日英翻訳を対象としたイディオム表現の評価指標の提案

廣瀬惟歩 [1]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　自然言語処理学研究室
{hirose.yuiho.ia8,taro}@is.naist.jp

概要

ニューラル機械翻訳（NMT）の課題の一つして，イディオムなどの非構成的な表現の翻訳が挙げられる．NMT システムは原文を単語単位で解釈して翻訳するため，非構成的な意味を有するイディオム表現に対しては誤訳が度々生じる．また，既存の自動評価指標は局所的な評価ができず，イディオム表現の翻訳性能の評価には適さないという問題点がある．本研究では，日本語と英語を対象に，イディオム表現の翻訳性能の評価に適した新たな自動評価指標を提案する．具体的には，目的言語側でイディオム表現を検出し，その個数と原言語側のイディオム表現の個数とを比較して，翻訳モデルの性能を評価する．実験の結果，BLEU や BERTScore でのスコアの高さと，全体の出力訳に含まれる正しいイディオム表現の割合は翻訳モデルによって傾向が異なることが判明した．

1　はじめに

Transformer[1] などのニューラル機械翻訳（NMT）システムにとって，複合語表現（MWE）の翻訳は現状の課題の一つである．MWE の中でも，イディオム表現は非構成的な意味を持ち，構成される単語からは意味を推測できない場合が多い表現である．NMT システムは翻訳対象となる文章の意味が構成的であることを前提として翻訳を行うため，イディオム表現を文字通りに解釈する傾向にある．その結果，イディオム表現を直訳もしくは省略する翻訳エラーが多々発生する．

また，イディオム表現の翻訳の評価に適した自動評価指標が存在しないことも課題の一つとして挙げられる．BLEU[2] を始めとするグローバル評価指標は，翻訳文の全体を考慮するものであり，局所的な評価ができないため，イディオム表現の翻訳性能の評価には適さない．イディオム表現の翻訳結果を自動的かつ定量的に評価するには，既存のグローバル評価指標とは別に，翻訳文の特定の箇所に絞った専用の評価指標が必要とされる．

イディオム表現に対する自動評価指標の先行研究として，Shao ら [3] は英語と中国語のペアを対象に，イディオム表現を直訳した際に生じるであろう単語を登録したリストを活用する方法を提案した．具体的には，翻訳モデルが出力した翻訳文のうち，このリストに登録された単語を含むものの割合を評価スコアとしている．しかし，この手法には，対応する単語を逐一手動でリストに登録しなければならないという欠点が存在する．

本研究では，日英翻訳を対象とした，イディオム表現の翻訳性能を評価するための新たな自動評価指標を提案する．具体的には，イディオム検出器によって出力訳のイディオム表現を自動的に抽出し，出力訳と参照訳に含まれるイディオム表現の割合の類似度で翻訳モデルを評価する．

イディオム検出器の構築，並びに既存の翻訳モデルでの日英翻訳に対する評価実験の結果，BLEU と BERTScore のスコアが最大である翻訳モデルと，提案手法での評価スコアが最大である翻訳モデルが異なると判明した．この結果は，BLUE や BERTScore といったグローバル自動評価指標では，機械翻訳の出力訳に含まれるイディオム表現を正しく認識できないことを示していると言える．

2　関連研究

イディオム表現の翻訳結果を BLEU などのグローバル評価指標で評価した研究は，Fadaee ら [4] によるものを始め，数多く存在する．Fadaee らの研究では，イディオム表現に特化したデータセットの機械翻訳結果の BLEU スコアが，イディオム表現を含まない標準データセットの機械翻訳結果よりも低下したことが示されている．

イディオム表現の自動評価指標に関する先行研究には，Zaninello ら [5] によるものが挙げられる．Zaninello らは英語とイタリア語のペアを対象に，

MEW に対する参照訳と出力訳を文字単位で考慮し，双方のレーベンシュタイン距離を算出して，データセット全体の評価スコアを得る手法を考案した．

他にも，Shao ら [3] は英語と中国語のペアを対象に，イディオム表現に対する自動評価指標として，独自で構築したブラックリストを用いた評価方法を提案した．このリストには，対象となるイディオム表現を直訳した際に生じると思われる単語が登録される．そして，イディオム表現を含む文章を機械翻訳モデルに翻訳させ，出力された全ての文字列のうち，リストに登録された単語を含むものの割合を評価スコアとしている．

