

機械翻訳の前処理のための言い換え辞書自動作成手法の提案

胡 尤佳^{1,a)} 岡田 真¹ 森 直樹¹

概要: 近年, ニューラルネットワークを用いる手法が自然言語処理の多くのタスクで成果を上げている. 機械翻訳の分野でも, ニューラル機械翻訳が登場し, これまでのフレーズベース機械翻訳や統計的機械翻訳と比べて翻訳の質が飛躍的に上がり, 流暢性の高い翻訳もできるようになった. しかし, 低頻度語や未知語が存在することにより, 翻訳の際に意味が通じなくなるという問題が依然存在している. このような低頻度語や未知語の問題を解決するために, 機械翻訳の前処理をするアプローチがある. 本研究では, 作成した機械翻訳のモデルを直接用いて, モデル作成時に用いた学習データからそのまま言い換え辞書を自動作成する手法を提案する. その後, その辞書を用いた低頻度語の高頻度語への言い換えを入力文の前処理として施すことにより, 意味を保持したまま翻訳精度を向上できると考える. 本稿では提案手法の有効性を実験により検証し, 考察した.

1. はじめに

近年, ニューラルネットワークを用いる手法が自然言語処理の多くのタスクで成果を上げている. 機械翻訳の分野でも, ニューラル機械翻訳が登場し, これまでのフレーズベース機械翻訳や統計的機械翻訳と比べて翻訳の質が飛躍的に上がり, 流暢性の高い翻訳もできるようになった.

しかし, ニューラル機械翻訳においても, 低頻度語や未知語が存在することにより, 翻訳の際に意味が通じなくなるという問題が依然存在している. 通常の機械翻訳の場合, 出力層における語彙数が制限されているため, 低頻度語は未知語と見なされ, 意味繋がりが中断され, 翻訳精度が悪くなる. このような低頻度語や未知語の問題を解決するために, 機械翻訳の前処理をするアプローチがある.

本研究では, 作成した機械翻訳のモデルを直接用いて, モデル作成時に用いた学習データからそのまま言い換え辞書を自動作成する手法を提案する. その後, その辞書を用いた低頻度語の高頻度語への言い換えを入力文の前処理として施すことにより, 意味を保持したまま翻訳精度を向上できると考える. 本稿では提案手法の有効性を実験により検証し, 考察した.

2. 先行研究

機械翻訳の前処理をすることで未知語を削減し, 翻訳精

度の向上を試みる先行研究はいくつか存在する.

Sennrich らは, 系列に対するデータ圧縮手法で, サブワード分割アルゴリズムである Byte Pair Encoding (BPE) を文字列に適用し, 未知語を頻出する部分文字列の系列に分解して学習することで未知語を削減した [1]. この手法では, 未知語は減らせるが, 貪欲に単語を分解するため, 頻出する部分文字列の意味を考慮しておらず, 意味を保持できない問題がある.

Li らは, 単言語データで Word2Vec [2] により単語分散表現のモデルを学習をし, 学習データにおける未知語を cos 類似度を基に同義となる既知語に置き換えた後, 翻訳モデルを学習している [3]. その後, テストデータの入力文にも言い換えを適用し, 学習したモデルで翻訳することで翻訳精度を上げた. Li らの手法では, 学習データとテストデータの両方で言い換えをしている.

関沢らは, 学習データの目的言語文に存在する低頻度語を, 大規模パラフレーズ辞書を用いて高頻度語に言い換えてから翻訳モデルを作成し, 翻訳する手法を提案した [4]. 関沢らの手法では, 目的言語文に低頻度語が存在する場合, その単語またはその単語を含むフレーズを高頻度な単語またはフレーズに言い換えることで, 言い換え前後の意味を保持したまま翻訳結果の未知語の削減を求める.

