SMTによる大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳における訳抜け削減効果

Effect on Reducing Untranslated Content by NMT integrated with Large Vocabulary Phrase Translation by SMT

筑波大学システム情報系知能機能工学域教授
宇津呂　武仁

1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。京都大学等を経て、2012 年より筑波大学システム情報系知能機能工学域教授。自然言語処理、機械翻訳、ウェブマイニングの研究に従事。

筑波大学大学院システム情報工学研究科知能機能システム専攻
木村　龍一郎

2017 年東京都市大学工学部電気電子工学科卒業。現在、筑波大学大学院システム情報工学研究科知能機能システム専攻博士前期課程在学中。機械翻訳の研究に従事。

筑波大学大学院システム情報工学研究科知能機能システム専攻
飯田　頌平

2018 年東京電機大学工学部情報通信工学科卒業。現在、筑波大学大学院システム情報工学研究科知能機能システム専攻博士前期課程在学中。機械翻訳の研究に従事。

筑波大学システム情報系情報工学域教授
山本　幹雄

1986 年豊橋技術科学大学大学院情報工学系修士課程修了。豊橋技術科学大学等を経て、2008 年より筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻教授。博士（工学）。自然言語処理、機械翻訳の研究に従事。

1　はじめに

ニューラル機械翻訳（Neural Machine Translation; NMT）はひとつの大きなニューラルネットワークで翻訳モデルを構成した機械翻訳である。翻訳モデルは対訳コーパスを用い、正解文を生成する条件付き確率を最大化するように訓練される。NMT は従来の機械翻訳手法と比較して流暢さで優れるものの、目的言語文出力の計算時間が使用する目的言語の単語数に依存するために、大規模語彙を含む翻訳に対応できない。この問題に対応する研究はいくつか存在するが[6][8]、これらの手法は未知語を単語単位で処理しているために、複合名詞の一部として出現する場合に想定通りに翻訳できない問題がある。この問題は専門用語を多く含む特許文の翻訳において特に顕著である。

このような背景から、低頻度語を含みやすい複合名詞をトークンに置き換えて翻訳モデルの訓練を行う手法が提案された[5]。トークンに置き換えられた複合名詞は統計的機械翻訳（Statistical Machine Translation; SMT）によって翻訳され、MT 出力に代入される。し

かしこの手法は、入力文を意味的にすべて翻訳することを保証できずに内容が訳抜けするという、NMT のもう一つの大きな課題に対応できていない。訳抜けに対応するために、出力文が入力文を生成する逆翻訳確率を訳抜けした内容の検出に用いる手法が提案された[3]。そこで、本稿では大規模語彙への対応手法と訳抜けの削減手法を組み合わせて翻訳を行い、翻訳結果の訳抜けの度合いと翻訳精度を評価した。

2 訓練・評価文

日英対訳特許文として、NTCIR-7 ワークショップで配布された 180 万文の対訳対のうち単語数が 40 語未満のもののみで構成された 110 万文を使用した（NTCIR-7 特許翻訳タスク[2] における日英対訳特許文。詳細は[5]参照）。日英対訳特許文のうち、1,000 文を評価文、1,000 文を開発文として抽出し、残りを訓練文として使用した。訓練文からは 2,785,108 個の対訳フレーズ対が抽出された。抽出されたフレーズ対の種類は 704,346 種で、日本語フレーズは 511,633 種、英語フレーズは 422,269 種だった。訓練文からは 2,539 個、2,171 種の対訳フレーズ対が抽出された。

3 NMT モデル・SMT モデルの訓練条件

ワードアラインメントとフレーズ翻訳テーブルを作成するための SMT モデル作成には、Moses Toolkit を使用した[4]。チューニングには開発文を使用した。NMT モデルを訓練する際のパラメータは[1] で使用されたものを用いた。エンコーダは前向き・後ろ向きの 3 層 LSTM、デコーダは前向き 3 層 LSTM で、各層の時限はいずれも 256 次元とした。入力単語の分散表現は 256 次元とした。原言語と目的言語の語彙は頻度上位 40,000 語として、その他の語は NMT モデルの語彙外の未知語とした。