Shao らの手法では手作業でこのブラックリストを構築しているが，Christos ら [6] はこの研究をもとに，自動アルゴリズムを活用して故意にイディオム表現の直訳エラーを起こし，それによって得られた単語を自動的にリストに登録する手法を提案した．

3 提案手法

本研究では，イディオム表現の自動検出器を作成し，それぞれの出力訳および参照訳の文中からイディオム表現を自動的に抽出して，全体の出力訳のうち，参照訳と同様のイディオム表現がどれほど含まれているかによって翻訳モデルの性能を評価する指標を提案する．

3.1 イディオム表現の自動検出

初めに，出力訳に含まれるイディオム表現の候補が実際に比喩的な意味であるのか，あるいは文字通りの意味であるのかを判別するためのイディオム検出器を作成する．自動検出器を作成するには，イディオム表現を含む例文を用意し，各文のイディオム表現の位置をアノテーションした上で学習を行う必要がある．本研究ではイディオム表現を含む例文のリストとして，Wiktionary のイディオムカテゴリ[1)]を活用した．このカテゴリには 7990 種類の英語のイディオム表現が掲載されており，その内 5519 種類のイディオム表現には用例も掲載されている．

これらの用例を抽出し，各英文の単語に BIO/BIEO タグ形式でイディオム表現の位置をアノテーションしたデータセットを用いて，BERT[7] の系列ラベリングモデルでタグの推論を行い，イディオム表現の自動検出器を作成した．

1) https://en.wiktionary.org/wiki/Category:English_idioms

表 1 検出実験時のデータセットの内訳

データ	文	イディオム有	イディオム無
train	23,134	11,567	11,567
test	2,568	1,284	1,284
all	25,702	12,851	12,851

3.2 評価手法

機械翻訳モデルの出力訳を $\hat{y}$，参照訳を y としたとき，それぞれの訳文から抽出されたイディオム表現の個数（$f(y)$ と定義）をもとに適合率，再現率，F1 スコアを計算する．すなわち，出力訳と参照訳のそれぞれの訳文に含まれるイディオム表現の数が同等であれば，それらは一致しているという仮定を置き，その上で対象となる翻訳モデルがどれほどイディオム表現を出力できるかを見てその性能を評価する．この仮定を置いた理由は，イディオム表現が検出されたとしても，該当する出力訳と参照訳で単語列が完全に一致することは稀であると判断したためである．訳文の総数を n とすると，提案手法の適合率（Pre.），再現率（Rec.）はそれぞれ (1)，(2) 式のように表される．

$$Pre. = \frac{\sum_{i=1}^{n} min(f(\hat{y}_i), f(y_i))}{\sum_{i=1}^{n} f(y_i)} \quad (1)$$

$$Rec. = \frac{\sum_{i=1}^{n} min(f(\hat{y}_i), f(y_i))}{\sum_{i=1}^{n} f(\hat{y}_i)} \quad (2)$$

4 実験

4.1 イディオム表現の検出

イディオム検出器の train 及び test データセットは，上述した Wiktionary のイディオムカテゴリに掲載された用例を抽出して構築した．具体的には，抽出によって得られた 12,851 文の英文を 9:1 の割合で train と test データセットに分け，双方のデータセットに対して，OpenSubtitle2016 コーパスから抽出したイディオム表現のない英文を追加した．イディオム検出器の学習，評価に用いたデータセットの内訳を表 1 に示す．

英文中のイディオム表現をモデルに認識させるため，英文の各単語に B, I, O, もしくは（BIEO 方式の場合）E のタグを割り当て，train データを用いてモデルに文中のイディオム表現のタグ及びスパンを学

表 2　イディオム検出の各スコア

	Tag-base				Span-base			
タグ方式	Acc.	Pre.	Rec.	F1-micro	Acc.	Pre.	Rec.	F1-micro
BIO	**0.967**	**0.749**	**0.735**	**0.742**	**0.967**	**0.585**	**0.643**	**0.617**
BIEO	0.965	0.729	0.713	0.721	0.965	0.577	0.636	0.605

イディオム: break the ice							
タグ方式	We	have	already	broken	the	ice	.
BIO	O	O	O	B	I	I	O
BIEO	O	O	O	B	I	E	O