3. 提案手法

我々は, 大規模パラフレーズ辞書に依存しない言い換えアプローチを提案する. 提案手法として, 原言語の学習データにおける単語の出現頻度に注目することで, パラフレーズ辞書を学習データ内の高頻度語と低頻度語から自動

¹ 大阪府立大学, 〒 599-8531 大阪府堺市中区学園町 1 番 1 号
Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku,
Sakai, Osaka 599-8531, Japan

^{a)} ko@ss.cs.osakafu-u.ac.jp

作成する手法を考える。その際にあらかじめ作成された翻訳モデルと、そのモデルの学習データとテストデータを用いる。以下では、日英翻訳モデルとその対訳データを例とする。図1に言い換え辞書自動作成の概要を示す。

具体的手順は以下になる。

まず、原言語である日本語学習データを用いて、Word2Vecにより単語の分散表現のモデルを学習する。

次に、日本語の単語の頻度情報から、出現頻度が閾値より大きい単語を高頻度語とし、閾値以下の単語を低頻度語と設定する。図1に示すように、辞書内で高頻度語と低頻度語に分ける。

その後、ある低頻度語と全ての高頻度語との間の $\cos$ 類似度を分散表現から求め、 $\cos$ 類似度の大きい高頻度語を上から順に並べて、低頻度語の類似度ランキングを作成する。ランキング上位のいくつかの単語をある低頻度語の言い換え候補とする。

図1の例では、ある低頻度語「議論する」について、この低頻度語と $\cos$ 類似度の大きい高頻度語を調べている。このとき、上位3単語「話す」、「言う」、「戦う」という類似度ランキングを作成している。この上位3単語を低頻度語「議論する」の言い換え候補とする。

次に、学習データ内で低頻度語が含まれる文において、低頻度語の部分を使い換え候補で言い換えてできた言い換え後の文と、言い換え前の原文を翻訳モデルで翻訳する。これにより言い換え候補数+1個の出力文が獲得される。これらの出力文について、BLEUスコアが最大となった文における言い換え候補を低頻度語の言い換え対象と決定し、言い換え辞書に登録する。

図1では、学習データに「テーマについて議論する」という文があり、この文の「議論する」を言い換えることで3つの文ができるので、それらを翻訳する。また、言い換え前の原文も翻訳し、合計4文の出力文を得る。その結果を正解データと比較し、BLEUスコアが最大となった「テーマについて話す」の文に用いた「話す」を、「議論する」の言い換えとして、言い換え辞書に登録する。この際、原文による出力文のBLEUスコアが最も大きくなった場合は、より良い言い換え候補がないものと考え、言い換え対象を登録しないこととする。

続いて、上述した低頻度語に関する手順を、全ての低頻度語について同様に処理することで、低頻度語に対する高頻度語の言い換え辞書を作成する。

4. 数値実験と実験結果

提案手法の有効性を確認するために、以下の実験をした。実験においては、以下の指標から有効性を評価する。

- 未知語の削減ができたか
- テストデータへの言い換え辞書の適用により BLEU スコアが向上したか

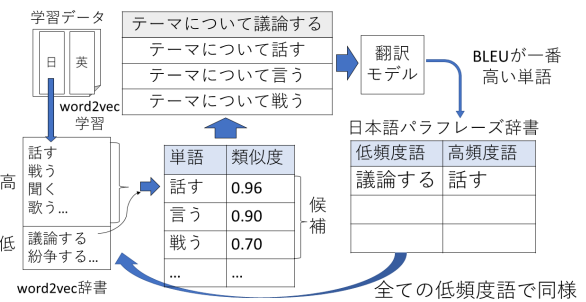


図1 言い換え辞書自動作成手法の概要

表1 実験データの内訳

内容	対訳文
Train	2797388
Test	2000
Dev	2000

4.1 データセット

今回の実験においては、翻訳モデルと言い換え辞書の作成に Japanese-English Subtitle Corpus (JESC) [5] のデータセットを用いた。このデータセットは、インターネット上からクロールされた映画とTV番組の字幕データを日英対応させることにより構築された日英対訳データであり、俗語、口語、説明文、物語開設の対訳など、既存のコーパスではあまり扱われてこなかった分野の文が多い特徴がある。実験においては、JESCの全データを用いた。表1に実験におけるデータの内訳を示す。

4.2 翻訳モデルの作成

今回の実験では、言い換え辞書作成時の翻訳と、テストデータの翻訳には、OpenNMT オープンツールキット [6] により、JESC 対訳コーパスの学習データを用いて学習したモデルを用いた。表2に学習時のパラメータを示す。

4.3 言い換え辞書の作成

まず、学習データの日本語文を形態素解析し、単語の分散表現の学習をした。単語の分散表現の学習には Gensim [7] により実装した Word2Vec を用いた。日本語学習データの形態素解析には MeCab [8] を用いた。表3に Word2Vec の学習時のパラメータを示す。

