教師だ。
変換テーブル: C	教師
変換テーブル: A	生徒
一致する学習文対 (日本語側)	私の姉は生徒だ。

手順 2 学習文対に英語側の変換テーブルを適用

手順 1 と同じ変換テーブルの B と D を学習文対の英語側に適用し, 出力候補文を作成する. 表 6 では学習文対中の “student” を “teacher” に変換している.

表 6 英語変換テーブルの適用例

一致した学習文対 (日本語側)	私の姉は生徒だ。
一致した学習文対 (英語側)	My sister is a student.
変換テーブル:B	student
変換テーブル:D	teacher
出力候補文	My sister is a teacher.

手順 3 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 の流れ図を示す。

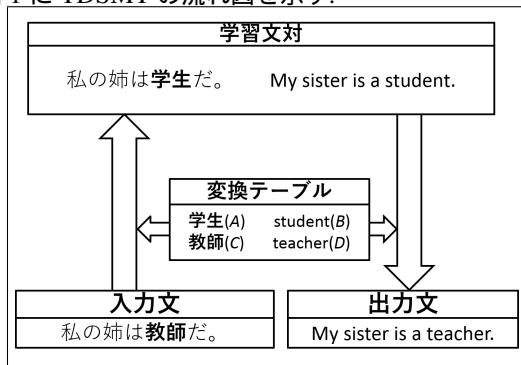


図 1 TDSMT の流れ図

2.2 問題点

従来手法の問題点は、入力文数に対して、出力文数が少ないことである。出力不可能な例として表 7, 8 があげられる。

表 7 出力不可能な例

入力文	その数値は小さくなった。
一致させたい学習文対 (日本語側)	その振幅は大きくなった。
一致させたい学習文対文対 (英語側)	The amplitude increased.
想定する出力文	The value reduced.

表 8 用意されている変換テーブル

	A	B	C	D
変換テーブル 1	振幅	amplitude	数値	value
変換テーブル 2	大きくな	increased	走る	run
変換テーブル 3	大きくな	grow	小さくな	reduced

上記の例が出力不可能となる流れを説明する。

1. 変換テーブル 1 を適用し入力文中の「数値」を「振幅」に変換する。
2. 変換テーブル 2 は「C: 走る」が学習文対の日本語側に存在しない。
3. 変換テーブル 3 は「B: grow」が学習文対の英語側に存在しない。

以上より、示した例は出力不可能である。本稿では、この問題を解決するために新しい手法を提案する。

3 相対的意味論に基づく変換主導型パターンベース統計機械翻訳 (TDPBSMT)

“相対的意味論に基づく変換主導型パターンベース統計機械翻訳”では文パターンを用いて翻訳を行う。この手法は、学習文対の代わりに文パターンを利用するので、出力文数が増加する。以下に TDPBSMT の手順を示す。手順の説明で用いる入力文と、作成する変換テーブルと、用意されている学習文対は、表 7, 8 と同一とする。

3.1 学習の手順

TDPBSMT における学習は「変換テーブルの作成」と「文パターンの作成」である。本節で作成手順を示す。

手順 1 変換テーブルを作成

TDSMT と同様の方法で変換テーブルを作成する。

手順 2 変換テーブルを分割

変換テーブルを変換テーブル AB と変換テーブル CD に分割する。

手順 3 文パターンを作成

学習文対の、変換テーブル AB に当たる単語を変数化し、文パターンを作成する。表 9 では 2 つの対訳単語を変数に置き換えている。

表 9 文パターンの作成例

	日本語側	英語側
学習文対	その振幅は大きくなった。	The amplitude increased.
変換テーブル $AB1$	A: 振幅	B: amplitude
変換テーブル $AB2$	A: 大きくな	B: increased
作成される文パターン	その $N0$ は $N1$ った。	The $N0$ $N1$.