図 1　タグを割り当てた例文

習させた．具体的には，イディオム表現の始めの単語には B を，それ以降のイディオム表現を構成する単語には I を，イディオム表現の終わりの単語には E を，それ以外の単語には O を割り当てた．英文にタグを割り当てた例を図 1 に示す．

なお，学習モデルは BERT モデル [7] の一種である Bert-base-multilingual-cased を使用した．学習時の設定は BIO/BIEO タグともに max_seq_length=128 とした．

4.2　イディオム表現の翻訳

イディオム表現の翻訳に用いるデータセットの構築には，OpenSubtitle2016 コーパス [8] を活用した．OpenSubtitle2016 は映画や TV 番組の字幕を収集したコーパスであり，英日データセットでは約 2,083,600 の対訳文が収録されている．このコーパスを本研究に用いた理由は，イディオム表現は基本的にカジュアルな会話や文章で用いられるものであり，字幕や小説に特化したコーパスであれば，イディオム表現が文中に出現する頻度が高いと判断したためである．

このコーパスからイディオム表現を含む英文を抽出するため，Wiktionary のイディオムカテゴリに掲載されたイディオム表現のうち，例文が付属した 5519 種類のイディオム表現を抽出し，それらと動詞の態を適宜変更したものを併せてイディオム辞書とした（例えば *break the ice* というイディオムがあれば，*broke the ice*，*breaking the ice*，及び *broken the ice* をリストに追加した）．この辞書を用いて抽出を行った結果，イディオム表現の候補を含む英文が 306,720 文得られた．これを 8:1:1 の割合で分け，245,376 文を train データに，30,672 文を dev 及び test データに割り当てた．

これらのデータセットを基に，beam-search を用いた Fairseq，sampling を用いた Fairseq，Huggingface を活用して日英データセットで fine-tuning を行った T5-Base モデルの 3 つで学習を行い，test データセットで日英翻訳を実施した．その上で，各翻訳結果に対して BLEU，BERTScore，提案手法の 3 つのスコアを計測し．異なる翻訳モデルに対して提案手法による評価スコアがどのような傾向を示すのかを確認した．なお，Fairseq と T5-Base の学習時の設定は双方ともに学習率 7e-4，epoch 数 30，最大トークン数 6000 とし，翻訳時の設定は beam-width=5 とした．

5　結果

5.1　イディオム検出器の評価

検出モデルの評価は，BIO タグ方式と BIEO タグ方式それぞれに対し，Tag-base と Span-base の両方で行った．イディオム表現の検出モデルによる予測結果の各スコアを表 2 に示す．

表 2 の Tag-base で，Pre. はモデルが予測した B，I，E タグのうち実際に正解である割合，Rec. は test データセット内の B，I，E タグのうちモデルが正しく予測できた割合を示す．Span-base では，Pre. はモデルが予測したスパンのうち正解と完全一致する割合，Rec. は test データセット内のスパンのうちモデルが予測したスパンと完全一致する割合を示す．

表 2 を見ると，BIO タグ方式のスコアが BIEO タグ方式のスコアを全体的に上回っており，正解率に関しては双方ともに 96%以上を記録していることが分かる．一方で，Span-base における再現率は最高でも 64%に留まっており，test データセットのうち 6 割ほどのイディオム表現のスパンしか検出できなかったことが窺える．

5.2　翻訳結果の評価

beam-search もしくは sampling を用いた Fairseq，並びに T5-Base モデルに対する BLEU，BERTScore，提案手法のスコアを表 3 に示す．BIO と BIEO の項は，翻訳モデルの出力訳に対して BIO もしくは BIEO タ

表 3　イディオム翻訳の評価結果

翻訳モデル	BLEU	BERTScore			提案手法 BIO			提案手法 BIEO		
		Pre.	Rec.	F1-micro	Pre.	Rec.	F1-micro	Pre.	Rec.	F1-micro
Fairseq beam-search	**11.9**	**0.797**	**0.781**	**0.788**	0.0627	0.959	0.118	0.0602	**0.963**	0.113
Fairseq sampling	5.19	0.735	0.741	0.738	**0.139**	**0.961**	**0.243**	**0.142**	0.961	**0.248**
T5-Base	2.00	0.695	0.678	0.685	0.00202	0.914	0.00403	0.00152	0.923	0.00303

イディオム: *put 〜 the hammer down*												
	She	always	put	the	top	up	and	the	hammer	down	.	
Ref.	O	O	B	I	I	I	I	I	I	I	O	
Hyp.	O	O	B	I	I	I	O	O	O	I	O	

図 2　タグ予測の例

翻訳モデル	翻訳結果
参照訳	You shouldn't have lost your temper.
beam-search	You don't get carried away.
sampling	It's not good to lose your temper.
T5-base	I'm going to get you out of here.