言い換えが有効な低頻度語の閾値を確認するため、低頻度語閾値を5, 10, 15, 20, 25, 30以下の6パターンに分け、その閾値においては、閾値以下の頻度の単語を低頻度語とし、閾値より大きい頻度の単語を高頻度語とした。

ある低頻度語に対し、意味的に類似する高頻度語を分散表現の $\cos$ 類似度から求め、20語の言い換え候補を用意した。その後、学習データ内で、ある低頻度語を含む文を1文抽出し、その低頻度語部分を言い換え候補の高頻度語に言い換えることで、言い換えた文20文を作成した。今回は、1文中に複数言い換え対象がある場合においても、全

表 2 翻訳モデル学習時のパラメータ

パラメータ	値
Encoder 構造	2 層 LSTM
隠れ層ユニット数 (第 1 層)	500
隠れ層ユニット数 (第 2 層)	500
Decoder 構造	2 層 LSTM
隠れ層ユニット数 (第 1 層)	500
隠れ層ユニット数 (第 2 層)	500
単語埋め込み次元数	500
原言語単語サイズ	50002
目的言語単語サイズ	50004
モデル	Global Attention
Batch Size	64
学習ステップ数	100000
最適化手法	Adam
Dropout	0.3
学習率	0.001
学習減衰率	0.5
学習減衰開始ステップ数	50000
学習減衰ステップ数	10000

て言い換えをした。これらの言い換えた文 20 文と言い換え前の原文 1 文を含む 21 文を、作成した翻訳モデルで翻訳することにより、出力文 21 文を得た。その後、それらの出力文の BLEU スコアを計算し、BLEU スコアが最大となる文に用いられていた言い換え候補を、その低頻度語に対応する高頻度語として決定し、言い換え辞書に登録した。この際、BLEU スコアが最も大きくなった文がいくつもあった場合は、その中で $\cos$ 類似度が最も大きいものを言い換え辞書に登録した。言い換え前の原文からの出力文よりも BLEU スコアが大きくなる文がなければ、その低頻度語の言い換え対象はないものとして、言い換え辞書に高頻度語の登録をしないものとした。BLEU スコアの計算には、Natural Language Toolkit(NLTK)[9]を用い、文ごとの BLEU スコアを計算した。

この操作を全ての低頻度語に対して同様に実行することで、各頻度の低頻度語に対応する高頻度語の言い換え辞書を作成した。

4.4 テストデータへの言い換え辞書の適用

作成した言い換え辞書をテストデータに適用し、入力文の前処理をした上で、翻訳モデルで翻訳した。表 4, 図 2, 図 3 に、言い換え辞書による前処理をしていないテストデータを Baseline として、低頻度語閾値 6 パターンそれぞれの閾値以下で作成した言い換え辞書により言い換えをしたテストデータと、BLEU スコアと未知語数の比較をした結果を示す。表 5 に、テストデータにおいて言い換えられた単語数と、言い換えられた文数、その中でも言い換えによって BLEU スコアが向上した文数を Better, 低下した文数を Worser, 変化のない文数を Unchanged として示

表 3 Word2Vec の学習時のパラメータ

パラメータ	値
モデル	Skip-gram
Window	3
ベクトルサイズ	200
Epoch	50
Batch Size	10000
頻度の閾値	0
単語数	23202878
語彙数	155191

表 4 テストデータへの言い換え辞書の適用による BLEU スコアと未知語数

低頻度語閾値	BLEU スコア	未知語数
言い換えなし	0.29064	667
5	0.29157	661
10	0.29151	630
15	0.29150	615
20	0.29202	604
25	0.29128	599
30	0.29046	592

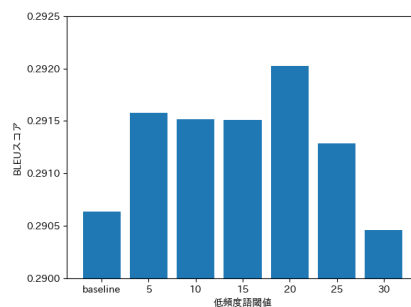


図 2 テストデータへの言い換え辞書の適用による BLEU スコア

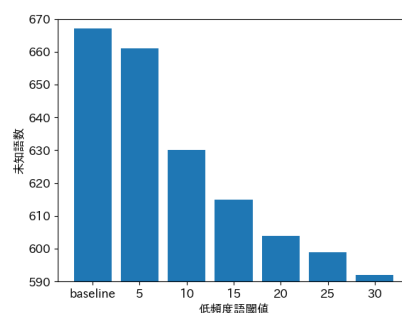


図 3 テストデータへの言い換え辞書の適用による未知語数

す。表 6, 表 7, 表 8 に, Better, Worser, Unchanged となった文の例をそれぞれ示す。

5. 考察

表 4, 図 2, 図 3 から、単語の分散表現から計算した類似度と BLEU スコアに基づいて作成した言い換え辞書による入力文への前処理により、未知語の削減が十分に期待

