対訳単語の対偶を考慮した文パターンの選択と NMT の効果

村上仁一
鳥取大学工学部
murakami@tottori-u.ac.jp

概要

パターン翻訳は，古典的な機械翻訳の方法である．従来，文パターンは，人手で作成される．しかし，簡単な文パターンは，単語辞書を利用して，自動的に作成可能である．そして，NMT において翻訳精度を向上させるため，学習データに文パターンを加える方法がある．しかし，翻訳精度は，ほとんど向上しない．この原因として，文パターンの作成方法に問題があると考えた．そこで，対訳単語の対偶を考慮して文パターンを選択して，学習データに加えた．作成した NMT の翻訳精度を評価したところ，文パターンの選択前と比較すると大幅に向上した．

1 はじめに

パターン翻訳は，古典的な機械翻訳の方法である．従来，文パターンは，人手で作成される．しかし，簡単な文パターンは，対訳単語辞書を利用して，自動的に作成可能である．そして，NMT の翻訳精度を向上させるために，学習データに文パターンを加える方法がある．しかし，翻訳精度は，ほとんど向上しない．この原因として，文パターンの作成方法に問題があると考えた．

文パターンは，対訳単語から見ると，基本的に対偶の関係にある．しかし，１つのソース側の単語がターゲット側の複数の単語に依存する場合，この関係が崩れる．そこで，対訳単語と文パターンの対偶の関係を維持するために，１つのソース側のパターンが，複数のターゲット側のパターンにならないように，制約をかける．この効果を調査する．

2 パターン翻訳

2.1 パターン翻訳の例

パターン翻訳は，古典的な翻訳方法である．そして，定型文ならば翻訳精度が高い．そのため，定型文が多い分野で，利用されてきた．例として，カナダの天気予報の英仏翻訳がある．しかし，定型的な文では翻訳精度が高いが，汎用的な文では，多くの問題が生じる．その問題の１つが，文パターンの作成方法である．従来，文パターンは，人手で作成される．このためコストが高い．しかし，簡単な文パターンならば，単語辞書を利用して，自動的に作成可能である．例を表 1 にあげる．

表 1 パターン翻訳の例

対訳単語 山 mountain
対訳文 彼は山に行く． He go to the mountain
文パターン 彼は X に行く． He go to the X
テスト文 彼は海に行く．
対訳単語 海 sea
出力文 He go to the sea

ある程度，単純で定型的な文であれば，この方法で高い翻訳精度が得られる．

2.2 パターン翻訳の問題点

パターン翻訳には，様々な問題がある．その１つが要素合成法の問題として知られている．１つの単語の訳が，全体の誤訳に繋がる．例を表 2 にあげる．

表 2 に示したように，1 つの日本文パターンに２つの英文パターンが対応するとき，翻訳誤りになることがわかる．通常，文パターンは，人手で作成される．人手で作成するとき，上記のような問題が生じないように文パターンを作成する．しかし，この作業は，人手の負荷が大きい．

表 2　要素合成法の問題）

対訳単語 彼女 She 彼女 her
対訳文 彼女は我を通した．　She had her own way.
文パターン X は我を通した．　X had her own way. X は我を通した．　She had X own way.
テスト文 彼は我を通した．
対訳単語 彼 He 彼 his
出力文（誤訳） He had her own way （彼は彼女の我儘を聞いた） She had his own way （彼女は彼の我儘を聞いた）

2.3　対訳単語の対偶と文パターン

本節では，まず，対訳単語の対偶と文パターンの関係を考察する．

1. 対訳単語
   対訳単語 “ 山:mountain”　を想定する．
2. 対訳単語の対偶
   “ 山　ならば mountain” の対偶は，”mountain でなければ山ではない．” になる．
3. 対訳文
   対訳文 ”彼は山に行く． He go to the mountain ” を想定する．
4. 対訳単語と文パターン
   対訳文において，対訳単語を変数とすると，以下の文パターンが得られる．
   “彼は X に行く．　He go to the X”
5. 対訳単語の対偶と文パターン
   以下の推論によって，対訳単語の対偶は，文パターンとして捉えることが可能である．
   (a) “He go to the X” は，” He go to the mountain” において”mountain でなければ” になる．
   (b) “彼は X に行く X” は，” 彼は山に行く．” において” 山ではない．” になる．
   (c) ゆえに，
   “ 山　ならば mountain”　の対偶は，
   “He go to the X” は “彼は X に行く X” になる．
   (d) したがって，
   文パターンは，対訳単語の対偶として捉えることが可能である．