図 3　翻訳モデルの出力訳の例

グで学習したイディオム検出器でイディオム表現の抽出を実行し，抽出されたイディオム表現をもとに提案手法での評価を行った結果を表す．

表 3 の提案手法の項目を見ると，BIO 及び BIEO ともに全ての翻訳モデルの適合率が低く，イディオム表現が含まれる英文を翻訳モデルがさほど出力できなかったことが窺える．T5-Base に関しては，BLEU と提案手法の適合率の双方で低い結果が出ている．一方で翻訳モデルごとの評価スコアを見ると，BLEU，BERTScore ともに beam-search を利用した Fairseq が最大であるのに対し，提案手法でのスコア算出法では，sampling を利用した Fairseq が最も高いスコアを出している．

6　分析

6.1　イディオム表現の検出

作成したイディオム検出器の性能を測る実験において，検出器の予測結果を見ると，図 2 のように，イディオム表現の途中に目的語などが入る場合にはタグを正しく推測できないケースが多く見られた．このような長大かつ語彙的変化を受けているイディオム表現を正確に検出させるためには，目的語の異なるイディオム表現のデータを充実させ，検出器の学習の段階でイディオム表現のパターンを認識させる必要があると思われる．

また，実際にはイディオム表現でない箇所を検出器が誤検出したケースも多々見られた．一方で，実際にはイディオム表現であるにも拘らず，データセットの構築段階でアノテーションされていなかったものを検出できていた場合も少数ながらあった．

6.2　イディオム表現の翻訳

イディオム表現の翻訳に関する実験において，各翻訳モデルの翻訳結果の一例を図 3 に示す．ここで，原言語側の日本語文は「あなたは短気を起こすべきじゃなかった」，目的言語側の英文は「*You shouldn't have lost your temper*」であり，含まれるイディオム表現は *lost your temper* である．表 3 に示されている通り，sampling を用いた Fairseq は参照訳とほぼ同様のイディオム表現を出力できている．

7　結論と今後の課題

本研究では，イディオム表現の機械翻訳という分野を対象に，全体の出力訳のうち，参照訳と同様のイディオム表現を含む割合をスコアとする評価指標を提案した．実験の結果，BLEU や BERTScore でのスコアが比較的高い翻訳モデルでも，イディオム表現を含む文章を出力するのは困難であることが判明した．

今後の課題として，イディオム表現の検出に関しては，語彙的変化を受けるイディオム表現や，途中に長い目的語を取るイディオム表現を正確に検出する方法を探る必要がある．

また，提案手法の評価指標についても，他の翻訳モデルに対して同様の実験を行い，イディオム表現に対する出力訳の傾向や，提案手法と既存の評価指標との相関関係を確認する必要がある．加えて，本研究では参照訳と出力訳に含まれるイディオム表現は一致しているという仮定を置いたが，将来的には BERT のベクトル表現を活用し，双方の距離の近さを見る必要もある．

謝辞

本研究は JSPS 科研費 JP21H05054 の助成を受けたものである．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. https://arxiv.org/abs/1706.03762.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Annual Meeting of the Association for Computational Linguistics**, 2002.

[3] Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. Evaluating machine translation performance on Chinese idioms with a blacklist method. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[4] Fadaee Marzieh, Bisazza Arianna, and Monz Christof. Examining the tip of the iceberg: A data set for idiom translation. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[5] Andrea Zaninello and Alexandra Birch. Multiword expression aware neural machine translation. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020.

[6] Christos Baziotis, Prashant Mathur, and Eva Hasler. Automatic evaluation and analysis of idioms in neural machine translation. 2022. https://arxiv.org/abs/2210.04545.

[7] Devlin Jacob, Chang Ming-Wei, Kenton Lee, and Toutanova Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, 2019.

[8] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, 2016.