類似度を利用した変換テーブルの精度向上

森本 世人
鳥取大学 工学部
b16t2117c@edu.tottori-u.ac.jp

村上 仁一
鳥取大学 工学部
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いた変換主導型統計機械翻訳 [1](Transfar Driven Statistical Machine Translation：以下，TDSMT) が提案されている．TDSMT では変換テーブルを用いて翻訳を行う．変換テーブルとは「A が B ならば C は D である」という A，B，C，D の相対性に基づいて関係を定義したテーブルである．ここで A と B は単語である．また，C と D は単語もしくは句である．

変換テーブルを自動的に作成する際に, 誤った変換テーブルを作成してしまうという問題点が存在する．そこで，従来手法では誤った変換テーブルを削除するために変換テーブルを生成した後に，枝刈りを行う．従来行われている枝刈りの手法は，A と B，C と D の対訳単語確率を用いた枝刈りである．しかし，枝刈りの精度は未だ不十分である．

本研究では，従来手法で行う枝刈りに加えて，A と C, B と D の類似度を用いて枝刈りを行うことを提案する．類似度とは注目単語の前後環境がどれだけ一致しているかと定義する．提案手法を用いることで変換テーブルの精度を向上できると考える [2].

2 変換テーブル作成方法

変換テーブルは対訳学習文からパターンを用いて自動作成する．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
対訳文 (1) と対訳単語確率 (IBM model 1[3]) 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 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, パターンを作成する.

手順 3 変換テーブルの作成
対訳文 (2) とパターンを照合する. パターンにおいて変数が一致した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	猫		
	英語	cat		
学習文対 (1)	日本語	私 は 猫 が 好き だ 。		
	英語	I like a cat .		
パターン (手順 2)	日本語	私 は X が 好き だ 。		
	英語	I like a X .		
学習文対 (2)	日本語	私 は 犬 が 好き だ 。		
	英語	I like a dog .		
変換テーブル (手順 3)	A	猫	B	cat
	C	犬	D	dog

3 従来手法

3.1 対訳単語確率 [4]

対訳単語確率とは入力言語の単語が出力言語の単語に訳される確率を示している．対訳単語確率が高いほどもっともらしい訳であると考えられる．値は IBM model 1 によって求める．

日本語単語 A が英語単語 B に訳される対訳単語確率を *prob(AB)* とする．入力言語と出力言語を入れ替えた対訳単語確率 *prob(BA)* も同様に計算する．

3.2 対訳単語確率を用いた枝刈り

以下に手順を示す.

手順 1 変換テーブルより *prob(AB)*, *prob(BA)*, *prob(CD)*, *prob(DC)* の値を計算する．
手順 2 単語ごとにそれぞれ，順位付けする．
手順 3 任意の順位を閾値として枝刈りを行う．

従来手法の問題点として，誤った変換テーブルが多く残る点がある．

4 提案手法

4.1 類似度

誤った変換テーブルは A と C や B と D の置き換え可能性が低いと考える．A と C や B と D の類似度で枝刈りをすると精度の向上が見込める．類似度が高い単語らは，置き換え可能性が高いと仮定する．本研究では，類似度とは前後単語の環境と定義する．よって注目単語における前後単語の一致度が高いほど類似度が高いと考える．

4.2 類似度を用いた枝刈り

以下に手順を示す．類似度の値を *sim()* とする．

手順 1 従来手法の枝刈りを行う
手順 2 変換テーブルより *sim(AC)*, *sim(BD)* の値を計算する．
手順 3 任意の値を閾値として枝刈りを行う．

4.3 類似度の計算

二つの日本語単語 A と C の類似度の値を $sim(AC)$ とすると $sim(AC)$ は以下の式 4.3 で計算する．

$$sim(AC) = \log_2\{\frac{count(A_{context} \cap C_{context})}{count(A_{context})} \times \frac{count(A_{context} \cap C_{context})}{count(C_{context})}\}$$

count(X):集合 X の単語の総数
$A_{context}$,$C_{context}$:単語 A と C の前後単語の集合

上記の式は One-hot の word2vec に類似している．なお類似度の計算においては 2 単語連続を 1 単語として用いる．式 4.3 を英語単語 B と D でも同様の計算を行い $sim(BD)$ とする．

5 実験条件

5.1 実験データ

電子辞書などの例文より抽出した単文コーパス [5] を用いる．使用するデータは対訳学習文約 160,000 文である．学習文より生成された枝刈り前の変換テーブルの数は約 2,064 万個である．枝刈り前の変換テーブルの精度は誤り率 76% である．