表 1　訓練・評価文

訓練文	開発文	評価文
1,167,198	1,000	1,000

4 文単位の翻訳性能の評価

4.1 自動評価

表 1 に BLEU による自動評価の結果を示す。ベースラインの SMT と比較すると、大規模語彙対応とリランキングの併用手法（0 節）は BLEU が約 5.7 向上したが、ベースラインの NMT と比較して BLEU の向上は見られなかった。

表 2　自動評価の結果（BLEU）

モデル	ja→en
ベースラインSMT[4]	32.3
ベースラインNMT	38.2
大規模語彙に対応したNMTモデル	39.8
大規模語彙に対応したNMTモデル + リランキング	38.0

4.2 人手評価

[7]における人手評価尺度である「一対評価」および「JPO 基準に基づく絶対評価」を用い、評価用対訳特許文から無作為に抽出された 100 文を評価対象として、著者が評価を行った。評価対象手法、および、ベースラインとなる手法との間の「一対評価」では、評価対象手法による翻訳精度が、ベースラインとなる手法による翻訳精度を上回った文の数を W、評価対象手法による翻訳精度が、ベースラインとなる手法による翻訳精度を下回った文の数を L、評価対象手法による翻訳精度が、ベースラインとなる手法による翻訳精度と同等となった文の数を T として、一対評価のスコア（値の範囲は、−100〜100）を次式で定義する。

$$score = 100 \times \frac{W-L}{W+L+T}$$

「JPO 基準に基づく絶対評価」においては、JPO 評価基準[1] に基づき、各翻訳文に対して人手で 1〜5 の値の範囲のスコアを付与し、その平均を「JPO 基準に基づく絶対評価」のスコアとする。「一対評価」結果を表 3 に、「JPO 基準に基づく絶対評価」結果を表 4 に示す。どちらの評価結果も大規模語彙対応とリランキングの併

表 3　ベースライン NMT に対する一対評価の結果（スコアの範囲は−100 以上 100 以下）

モデル	ja→en
大規模語彙に対応したNMTモデル	18
大規模語彙に対応したNMTモデル + リランキング	37

1　https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf

表 4　JPO 基準に基づく絶対評価の結果
（スコアの範囲は 1 以上 5 以下）

モデル	ja→en
ベースラインNMT	4.1
大規模語彙に対応したNMTモデル	4.2
大規模語彙に対応したNMTモデル + リランキング	4.5

用手法が最も高かった。

5　訳抜け削減における効果の検証

5.1　訳抜け検出

後藤らは、訳抜けした内容を検出する翻訳スコアに基づいてリランキングすることによって、BLEU に関して翻訳評価結果を改善するだけでなく、NMT の訳抜けした内容を検出できることを明らかにした[3]。訳抜けした内容を検出する翻訳スコアの中から、我々は特に影響の大きい逆翻訳確率を採用した。より具体的には、ベースラインの NMT モデルを用いて、大規模語彙に対応した NMT モデルとベースライン NMT の両方の逆翻訳確率を評価した。逆翻訳確率に基づいて訳抜けした内容の多寡を予測した結果、大規模語彙に対応した NMT モデルはベースライン NMT と比較して改善した。

5.1.1　逆翻訳確率

逆翻訳確率は、MT 出力からその入力文へ強制的に翻訳し直すときの確率として定義される。入力単語の内容が MT 出力に欠落している場合、入力単語の逆翻訳確率は小さくなると予想される。この関係から、逆翻訳確率を訳抜けした内容を検出するための手掛かりとして使用した。n-best MT 出力 y^d（$1 \le d \le n$）の、入力単語 x_j（$1 \le j \le N$）についての逆翻訳確率スコア（BT-P）b_j^d を次のように定義する。

$$b_j^d = -\log p(x_j \mid x_{<j}, y^d)$$

BT-P はベースライン NMT モデルと、大規模語彙に対応した NMT モデルの両方について、フレーズトークンのない訓練セットで英語から日本語へ訓練したベースライン NMT モデルを用いて計算する。この BT-P の公式化においては、「翻訳の存在」という以下の仮定を用いる。