表 5 テストデータにおける言い換えられた単語数と文数と文の内訳

低頻度語閾値	単語	文	Better	Worse	Unchanged
5	84	78	25	18	35
10	136	123	42	32	49
15	177	162	59	49	54
20	215	194	77	58	59
25	265	237	88	72	77
30	308	275	96	88	91

できると言える。また、低頻度語閾値 25 までは、言い換え辞書による BLEU スコアの向上が見られ、低頻度語閾値 20 と設定した際に最も BLEU スコアが向上する結果となった。よって今回の実験では、適切な閾値が 20 前後であったと考えられる。

また、表 5 より、言い換え辞書により BLEU スコアが向上した文数が、低下した文数よりいずれも大きい結果となり、翻訳精度向上の上で言い換え辞書が有効であったとみられる。その一方で、BLEU スコアが言い換え前後で変化しない文も多くあった。

表 6 より、BLEU スコアが向上した文の例をみると、動詞や形容詞などの多義的な意味のある低頻度な単語が、より高頻度の似た意味を持つ単語に置き換えられることで、言い換え前では未知語として翻訳されてしまう単語が、未知語にならずに意味を持つ単語として翻訳でき、前後の意味つながりを保持したまま翻訳できるようになったケースが多くみられた。

また、表 7 より、BLEU スコアが低下した例に関しては、低頻度語に多く含まれていた固有名詞など、多義性のない単語が言い換えられる、もしくは言い換えられることにより固有名詞になるなどのケースが多くみられた。その一方で、「謳わ」を「呼ば」に言い換えるように、意味を保持した言い換えができていても関わらず BLEU スコアが低下した場合もあった。このように、BLEU スコアは N-gram に基づいた評価手法となるため、意味を真に考慮した出力文の評価ができない場合もあると言える。

表 8 より、今回多く出た言い換えによって変化しない文の多くは、言い換え前後でどちらの出力文中の単語も参照文中の単語と一致せず、BLEU スコアが 0 になるという事例が多くみられた。BLEU スコアが低下した文と同様、固有名詞の言い換えによる文や意味を保持した言い換えができていない文に関しても、BLEU スコアが変化しないケースがみられた。

6. まとめと今後の課題

今回の実験から、BLEU スコアを基準として学習データから作成した言い換え辞書のテストデータへの適用により、翻訳後の BLEU スコアの向上が見られ、言い換え辞書の有効性が確認できた。極めて出現頻度の低い頻度 1, 2 などの低頻度語には多くの固有名詞が含まれており、これ

らの言い換えにより BLEU スコアが低下する、もしくは変化せず、意味を保持できない上で精度が向上しない場合が多くみられた。言い換えにより意味的な保持ができないため、これらの多義性のない固有名詞を抽出した上で、動詞や形容詞などの多義性のある単語のみを、品詞に注目し言い換えることが今後必要となると考えられる。

また、今回は言い換え辞書の作成とテストデータの評価の際に BLEU スコアを基準としたが、BLEU スコアは N-gram に基づいた表層的な評価であり、真に意味を考慮した出力文の評価ができない場合があると考えられる。今後は、文の分散表現を得る手法 [10][11] を用いて、参照文と出力文の分散表現の類似度を計算し、意味的に類似した文の評価をした上で、言い換え辞書の作成とテストデータにおける評価もする予定である。