5.2 評価方法

枝刈り後の変換テーブルをランダムに 100 個抽出する．抽出した変換テーブルに対して人手で正解か否かで 2 値評価する．変換テーブルの評価は置き換え可能性の評価が難しい．よって，翻訳に用いる際最も重要となる C と D の対訳関係を評価する．

6 実験

6 章の実験の目的は，変換テーブルの精度向上である．そして，提案手法と従来手法を比較する．

6.1 閾値

実験 1 で用いた閾値を表 2 に示す．

表 2 実験で用いた閾値

	対訳単語確率	類似度
従来手法	16 位	用いない
提案手法	16 位	-39.0

6.2 実験結果

評価結果を表 3 に示す．

表 3 従来手法と提案手法の誤り率

	提案手法	従来手法
変換テーブルの数	701,828	3,698,524
誤り率	3%	28%

表 5.2 の結果より，提案手法による変換テーブルの精度向上が確認できた．提案手法により，誤り率が約 28% から 3% に向上する．一方で変換テーブルの数が約 81% 減少した．

6.3 出力例

提案手法の出力例を表 4 と表 5 に示す．表 4 は正解と評価される変換テーブルである．表 5 は誤りと評価される変換テーブルである．

従来手法では出力し，提案手法では出力できなかった例を表 6 と表 7 に示す．表 6 は誤りと評価される変換テーブルである．表 7 は正解と評価される変換テーブルである．

表 4 提案手法における正解の出力

A	ドア	B	door
C	犬の首輪	D	collar of the dog
A 原文	ドアが外れた。		
B 原文	The door has got unhinged.		
C 原文	犬の首輪が外れた。		
D 原文	The collar of the dog has got loose.		
AB(順位)	1	BA(順位)	1
CD(順位)	2	DC(順位)	2
AC	-17.5	BD	-23.0

表 5 提案手法における誤りの出力

A	声	B	voice
C	鼻水	D	nose
A 原文	彼は声が詰まった。		
B 原文	His voice choked.		
C 原文	彼は鼻水が出た。		
D 原文	His nose watered.		
AB(順位)	1	BA(順位)	1
CD(順位)	1	DC(順位)	7
AC	-23.2	BD	-25.3

表 6 提案手法における変換テーブルの改善例

A	英語	B	English
C	あひるは水かき	D	its webbed feet
A 原文	英語の勉強を怠けている。		
B 原文	He is lazy in the study of English.		
C 原文	あひるは水かきで水を掻いている。		
D 原文	The duck is paddling in the water with its webbed feet.		
AB(順位)	1	BA(順位)	1
CD(順位)	2	DC(順位)	2
AC	-47.2	BD	-47.9

表 7 提案手法における変換テーブルの改悪例

A	サングラス	B	sanglasses
C	腱	D	tendon
A 原文	サングラスをかけた。		
B 原文	She put on sunglasses.		
C 原文	腱を痛めた。		
D 原文	I've pulled a tendon.		
AB(順位)	1	BA(順位)	1
CD(順位)	1	DC(順位)	1
AC	-44.5	BD	-43.9

提案手法は間違った変換テーブルを出力から削除できる (表 6)．一方で，正解の変換テーブルも削除している (表 7)．

7 考察

7.1 単語と句に分けた調査

枝刈りした変換テーブルにおける単語と句の内訳を調査する．方法として，6 章で枝刈りした変換テーブルの C と D に当たる部分が単語-単語，単語-句，句-単語，句-句でグループ分けする．

7.1.1 枝刈り前の変換テーブル

枝刈り前の変換テーブルにおけるグループごとの変換テーブルの数を表 8 に示す．

7.1.2 枝刈り後のテーブルの調査

グループごとの変換テーブルの数を表 9 に示す．

表 8 グループごとの変換テーブルの数

単語-単語	9,443,152
単語-句	1,960,034
句-単語	4,126,333
句-句	5,115,214

表 9 グループごとの変換テーブルの数

	提案手法	従来手法
単語-単語	597,256	1,969,265
単語-句	25,787	147,918
句-単語	49,616	239,873
句-句	29,169	1,341,468

次に，表 9 の各グループを評価した．グループごとにランダムに 100 個ずつ抽出して評価した．評価結果を表 10 に示す．

表 10 誤り率

	提案手法	従来手法
単語→単語	2%	9%
単語→句	3%	10%
句→単語	2%	5%
句→句	11%	38%

表 10 より，句-句のグループでは誤り率が 38% から 11% に向上した．

7.1.3 出力例

7.1 節における句の出力例を表 11 と表 12 に示す．表 11 は正解と評価した例である．表 12 は誤りと評価した例である．

表 11 提案手法における句の正解例

A	使える	*B*	useful
C	我が校の名誉である	*D*	an honor to our school
A 原文	彼は使える。		
B 原文	He is useful.		
C 原文	彼は我が校の名誉である。		
D 原文	He is an honor to our school.		