2.4　文パターンと単語の対偶の問題点

文パターンを対訳単語の対偶と考えたとき，矛盾する文パターンが存在する．以下に例をあげる．このような文パターンは，削除すべきである．

1. 対訳文
   “ 彼女は我を通した．　She had her own way.”
2. 対訳単語
   “彼女 She”
3. 文パターン
   対訳文と対訳単語から，以下の文パターンが作成される．
   “ X は我を通した．　X had her own way.”
4. 対訳単語の対偶
   “彼女は She ならば，” “ She でない　ならば彼女　でない．”
5. 文パターンと対訳単語の矛盾
   文パターン “ X は我を通した．　X had her own way.” は，パターン中に her が存在するため，この文パターンは矛盾する．

3　文パターンと単語の対偶の問題点の解決方法

上記 2.4 節で示した文パターンは，対訳単語の対偶の観点から棄却すべきである．この根本的な原因は，対訳文において，日本語１単語に英文中の対訳単語が複数存在することに起因する．この解決方法として，文パターンを作成するとき以下の条件をつける．

1. “１：多” パターンの削除
   日本語の１パターンは，英語において複数パターンを持たない．削除すべき例を表 3 に示す．

表 3　1: 多パターンの例

X は我を通した．　X had her own way. X は我を通した．　She had X own way.

2. “多:1” パターンの削除
   英文１パターンは，日本語英語において複数パターンを持たない．削除すべき例を表 4 に示す．

表 4　多:1 パターンの例）

X は我を通した．　She had X own way. 彼女は X を通した．　She had X own way.
（以下の単語辞書を想定） 彼女 She　　我 her

4 翻訳実験に用いた文パターンの作成方法

4.1 文パターンの自動作成

本研究で利用する文パターンは論文 [1] の方法で自動的に作成する．概要を以下にしめす．

1. 対訳文を準備
2. 対訳単語
   IBM model3 を利用して，対訳文から対訳単語を作成
3. 対訳文パターン-0
   対訳単語と対訳文から対訳文パターン-0 を作成．ただし，変数は単語
4. 対訳句-1
   対訳パターン-0 と対訳文から，対訳句-1 を作成
5. 対訳文パターン-1
   対訳句-1 と対訳文から対訳文パターン-1 を作成なお，変数は，句および単語
6. 対訳句-2
   対訳文パターン-1 と対訳文から，対訳句-2 を作成
7. 対訳文パターン-2
   対訳句-2 と対訳文から対訳文パターン-2 を作成

4.2 文パターンの選択 (対訳文パターン-3)

本実験では，4.1 節で作成した文パターンから，3 章で示したパターンの選択を行う．具体的には，得られた文パターン（対訳文パターン-0, 対訳文パターン-1, 対訳文パターン-2）から，3 章のフイルターを通して，対訳単語の対偶に矛盾のない文パターン (対訳文パターン-3）を選択する．

4.3 文パターンの作成実験

作成した文パターンを表 5 にまとめる．ただし，(対訳文パターン-3）は，文パターンの単語の長さを最大 16 最小 3 の制限をつけている．

表 5 実験に用いた学習データ

対訳文 （Baseline,+Pattern,Propose で利用） 単文　約 16 万 + 複文　約 9 万 [2] + JPARACRAWL テスト文類似文 約 22 万
生成した文パターン (+Pattern で利用） (対訳文パターン-0+ 対訳文パターン-1 + 対訳文パターン-2) 1,419,793
選択した文パターン (Propose で利用) (対訳文パターン-3) （3 章のフイルターを利用） 524,437

作成された文パターン (対訳文パターン-3) の例を表 6 に示す．

人手でランダムに 100 パターンを調査したところ，パターンの精度は 90%を超えている．ただし，パターンの正誤の判断は，人によって基準が異なるため，値に信頼性を求めることは困難である．

表 6 文パターンの例

彼 の *X* の 方針 から の 逸脱 を すぐ に 見つけて 非難 した 。 They were quick to see and condemn his deviation from the *X* line .
X に も その 兵士 は 足 を 砲弾 で 撃ち落とさ れた 。 The *X* soldier had his leg shot off .
その 主題 に *X* は わずか な 関心 しか もて ない 。 The subject has only slight interest for *X* .
ぼく は 借金 を 返さ ない ので *X* を 訴えた 。 I sued *X* , as he did not pay back the money .
彼 は この 2、3 日 *X* に 変調 を きたし て いる 。 He has been in poor *X* for the last few days .
そんな 陰謀 に 加わる こと は 断じて *X* ない 。 I *X* never take part in such a plot .
彼 は *X* の 向こう 側 に 住ん で いる 。 He lives on the otherside of the *X* .
X が 未経験だ と いう こと を 考慮 して ください 。 Please make allowance for *X* inexperience .