仮定：翻訳の存在

任意の入力単語 x_j（$1 \le j \le N$）の翻訳は、目的言語側に対訳が全く存在しない場合を除いて、n-best MT 出力 y^d（$1 \le d \le n$）のどこかに存在する。

したがって、n-best MT 出力のうち最小のスコアをもつ出力は x_j の内容を最も含んでいる確率が高いと考えられ、$\min_{1 \le d' \le n} b_j^{d'}$ は x_j の内容を含む出力のスコアと考えることができる。

次に、y^d から x_j の内容が欠落したスコアとして、逆翻訳確率の比に基づく BT-R スコア q_j^d を考え、次のように定義する。

$$q_j^d = b_j^d - \min_{1 \le d' \le n} b_j^{d'}$$

これは、各 MT 出力のスコアと、x_j の内容を含む確率が最も高いもののスコアである n-best MT 出力の最小スコアとの差である。

最後に、このスコアを入力文のすべての入力単語 $x = (x_1, \cdots x_N)$ について足し合わせることによって、入力文 x に対する MT 出力 y^d の逆翻訳確率の比に基づく BT-R スコアは次のように得られる。

$$\text{BT-R}(x, y^d) = \sum_j q_j^d$$

本稿では BT-R スコアを n-best MT 出力それぞれで計算し、もっとも値の小さい出力を最終的な翻訳文としてリランキング結果として採用した。

5.1.2　訳抜け検出結果

評価文についての逆翻訳確率の比に基づく BT-R スコアを測定した。

（ⅰ）ベースライン NMT モデル、
（ⅱ）大規模語彙に対応した NMT モデル、
（ⅲ）大規模語彙に対応した NMT モデル＋BT-R スコアによるリランキング、

の各モデルに対して、評価文における BT-R スコアの平均をそれぞれ表 5(a)に示す。「（ⅱ）大規模語彙に対応した NMT モデル」は、「（ⅰ）ベースライン NMT」モデルよりも低い BT-R スコアを達成した。この結果から、「（ⅱ）大規模語彙に対応した NMT モデル」による MT 出力は、「（ⅰ）ベースライン NMT モデル」による MT 出力よりも訳抜けした内容が少ないと考えられる。また、「（ⅲ）大規模語彙に対応した NMT モデル＋BT-R スコアによるリランキング」においては、さらに低い BT-R

表５　評価文における訳抜け予測の評価
(a) 評価文における文ごとの BT-R スコアの平均

モデル	ja→en
ベースラインNMT	16.3
大規模語彙に対応したNMTモデル	14.0
大規模語彙に対応したNMTモデル + リランキング	6.9

(b) 評価文におけるベースライン NMT と大規模語彙に対応した NMT モデルの間の BT-R スコアの差
BT-R(x, y^d) −BT-R(x, y'_d) の分布 (%)

<0					>0				
<−20	−20〜−10	−10〜−5	−5〜−1	−1〜0	0〜1	1〜5	5〜10	10〜20	>20
4.9	8.4	12.3	19.1	12.9	12.9	14.8	8.0	4.4	2.3
57.6					42.4				

(c) 評価文におけるベースライン NMT と大規模語彙に対応した NMT モデル＋リランキングの間の BT-R スコアの差
BT-R(x, y^d) −BT-R (x, y'^d) の分布 (%)

<0					>0				
<−20	−20〜−10	−10〜−5	−5〜−1	−1〜0	0〜1	1〜5	5〜10	10〜20	>20
11.8	20.3	23.0	27.7	10.5	3.1	2.9	0.4	0.3	0.0
93.3					6.7				

スコアを達成した。この結果から、BT-R スコアを用いたリランキングにより、訳抜けした内容をさらに減らすことができていると考えられる。

次に、各評価文 x に対して、上述の(ⅱ)および(ⅲ)の提案手法による NMT モデルとベースライン NMT モデルとの BT-R スコアの差を測定する。

- BT-R(x, y^d) を(ⅱ)または(ⅲ)の提案手法による NMT モデルによる MT 出力の BT-R スコア、
- BT-R(x, y'^d) を(ⅰ)のベースライン NMT モデルによる MT 出力の BT-R スコア、