表 12 提案手法における句の誤り例

A	を	*B*	the
C	を大規模に	*D*	dialects on a large
A 原文	頭に一撃を受けた。		
B 原文	I received a blow on the head.		
C 原文	方言の調査を大規模に行なった。		
D 原文	We made a survey of dialects on a large scale.		

7.2 変換テーブルの評価 (*A* と C，*B* と *D* の置き換え可能性)

今回は簡易的に評価を行うために *C* と D の部分の訳が適切であるかで評価を行った．しかし変換テーブルの構成を考えると *A* と B の訳の適切さや *A* と C，*B* と *D* の置き換え可能性も考慮に入れて評価する．そこで，6 章で用いた変換テーブルに注目して評価を行う．

7.2.1 評価結果

評価結果を表 13 に示す．

表 13 従来手法と提案手法の誤り率

	提案手法	従来手法
誤り率	3%	34%

7.2.2 評価例

7.2 章で評価した変換テーブルの例を示す．表 14 は正解と評価した例である．

表 14 評価例

A	少し	*B*	little
C	調査	*D*	reseach
A 原文	その会社は調査部を設けた。		
B 原文	The firm has instituted a research department.		
C 原文	英語は少し話せます。		
D 原文	I speak a little English.		

表 14 において対訳関係は正しいと考える．ここで置き換え可能性を考える．「少し」は副詞であり，「調査」は名詞である．副詞は後ろに何を伴っても良い．よって *A* と *C* は置き換え可能性があると考えられる．

表 15 は正解と評価した例である．

表 15 評価例

A	限り	*B*	limit
C	では	*D*	at
A 原文	地下資源には限りがある。		
B 原文	There is a limit to underground resources.		
C 原文	成功へはめったに一足飛びでは届かない。		
D 原文	Success is rarely reached at a single leap.		

表 14 と同様に表 15 も対訳関係は正しいと考える．同様に，置き換え可能性を考える．「で は」は主に接続助詞と係助詞の連語であり，「限り」は名詞である．接続助詞の前単語は名詞を取る．名詞は複合名詞の形で名詞につながる可能性がある．よって *A* と *C* は置き換え可能性があると考える．

表 14 と表 15 で示した例のように品詞が異なった変換テーブルの置き換え可能性を評価するのが難しい．

7.3 パターンについて

変換テーブルの作成において，複数の変数によるパターンが用いられている．表 16 と 17 に変換テーブルのパターンの例を示す．

表 16 は正解と評価される変換テーブルである．

表 16 パターン例

A	落ちる	*B*	fall
C	犬を激しくひっかいた	*D*	scratched wildly at the dog
A 原文	葉 | が | 落ちる。		
B 原文	The |leaves |fall.		
C 原文	猫 | が | 犬を激しくひっかいた。		
D 原文	The |cat |scratched wildly at the dog.		
日本語パターン	X2 X1 X3		
英語パターン	X1 X2 X3		

表 16 では「が」と「The」のように訳されない単語を仮定した上で，SV の形でパターンが生成され上手く対応されている．(ここで *D* の原文は SVO 形ではあるが V と O を一つの句として扱っている)

表 17 は誤りと評価される変換テーブルである．

表 17 と同様に *D* の原文が命令形になっているとパターンが誤ってしまう例が多い．

7.4 変換テーブルの数を増加させた実験

7.4 節では，変換テーブルの数を提案手法と従来手法で同程度に揃えた上で精度を調査する．そして，提案手法と従来手法を比較する．

表 17 パターン例

A	本	*B*	book
C	犬をひも	*D*	a leash
A 原文	本 | を | 買っ | た。		
B 原文	I |bought |a |book.		
C 原文	犬をひも | に | つないで | ください。		
D 原文	Put |your dog |on |a leash.		
日本語パターン	X4 X2 X3 X1		
英語パターン	X1 X3 X2 X4		

7.4.1 閾値

7.4 の実験では求められた類似度を昇順に順位付けする．さらに，類似度の順位を閾値とする．実験で用いた閾値を表 18 に示す．

表 18 実験で用いた閾値

	対訳単語確率	類似度
従来手法	16 位	用いない
提案手法	512 位	2048 位

7.4.2 実験結果

評価結果を表 19 に示す．

表 19 従来手法と提案手法の誤り率

	提案手法	従来手法
変換テーブルの数	2,993,145	3,698,524
誤り率	20%	28%

表 19 の結果より，変換テーブルの数を同程度にした際も，提案手法による変換テーブルの精度向上が確認できた．提案手法を用いることにより誤り率が 28% から 20% に向上する．