3.2 翻訳の手順

本節で翻訳の手順を示す。

手順 1 入力文に日本語側の変換テーブルを適用

入力文中の変換テーブル CD に存在する語句を変数に変換し、文パターンに一致させる。表 10 に例を示す。

表 10 入力文への変換テーブル CD 適用例

入力文	その数値は小さくなった。
変換テーブル $CD1$	C: 数値
変換テーブル $CD3$	C: 小さくな
一致する文パターン (日本語側)	その $N0$ は $N1$ った。

手順 2 文パターンに変換テーブルの英語側を適用

文パターンの変数部に変換テーブル CD の英語側を代入し、出力候補文を作成する。表 11 に例を示す。

表 11 文パターンへの変換テーブル CD 適用例

一致した文パターン (日本語側)	その $N0$ は $N1$ った。
一致した文パターン (英語側)	The $N0$ $N1$.
変換テーブル $CD1$	D: value
変換テーブル $CD3$	D: reduced
出力候補文	The value reduced.

手順 3 最終的な出力文の決定

TDSMT と同様の方法で最終的な出力文を決定する。

図 2 に TDPBSMT の流れ図を示す。

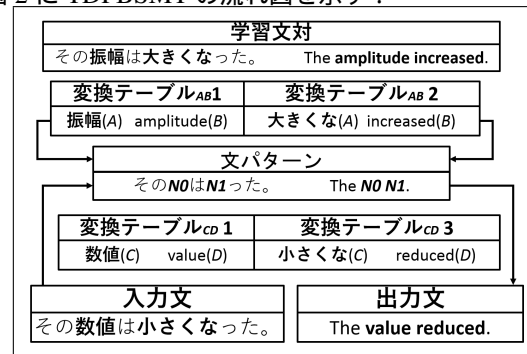


図 2 TDPBSMT の流れ図

また、実際の出力例を表 12 に示す。

表 12 TDPBSMT の出力例

入力文	視界は極めて良好であった。
文パターン (日本語側)	N02 は N00 N01 であった。
文パターン (英語側)	The N02 was N00 N01 .
変換テーブル $CD(N0)$	C : 極めて D : extremely
変換テーブル $CD(N1)$	C : 良好 D : good
変換テーブル $CD(N2)$	C : 視界 D : sight
出力文	The sight was extremely good .

4 実験

4.1 実験目的と方法

TDPBSMT と TDSMT を比較する。入力文数に対する翻訳可能な文の割合 (カバー率) と、翻訳精度を調査する。翻訳精度の調査は自動評価と人手評価で行う。

4.2 実験条件

4.2.1 実験に使用するデータ

本研究では、実験に電子辞書などの例文より抽出した単文コーパス [3] を使用する。使用するデータの内訳を表 13 に示す。

表 13 実験データの内訳

学習文対	160000 文
入力文	100 文

4.2.2 カバー率の調査の実験条件

TDPBSMT と TDSMT を用いて、入力文を 100 文として、得られた出力文数を調査する。

4.2.3 翻訳精度の調査の実験条件

TDPBSMT と TDSMT を用いて、得られた出力文を自動評価と人手評価で評価する。また、人手評価は対比較評価で行う。人手評価の基準は次の 4 つである。

- TDPBSMT : TDPBSMT の方が優れている。
- TDSMT : TDSMT の方が優れている。
- 差なし : それぞれの翻訳結果は異なるが、翻訳精度には差がない。
- 同一 : それぞれの翻訳結果が一致している。

4.3 実験結果

4.3.1 カバー率の調査の結果 (出力文数/入力文数)

カバー率の調査の結果を表 14 に示す。

表 14 カバー率調査結果

TDPBSMT	TDSMT
95/100	29/100

表 14 より TDPBSMT は TDSMT と比較して、カバー率が向上したことが分かる。

4.3.2 翻訳精度の調査の結果

4.3.2.1 自動評価の結果 (29 文)

TDSMT で出力を得られた 29 文を対象に、自動評価を行った結果を表 15 に示す。

表 15 自動評価結果 (29 文)

	BLEU	METEOR	TER
TDPBSMT	0.20	0.49	0.60
TDSMT	0.22	0.53	0.54

表 15 より TDPBSMT と TDSMT の翻訳精度にほぼ差はない。

4.3.2.2 人手評価の結果 (29 文)