と定義すると、BT-R スコアの差は以下のように定義される。

$$\text{BT-R}(x, y^d) - \text{BT-R}(x, y'^d)$$

評価文に対する BT-R スコアの差の分布を表５(b)および(c)に示す。

表５(b)の、「(ⅱ)大規模語彙に対応した NMT モデル」と「(ⅰ)ベースライン NMT モデル」との間の BT-R スコアの差の分布においては、評価文の 57.6% について、提案手法による NMT モデルの BT-R スコアはベースライン NMT と比べて小さかった。また、ベースライン NMT の BT-R スコアが提案手法による NMT モデルと比べて５以上小さかったものは 14.7% であったが、提案手法による NMT モデルの BT-R スコアがベースライン NMT と比べて５以上小さかったものは 25.6% で、10.9% 大きかった。

一方、表５(c)の、「(ⅲ)大規模語彙に対応した NMT モデル＋リランキング」と「(ⅰ)ベースライン NMT モデル」との間の BT-R スコアの差の分布においては、評価文の 93.3% について、提案手法による NMT モデルの BT-R スコアはベースライン NMT と比べて小さかった。また、ベースライン NMT の BT-R スコアが提案手法による NMT モデルと比べて５以上小さかったものは 0.7% であったが、提案手法による NMT モデルの BT-R スコアがベースライン NMT と比べて５以上小さかったものは 55.1% で、54.4% 大きかった。

5.2　入力日本語文中の単語訳抜け数の人手評価

無作為に選んだ 100 文の評価文について、日本語から英語への翻訳タスクにおいて、英語に翻訳されない入力日本語文中の単語の数を数えた。表６(a)に示すように、訳抜け単語数は、大規模語彙に対応した NMT モデルによりベースライン NMT の約 70% に減少し、さらに、逆翻訳確率でリランキングすることにより、ベースライン NMT の約 40% 程度に減少した。評価対象 100 文における訳抜け単語数の分布を表６(b)に示す。また、各モデルによる翻訳例、および、各翻訳例における訳抜け部分（「入力日本語文」欄の下線部に示す）の比較結果を表７に示す。これらの結果より、訳抜け箇所数は、大規模語彙に対応した NMT モデルを逆翻訳確率でリランキングした場合が最も少なく、大規模語彙に対応した NMT モデル、ベースライン NMT の順に多くなることが分かる。

この結果では、逆翻訳確率によるリランキングを行っ

表６　入力日本語文中の単語訳抜け数の人手評価
（評価文 100 文）
(a) 入力日本語文中の単語訳抜け数

モデル	ja→en
ベースラインNMT	73
大規模語彙に対応したNMTモデル	51
大規模語彙に対応したNMTモデル + リランキング	31

(b) 単語訳抜け数の分布（%）

モデル	単語訳抜け数										
	0	1	2	3	4	5	6	7	8	9	≧10
ベースラインNMT	64	24	6	2	0	1	1	0	1	0	1
大規模語彙に対応したNMTモデル	74	14	4	4	3	1	0	0	0	0	0
大規模語彙に対応したNMTモデル + リランキング	83	10	3	3	0	0	1	0	0	0	0

表 7　各 NMT モデルにおける訳抜け箇所およびその削減例
（訳抜け箇所を下線部に示す）
（a）英語参照訳

英語参照訳	in an air bag device for a driver , it is favorable for a driver that the air bag 16 quickly and largely extends vertically and horizontally (rightward and leftward in fig . 2) .