7.4.3 出力例

従来手法では枝刈りできず，提案手法では枝刈りできた変換テーブルの例を表 20 と 21 に示す．表 20 と 21 は誤りと評価される変換テーブルである

表 20 提案手法による変換テーブルの改善例

A	炒っ	*B*	roasted
C	し	*D*	with
A 原文	豆をこんがりと炒った。		
B 原文	He roasted the beans brown .		
C 原文	平手でその子を強く打とうとした。		
D 原文	He swiped at the child with his open hand .		
AB(順位)	1	*BA*(順位)	1
CD(順位)	14	*DC*(順位)	11
AC(順位)	91,907	*BD*(順位)	38,529

表 21 提案手法における変換テーブルの改善例

A	あけ	*B*	Open
C	し	*D*	He
A 原文	口 を 大きく あけ なさい 。		
B 原文	Open your mouth wide .		
C 原文	会議 を 中座 し た 。		
D 原文	He left the meeting halfway through .		
AB(順位)	2	*BA*(順位)	4
CD(順位)	2	*DC*(順位)	10
AC(順位)	67522	*BD*(順位)	51169

この節の例については 7.5 章で考察する

7.5 形態素解析の問題

7.4 章で得た出力例を考察する．表 20 と表 21 はいずれも誤りと評価した例である．

例のように，*C* が「し」となる変換テーブルは従来手法で枝刈りをした場合 6482 個存在する．一方提案手法で枝刈りをすると 44 個となる．中身として従来手法では「し」の訳として「with」や「to」などが多くある. 提案手法では「did」や「has」,「made」などが列挙される．

日本語において「し」は動詞に付属して「〜 し た」という 3 単語の形で用いられることが多い．しかし，英語では「〜 し た」を 1 単語の動詞の過去形で訳される場合が多い．また，英語文において「し」の対訳単語はない場合も多い．よって「し」の様な単語において間違った変換テーブルを作成してしまう場合がある．提案手法では *A* と *C* の類似度を用いて枝刈りを行うので多くの間違った変換テーブルを削除できると考える．

7.6 閾値の選択

6 章では閾値を類似度の数値とした．7.4 章では閾値を類似度の順位とした．表 3 と表 19 より閾値を類似度の数値とした実験がより高い精度を示した．類似度の数値を閾値としたことで改善できる例を表 22 に示す．表 22 は誤りと評価される例である．

表 22 誤りが削除された例

A	頭	*B*	head
C	そのオペラは	*D*	every listener off
A 原文	彼は食料を仲間と分けた。		
B 原文	The opera carried every listener off his feet.		
C 原文	そのオペラは聴衆を熱狂させた。		
D 原文	He shared his food with his friends.		
AB(順位)	1	*BA*(順位)	1
CD(順位)	2	*DC*(順位)	1
AC	-48.1	*BD*	-79.7
AC(順位)	3	*BD*(順位)	1

一方，閾値を類似度の順位としたほうが良い例を 23 に示す．表 23 は正解と評価される例である．

表 23 正解が削除された例

A	蒸気	*B*	steam
C	波打ち際	*D*	shoreline
A 原文	電気が蒸気に代わった。		
B 原文	Electricity has replaced steam.		
C 原文	波がひたひたと波打ち際に打ち寄せた。		
D 原文	Waves lapped the shoreline.		
AB(順位)	1	*BA*(順位)	1
CD(順位)	1	*DC*(順位)	2
AC	-79.7	*BD*	-46.4
AC(順位)	3	*BD*(順位)	75

表 22 において「listener」という表現がある．「listener」という表現は学習文内で 6 度しか出現しない表現である．表 23 において「shoreline」という表現がある．「shoreline」という表現は学習文内で 3 度しか出現しない表現である．学習文内での出現頻度が低頻度な表現は類似度の順位が高さと類似度の値が高さが一致しない．

8 おわりに

本研究では，変換テーブルの精度向上を目的とする手法を提案した. 提案手法は, 変換テーブルの同言語間の類似度を用いて枝刈りする手法である．実験結果より, 変換テーブル数の誤り率が 28% から 3% に改善できた．また句-句のグループでも 38% から 11% に改善できた．よって，提案手法の有効性が確認できた．しかし，変換テーブルの数が 3,698,524 個から 701,828 個に減少する．

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， 2018 年 3 月

[2] 中島 綜太，村上 仁一 “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”，言語処理学会第 26 回年次大会, P4-36, 2020.

[3] 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*.

[4] 興梠 玲架，村上 仁一 “パターンに基づく統計機械翻訳において変数部の総和を使った対訳句の抽出”，卒業論文，2016 年 3 月

[5] 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.