5 翻訳実験

NMT の学習データに，作成した文パターンを追加して，翻訳精度の変化を調査した．

5.1 実験条件

翻訳実験はベースライン (Baseline) とパターン追加 (+ Pattern) と提案手法 (Propose) の計 3 種類おこなった．各実験条件を表 7 に示す．

表 7 翻訳の学習データ量

A	ベースライン (Baseline)
	単文　約 16 万 + 複文　約 9 万 + JARACRAWL テスト文類似文 約 22 万
B	パターン追加 (+ Pattern)
	単文　約 16 万 + 複文　約 9 万 + JARACRAWL テスト文類似文 約 22 万 + 文パターン約 144 万文 (対訳文パターン-0 + + 対訳文パターン-1 + 対訳文パターン-2)
C	提案手法 (Proposed)
	単文　約 16 万 + 複文　約 9 万 + JARACRAWL テスト文類似文　約 22 万 + 文パターン約 52 万文　(対訳文パターン-3)

その他の条件を以下に示す．

表 8 翻訳の実験条件

テスト文	100 文（複文）
使用したソフト	OpenNMT Ver 2.3.0
使用したパラメータ	基本は default
vocabulary size	200,000

なお，翻訳方法は，多数決モデルを利用し，正順計 4 方向，初期値 4，N-best 4, テスト類似文は TF で

4 種類 (累積 1，累積 4，累積 16，累積 256) を利用してテスト文ごとに最適化している [3].

5.2 実験結果

実験の出力例を表 10 に示す．また，自動評価による実験結果を表 9 に示す．なお，参考のために google 翻訳の結果も載せる．

表 9 実験結果 (自動評価)

	BLUE	meteor	TER	RIBES	STR [4]
C) Proposed	0.322	0.573	0.513	0.813	0.10
A) Baseline	0.265	0.516	0.557	0.804	0.11
B) +Pattern	0.288	0.539	0.536	0.811	0.10
google	0.260	0.554	0.577	0.820	0.05

この結果から，以下のことが示される．

1. 文パターンの選択の効果
   提案手法と，“+ Pattern” を比較すると，翻訳精度が，大幅に向上している．これは，文パターンを対訳単語の対偶を考慮して選択（3 節）する有効性が示される．
2. 文パターンの効果
   ベースラインと “+ Pattern” を比較すると，学習データに文パターンを追加することにより，翻訳精度が，ある程度，向上することが示される．

6 考察

6.1 文パターンの選択の有効性と Masked Languge Model

実験結果から，文パターンを選択して学習することにより，翻訳精度が大きく向上することがわかる．この最大の原因は，3 節に示した仮説にあると考えている．今回の実験では，要素合成法の問題が生じる可能性のある文パターンを，できるだけ排除した．この効果が出たと思う．

なお，1 変数の文パターンは，一種の Masked Language Model と考えることができる．つまり，BERT や BART などの手法の類似性がある．この手法の効果が内在していると考えている．

6.2 google との対比較

人手による google との対比較をおこなった．この結果から，人手評価では，google のほうが良かった．しかし，google の学習データは，非常に巨大である．また今回使用したテスト文は google において closed data になっている可能性も高い．

6.3 テスト文の対偶と巨大な学習データ

機械翻訳において，学習データは多いほど翻訳精度が向上すると考えれている．この根拠を対訳文の対偶から考察してみた．

1. 対訳文
   対訳文として，“彼女は我を通した　She had her own way” を想定する．
2. 対訳文の対偶
   対訳文の対偶は，“She had her own way” でないならば，“彼女は我を通した” ではない．
3. 他の学習データ
   よって学習データば，テスト文の対偶とみなせる．
4. 学習データ量と対偶の関係
   したがって，学習データは多いほど，テスト文の対偶が多く収集される．その結果翻訳精度が向上する．

以上の観点から，翻訳精度を向上させるに，最も必要なのは，正確で大量の対訳データの収集である．翻訳アルゴリズムは，さほど重要でないかもしれない．

6.4 評価の問題 自動評価と人手評価

自動評価では，提案手法が，google を大幅に，上回っている．しかし，人手評価では，google が良かった．この問題は自動評価の問題と考えている．自動評価は基本的に部分的な正解の平均である．しかし，人手評価は文全体の意味の評価である．この違いが出現したと思う．なお，人手評価にも問題がある．最近の学生は google 翻訳に慣れている．そのため，テスト文において，reference（人手翻訳）よりも，google 翻訳が良いと判断する学生が多い．

7 まとめ

パターン翻訳は，古典的な機械翻訳の方法である．そして，NMT の学習に文パターンを加える方法がある．しかし，翻訳精度は，文パターンを追加しても，あまり向上しなかった．この原因として，追加した文パターンに要素合成法の問題があると考えた．そこで，対訳単語の対偶を考慮して文パターンの選択をおこなった．

提案したシステムの翻訳精度を評価したところ，文パターンの選択前と比較すると BLEU 値が 0.288 から 0.322 に大幅に向上した．つまり提案手法の有効性を示した．

謝辞

評価者の大学院生の名村 太一と松本 武尊 両氏に深く感謝します.