TDSMT で出力を得られた 29 文を対象に、対比較評価を行った結果を表 16 に示す。

表 16 人手評価結果 (29 文)

TDPBSMT	5
TDSMT	5
差なし	13
同一	6

表 16 より TDPBSMT と TDSMT の翻訳精度に差はないことが分かる。表 17 から 19 は実際の出力例を示す。表 17 と 18 の出力文中の下線部は、人手評価の判断基準とした部分である。

表 17 TDPBSMT とした例

入力文	良心が彼女を苦しめた。
TDPBSMT	<u>She</u> suffered from a conscience.
TDSMT	My conscience worried her.
参照文	Her conscience stung her.

TDPBSMT と TDSMT の出力文を比較すると、TDPBSMT の主語が「She」で、TDSMT の主語が「My conscience」である。従って、入力文の主語との整合性において、TDPBSMT が TDSMT より優れていると判断した。

表 18 TDSMT とした例

入力文	友達とハイキングに行きました。
TDPBSMT	I went on hiking friends.
TDSMT	I went hiking <u>with</u> friends.
参照文	I went hiking with some friends.

TDPBSMT と TDSMT の出力文を比較すると、TDPBSMT は「with」が存在せず、TDSMT は存在している。従って、適切な前置詞の有無において、TDSMT が TDPBSMT より優れていると判断した。

表 19 差なしとした例

入力文	靴のひもがほどけた。
TDPBSMT	My shoelaces came undone.
TDSMT	The shoelace came untied.
参照文	My shoelace has come undone.

4.4 実験結果のまとめ

実験結果から TDPBSMT と TDSMT の翻訳精度に差はなかった。一方、TDPBSMT のカバー率は、TDSMT と比較して向上している。よって提案手法の有効性が示された。

5 考察

5.1 TDPBSMT と Moses の翻訳精度の比較

本節で、提案した TDPBSMT と、一般に利用される句に基づく機械翻訳 (Moses)[4] の翻訳精度の比較を行う。

5.1.1 使用したデータ

使用したデータの内訳を表 20 に示す。学習文対と入力文は、表 13 のデータと同一である。

表 20 実験データの内訳

学習文対	160000 文
ディベロップメント文	1000 文
入力文	100 文

5.1.2 実験結果

TDPBSMT と Moses の実験結果を、表 21 と表 22 に示す。評価に用いる文数は TDPBSMT で出力が得られた 95 文である。

表 21 TDPBSMT と Moses の自動評価 (95 文)

	BLEU	METEOR	TER
TDPBSMT	0.20	0.49	0.60
Moses	0.17	0.51	0.60

表 22 TDPBSMT と Moses の人手評価結果 (95 文)

TDPBSMT	15
Moses	14
差なし	58
同一	8

表 23 から 25 に実際の出力例を示す。

表 23 TDPBSMT とした例

入力文	あの一件は無事に済んだ。
TDPBSMT	The affair is over safely .
Moses	The safely .
参照文	The affair was settled without mishap .

表 24 Moses とした例

入力文	その少年はパイロットを夢見ている。
TDPBSMT	The boy has a 夢見 pilot .
Moses	The boy dreams of becoming a pilot .
参照文	The boy is dreaming of becoming a pilot .

表 25 差なしとした例

入力文	公園は川まで広がっている。
TDPBSMT	The park is spread to the river.
Moses	The park is extended as far as the river.
参照文	The park reaches to the river.