参考文献

- [1] Sennrich, R., Haddow, B. and Birch, A.: Neural machine translation of rare words with subword units, *arXiv preprint arXiv:1508.07909* (2015).
- [2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013).
- [3] Li, X., Zhang, J. and Zong, C.: Towards zero unknown word in neural machine translation, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, AAAI Press, pp. 2852–2858 (2016).
- [4] 関沢祐樹, 梶原智之, 小町守: 目的言語の低頻度語の高頻度語への言い換えによるニューラル機械翻訳の改善, 言語処理学会第 23 回年次大会発表論文集, pp. 982–985 (2017).
- [5] Pryzant, R., Chung, Y., Jurafsky, D. and Britz, D.: JESC: Japanese-English Subtitle Corpus, *Language Resources and Evaluation Conference (LREC)* (2018).
- [6] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A. M.: Opennmt: Open-source toolkit for neural machine translation, *arXiv preprint arXiv:1701.02810* (2017).
- [7] Rehurek, R. and Sojka, P.: Gensim-python framework for vector space modelling, *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, Vol. 3, No. 2 (2011).
- [8] 工藤拓, 山本薫, 松本裕治ほか: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告自然言語処理 (NL), Vol. 2004, No. 47 (2004-NL-161), pp. 89–96 (2004).
- [9] Loper, E. and Bird, S.: NLTK: the natural language toolkit, *arXiv preprint cs/0205028* (2002).
- [10] Logeswaran, L. and Lee, H.: An efficient framework for learning sentence representations, *arXiv preprint arXiv:1803.02893* (2018).
- [11] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. et al.: Universal sentence encoder, *arXiv preprint arXiv:1803.11175* (2018).

表 6 Better の例

	文	BLEU
言い換え前	これらの問題すべてに解決策を <u>見い出さ</u> なければいけません	
出力文	all of these problems have to be <unk>	0.45241
言い換え後	これらの問題すべてに解決策を <u>見つける</u> なければいけません [閾値 20, 頻度 1]	
出力文	we have to find solutions to all these problems.	0.81435
参照文	we have to find ways to solve all these problems.	
言い換え前	前面の偏向スクリーンに全パワーを <u>切り替えろ</u>	
出力文	change all power on the front of the <unk>	0.17310
言い換え後	前面の偏向スクリーンに全パワーを <u>預け</u> [閾値 20, 頻度 18]	
出力文	and put all power on the front screen	0.18393
参照文	switch all power to front deflector screen. switch all power to front deflector screen.	
言い換え前	「宇宙を <u>握りつぶし</u> 、丸めて」	
出力文	<unk> <unk>	0
言い換え後	「宇宙を <u>切り出し</u> 、丸めて」 [閾値 30, 頻度 24]	
出力文	cut the universe off.	0.26959
参照文	to have squeezed the universe into a ball.	
言い換え前	こう <u>感慨深い</u> ものがあるよね。	
出力文	there's something like this.	0.47768
言い換え後	こう <u>良い</u> ものがあるよね。 [閾値 20, 頻度 2]	
出力文	there's something good like this.	0.49535
参照文	there's something deeply moving about it.	
言い換え前	あの津波を <u>思いだし</u> てください	
出力文	please have a <unk>	0
言い換え後	あの津波を <u>教えて</u> ください [閾値 20, 頻度 8]	
出力文	please tell me that tsunami	0.04037
参照文	and, you know, if you think about this with the tsunami	

表 7 Worser の例

	文	BLEU
言い換え前	この前エルロンのことを考えていた	
出力文	i was thinking about <unk>	0.57142
言い換え後	この前 <u>ない</u> のことを考えていた [閾値 20, 頻度 1]	
出力文	i was thinking about what i was doing.	0.44444
参照文	i was thinking about	
言い換え前	世界一素晴らしいデザインの新聞と <u>謳</u> われたのです	
出力文	it was made of the greatest design paper in the world.	0.33333
言い換え後	世界一素晴らしいデザインの新聞と <u>呼</u> ばれたのです [閾値 20, 頻度 9]	
出力文	was called the greatest design paper in the world.	0.3
参照文	they name it the world's bestdesigned newspaper.	

表 8 Unchanged の例

	文	BLEU
言い換え前	<u>きびしい</u> っすね	
出力文	it's <unk>	0
言い換え後	<u>難しい</u> っすね [閾値 20, 頻度 8]	
出力文	it's difficult	0
参照文	tough kids, huh?	
言い換え前	そんなこと知ったこっちゃないよね？ <u>神田</u> さん	
出力文	don't you know that, <unk>	0
言い換え後	そんなこと知ったこっちゃないよね？ <u>裕</u> さん [閾値 25, 頻度 21]	
出力文	don't you know that, <unk>	0
参照文	no use crying over spilled milk, right mr. kanda?	