参考文献

[1] 村上 仁一森本 世人. 似度を利用した変換テーブルの精度向上. 言語処理学会第 27 回年次大会, Vol. P3-7, , 2021.

[2] 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.

[3] 村上仁一. 多数決による自己回帰モデルに基づく機械翻訳. 言語処理学会第 30 回年次大会, No. C10-4, 2024.

[4] 村上仁一. 機械翻訳における文一致率による評価. 人工知能学会全国大会論文集 第 27 回, 2013.

表 10 実験結果　出力文　まとめ

入	テスト文
B	Baseline
P	Proposed
+	+Pattern
G	Google 翻訳

入	～ で の デモ は 過去 の 抗議 デモ など ちっぽけ に 見えて しまう ほど の 規模 に なる 見込み だ .
B	
P	The demo in is expected to be the size of the past protests .
G	The demonstration in is expected to be of such a scale that past protests will seem small .
+	The demo below is expected to be the size of products such as past protest demonstrations .
入	送り 状 第 2929 号 に 従って 船積み した 商品 の 損傷 の ため 、 貴 殿 に 迷惑 を おかけ し 申し訳 ございませ ん .
B	We apologize for the inconvenience caused by the shipment of the shipment of the shipment shipment to you .
P	We apologize for any inconvenience this may cause you .
G	We apologize for the inconvenience caused to you due to the damage to the goods shipped according to invoice number 2929 .
+	We apologize for any inconvenience this may cause you .
入	進退 伺い を 出したら その 儀 に 及ば ず と の こと であった 。
B	When I spoke to the story , I did not accept my speech .
P	It was not worth the vanity to go to the wall .
G	I was told that if I asked them to move forward , I would not be able to do so .
+	And that was not the way to complain of .
入	ご存じ の ように 、 自動車 業界 は 法律 の 要求 を 満たす こと が でき ない と 主張 して い ます .
B	The fact is that the automotive industry cannot meet the requirements of the law .
P	As you know , the automotive industry claims that the automotive industry cannot meet the demands of the law .
G	As you know , the auto industry claims that it cannot meet the demands of the law .
+	As you know , argue that the automotive industry cannot meet legal requirements .
入	その 国 は この 戦い を 前 から 予知 して それ に 備えて きた 。
B	The country has been looking forward to it .
P	That country has predicted this battle and prepared for it .
G	The nation has foreseen this battle and prepared for it .
+	The country has been preparing for it from before the war .
入	みだり に 干渉 して は 有害 無益だ .
B	Your interference will do more harm than good .
P	Good interference will do more harm than good .
G	Unnecessarily interfering is harmful and useless .
+	Your interference will do more harm than good .
<入	軍隊 を 出して ストライキ を 押さえ 付けた .
B	They forced the army to hold off the strike .
P	They put out the army and put down the strike .
G	I sent in the army to suppress the strike .
+	He held the strike on a strike .
入	無実 が 証明 さ れれば 死んだ 父 も 成仏 でき ます 。
B	My father ’s soul will rest in peace if he is exonerated .
P	My father ’s soul will rest in peace if he is exonerated .
G	If my innocence is proven , even my dead father can be enlightened .
+	My father ’s soul will rest in peace if he is exonerated .
入	この 花 が 咲く と もう 春 です 。
B	These flowers bloom in spring .
P	These flowers bloom in spring .
G	When this flower blooms , it’s already spring .
+	These flowers bloom in spring .
入	天才 の する こと は 普通 の 物差し で は 、 計れ ない 。
B	Genius is not possible by ordinary means .
P	A genius can not be measured by ordinary yardstick .
G	You can’t measure what a genius does with an ordinary ruler .
+	A genius can be measured by ordinary measure .
入	犬 が い なく なった ので 子供 たち は すっかり しょげて いる 。
B	The children are completely spoiled by the dogs .
P	The children are completely disheartened because there are no dogs .
G	The children are devastated because the dog is gone .
+	The children are completely disheartened because the dog is gone .
入	きみ の あいまいな 説明 で は あの 娘 が だれ だ か わから なかった 。
B	I couldn’t recognize the girl from the vague description of her you gave me .
P	I couldn’t recognize the girl from the vague description of her you gave me .
G	I couldn’t figure out who that girl was from your vague description .
+	I couldn’t recognize the girl from the vague description of her you gave me .
入	そのような 危機 に 直面 して も 彼女 が 冷静な の に 驚いた 。
B	Even when she saw such a crisis , she was surprised .
P	I was surprised at her calm in the face of such a crisis .
G	I was surprised at how calm she was in the face of such a crisis .
+	Even if she faced such a crisis , she was surprised .