表 21 と表 22 の結果より, TDPBSMT と Moses の翻訳精度に差はなかった。

5.2 提案手法の問題点

本節では提案手法の誤り解析を行う。

5.2.1 翻訳確率について

TDPBSMT では複数の出力候補文が得られた際に, 各種確率を用いて最終的な出力文を決定する。その計算式を以下に示す [5]。

$$\log P = \log P_v + \log P_p + \log P_m \quad (1)$$

P : 翻訳確率

P_v : 学習文対中の単語の出現回数に基づく確率

P_p : 文パターンに付与された確率

P_m : 言語モデルで生成された確率

5.2.2 誤出力の結果

解析する提案手法の誤出力の結果を表 26 に示す。また, 翻訳確率を表 27 に示す。この出力結果は複数作成された出力候補文の中で翻訳確率が第 1 位である。

表 26 提案手法の誤出力の結果

入力文	その翌日のバスの切符を買った。
参照文	He booked himself for the following day's bus.
文パターン (日本語側)	$N03 N00 N04 N05 N01 N02$ た。
文パターン (英語側)	$N03 N02 N01 N04 N00 N05$.
変換テーブル $CD(N0)$	C : 翌日 D : next
変換テーブル $CD(N1)$	C : の切符 D : ticket for
変換テーブル $CD(N2)$	C : を買った D : bought a
変換テーブル $CD(N3)$	C : その D : I
変換テーブル $CD(N4)$	C : の D : the
変換テーブル $CD(N5)$	C : バス D : bus
出力文	I bought a ticket for the next bus.

表 27 出力文の翻訳確率

出力候補文	I bought a ticket for the next bus.
$\log P_v$	-48.1586
$\log P_p$	-3.5850
$\log P_m$	-2042.5963
$\log P$	-2094.3399

表 26 の出力文からは「翌日」を表す英語句が存在しない。また, 表 27 を見ると, 言語モデルの確率 ($\log P_m$) が, 他の 2 つの確率 ($\log P_v$ と $\log P_p$) と比較して非常に小さい。

5.2.3 比較的正しい出力候補文の結果

比較的正しい出力候補文「I bought a ticket for bus in the next day.」が作成されていた。また, 複数作成された翻訳候補文の中で翻訳確率が第 3 位である。文の作成結果を表 28 に示す。また, 翻訳確率を表 29 に示す。

表 28 出力候補文の作成結果

文パターン (日本語側)	その $N02 N00 N04 N01 N03$ た。
文パターン (英語側)	$I N03 N01 N04 N00$ the $N02$.
変換テーブル $CD(N0)$	C : の D : in
変換テーブル $CD(N1)$	C : の切符 D : ticket for
変換テーブル $CD(N2)$	C : 翌日 D : next day
変換テーブル $CD(N3)$	C : を買った D : bought a
変換テーブル $CD(N4)$	C : バス D : bus
出力候補文	I bought a ticket for bus in the next day.

表 29 出力候補文の翻訳確率

出力候補文	I bought a ticket for bus in the next day.
$\log P_v$	-27.9219
$\log P_p$	-3.9069
$\log P_m$	-2290.6281
$\log P$	-2322.4569

表 28 の出力候補文は翻訳確率 ($\log P$) が表 26 の出力文と比較して小さかったため, 出力文として選択されなかった。また, P_v は比較的正しい出力候補文 (表 29) の方が, 誤出力文 (表 27) より大きい。

5.2.4 解析

表 27 と 29 から以下の 2 点に分かる。

- P_m は P_v と P_p より非常に小さい。
- 比較的正しい出力候補文の P_v の方が, 誤った出力文の P_v より大きい。

従って, 比較的正しい出力候補文を選択する方法として 2 つの方法がある。

1. 別の種類の言語モデルを使用する。
2. P_v , P_p , P_m に重みを付ける。

今後, この 2 つの方法に取り組んでいきたい。

6 おわりに

従来手法のカバー率の向上を目的として「相対的意味論に基づく変換主導型パターンベース統計機械翻訳」を提案した。実験により, 提案手法の有効性が示せた。

参考文献

- [1] 安場裕人, 村上仁一. “変換主導型翻訳の提案” 自然言語処理学会第 24 年次大会, 2018 年 3 月
- [2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. “The mathematics of statistical machine translation: Parameter Estimation”, Computational Linguistics, 1993.
- [3] 村上仁一, 藤波達. “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] Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Corbett Moran, Evan Herbst “Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding”, 2006 Language Engineering Workshop, September 3, 2007
- [5] 松本大輝, 村上仁一. “翻訳における分野依存性を軽減する言語モデルの調査” 自然言語処理学会第 25 年次大会, 2019 年 3 月 (予定)