（b）ベースライン NMT モデル

入力日本語文	運転席用エアバッグ装置においては、エアバッグ16が運転手にとって上下及び左右方向（第2図の左右方向)に大きく且つ素早く展開することが好ましい。
MT出力	in the air bag apparatus for the driver seat , it is preferable that the air bag 16 is in a large and left direction (left and right in fig . 2) , and is rapidly deployed and developed .
BT-R スコア	34.7

（c）大規模語彙に対応した NMT モデル

入力日本語文	運転席用エアバッグ装置においては、エアバッグ16が運転手にとって上下及び左右方向（第2図の左右方向)に大きく且つ素早く展開することが好ましい。
フレーズトークンを含むMT出力	in the $T_1{}'$, it is preferable that the $T_2{}'$ 16 is larger and quickly in the $T_3{}'$ and $T_4{}'$ (the $T_5{}'$ of fig . 2) .
MT出力	in the driver 's seat airbag device , it is preferable that the airbag 16 is larger and quickly in the driver and lateral direction (the lateral direction of fig . 2) .
BT-R スコア	30.1

（d）大規模語彙に対応した NMT モデルとリランキングの併用

入力日本語文	運転席用エアバッグ装置においては、エアバッグ16が運転手にとって上下及び左右方向（第2図の左右方向)に大きく且つ素早く展開することが好ましい。
フレーズトークンを含むMT出力	in the $T_1{}'$, it is preferable that the $T_2{}'$ 16 is largely and quickly developed to the upper and lower $T_4{}'$ (the $T_5{}'$ of fig . 2) for the $T_3{}'$.
MT出力	in the driver 's seat airbag device , it is preferable that the airbag 16 is largely and quickly developed to the upper and lower lateral direction (the lateral direction of fig . 2) for the driver .
BT-R スコア	22.4

ていないにも関わらず、大規模語彙に対応した NMT モデルの段階で訳抜け単語数が減少している。大規模語彙に対応した NMT モデルでは、フレーズの一部としてそれらの訳抜け単語（の原言語単語）を抽出し、NMT のデコーディングの前にそれらのフレーズをトークンに置き換え、抽出されたフレーズは SMT によって翻訳された後、MT 出力に挿入される。大規模語彙に対応した NMT モデルにおいては、以上の仕組みにより訳抜け単語の数が抑えられる。

6　おわりに

本稿では、訳抜けした内容の削減という観点から、大規模語彙に対応した NMT モデル[5] の効果を検証した。日本語から英語への翻訳タスクにおいて、訳抜けした内容の自動検出方式の適用結果を示し、実際の訳抜け単語数の評価結果を示した。その結果、大規模語彙に対応した NMT モデルによって、ベースライン NMT モデルと比較して訳抜け単語数が減少することを示した。さらに、大規模語彙に対応した NMT モデルに対して逆翻訳確率に基づくリランキング方式[3] を適用することにより、BLEU において改善は見られないものの、訳抜け単語数がさらに減少することを示した。今後は、サブワード単位に基づく手法[8] との比較を行う。

参考文献

[1] Bahdanau, D., Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, in Proc. 3rd ICLR (2015)

[2] Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T. Toward the evaluation of machine translation using patent information. In Proc. 8th AMTA, pages 97-106, (2008)

[3] Goto, I. and Tanaka, H.. Detecting untranslated content for neural machine translation, In Proc. 1st NMT, pp. 47--55, (2017).

[4] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, in Proc. 45th ACL, Companion Volume, pp. 177-180 (2007)

[5] Long, Z., Utsuro, T., Mitsuhashi, T., and Yamamoto, M.: Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation, in Proc. 3rd WAT, pp. 47-57 (2016)

[6] Luong, M., Sutskever, I., Vinyals, O., Le, Q. V., and Zaremba, W.: Addressing the rare word problem in neural machine translation, in Proc. 53rd ACL, pp. 11-19 (2015)

[7] Nakazawa, T., Mino, H., Goto, I., Neubig, G.,

Kurohashi, S., and Sumita, E.: Overview of the 2nd Workshop on Asian Translation, in Proc. 2nd WAT, pp. 1–28 (2015)

[8] Sennrich, R., Haddow, B. and Birch, A. Neural machine translation of rare words with subword units. In Proc. 54th ACL, pp. 1715--1725, (2016) .