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

関沢 祐樹　　梶原 智之　　小町 守
首都大学東京
{sekizawa-yuuki, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

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、これまでの統計的機械翻訳と比べて流暢性の高い出力ができるという利点があり、ニューラル機械翻訳 [1] が活発に研究されている。しかし、ニューラル機械翻訳は語彙次元の分類問題を順番に解く生成タスクであり、出力層が高次元となり計算量が多いという課題がある。そこで、ニューラル機械翻訳では通常、出力層の語彙制限によって計算量を削減する。そのため、目的言語の語彙はトレーニングの際に高頻度語のみ（例えば上位 30,000 語 [1]）に制限され、その他の低頻度語は未知語（OOV）となり、まとめて"<unk>"などの特殊記号に置き換えられる。この OOV は意味を持たない記号であるため、出力文の内容語が OOV となることで妥当性が失われ、機能語が OOV となることで流暢性が失われる。

ニューラル機械翻訳の OOV の削減を試みる先行研究として、Mi ら [2] はトレーニングに使用する語彙を文ごとに選択することで、トレーニングの計算量を減少させ、全体の語彙を拡張した。しかし、この手法では翻訳のトレーニング方法を変更する必要がある。また、Luong ら [3] は OOV との対応関係にある原言語の単語を翻訳辞書を用いて直接翻訳する後処理を提案した。この手法では、トレーニングデータを用いて原言語と目的言語の単語アライメントを取る必要がある。さらに、Sennrich ら [4] は、系列に対するデータ圧縮手法である Byte Pair Encoding (BPE) を文字列に適用し、単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 OOV を削減した。この手法では、意味を考慮せずに単語を部分文字列に分解する。

本研究では、トレーニングデータにおいて目的言語の OOV に該当する低頻度語を同義な高頻度語に言い換えることによって、OOV へ翻訳する事例を削減する前処理を提案する。本手法の利点は以下である。

- 前処理である（トレーニング方法を変更しない）ため、任意のニューラル機械翻訳手法をブラックボックスとして適用できる。
- 対訳辞書や単語アライメントの必要がない。
- 低頻度語の意味を保ったまま目的言語の単語に翻訳される（OOV への翻訳事例が削減される）。

ASPEC の日英翻訳コーパスと Bahdanau ら [1] の Attention に基づくニューラル機械翻訳モデルをベースラインとして用いた評価の結果、提案手法が OOV の出現数を 17.3%削減し、BLEU を 0.08 ポイント改善することを確認した。

2 先行研究

ニューラル機械翻訳のトレーニング方法の変更によって OOV の削減を試み、翻訳の精度を向上させる先行研究が存在する。Jean ら [5] は、トレーニングにおいて対訳コーパスを分割し、分割された対訳コーパスを用いたトレーニングにおいて、使用する語彙を目的言語側の語彙からサンプリングし、得られた一部分の語彙を用いてトレーニングを行うことでトレーニングの計算量を減少させ、全体の語彙を広く取ることで OOV の削減を試みた。Mi ら [2] はトレーニングに使用する語彙を文ごとに選択し、トレーニングの計算量を減少させ、全体の語彙を拡張した。この手法では、翻訳前にあらかじめアライメントを取り、アラインされる単語単位の翻訳および、フレーズ単位の翻訳をトレーニング時に選択するため、計算量が減少する。Luong ら [6] は文字ベースの学習によって OOV を減少させた。これらの手法はトレーニング方法を変更する必要がある。本研究では、トレーニング方法を変更

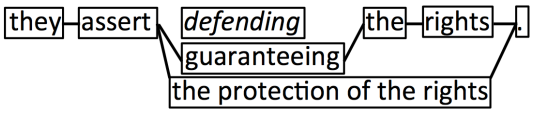

図 1: ビタビアルゴリズムによる言い換え例

せず、トレーニングデータにおける目的言語の語彙的言い換えによって前処理のみで翻訳結果の OOV を削減する。

一方、文の複雑さを削減するために機械翻訳の前処理において言い換えを行う先行研究も存在する。Štajner ら [7] は機械翻訳の前処理として入力文の語彙と文法を平易に言い換えた。本研究では、入力文のテキスト平易化を用いず、語彙の言い換えのみを用いて OOV の削減を試みる。また、トレーニングの前処理や後処理によって OOV の削減を試みる先行研究も存在する。Luong ら [3] は OOV との対応関係にある翻訳前の単語を翻訳辞書を用いて直接翻訳する後処理を提案した。この手法は、あらかじめ翻訳文対のアライメントを取る必要があり、翻訳辞書はトレーニングデータにおける単語アライメントの頻度によって構築されるため、低頻度語は原文表記のまま出力される。本研究ではアライメントを用いず、目的言語のみの言い換えによって OOV を削減する。Sennrich ら [4] は、系列に対するデータ圧縮手法である Byte Pair Encoding (BPE) を文字列に適用し、OOV を頻出する部分文字列の系列に分解して学習を行うことで OOV を削減した。この手法では、頻出するユニットの意味を考慮せず、貪欲に単語を分解する。本研究では、トレーニングデータにおける目的言語の語彙的言い換えを行うため、言い換え前後の意味を保持しつつ翻訳結果の OOV の削減が期待できる。提案手法は前処理なので、後処理と組み合わせることが可能であり、更なる性能改善が期待できる。

3 トレーニングデータの低頻度語の言い換え

本研究では、ニューラル機械翻訳の OOV を減らすために、トレーニングデータの目的言語文に存在する低頻度語を高頻度語に言い換えてから翻訳する手法を提案する。我々は言い換え対および言い換え確率が登録されている言い換え辞書を用いて 2 つのアプローチで低頻度語を高頻度語に繰り返し言い換える。まず 3.1 節では、妥当性（adequacy）を重視し、語句の言い換え確率を最大化する言い換えを行う。次に 3.2 節では、流暢性（fluency）を重視し、言語モデル確率を最大化する言い換えを行う。

3.1 言い換え確率を最大化する言い換え

この手法では、言い換え後の文の妥当性を重視して言い換え候補を選択する。トレーニングデータの目的言語側の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繰り返し言い換える。ただし、複数の言い換え候補が存在する場合、最も高頻度な候補を選択するのではなく、最も言い換え確率の高い候補を選択することで言い換え後の文の妥当性を高める。

以下は言い換えの例である。原文の低頻度語 quarrels は 1 回目の言い換えで高頻度語 discussions へと言い換えられる。また、低頻度語 pedagogues は 1 回目の言い換えで低頻度語 educators へと言い換えられ、2 回目の言い換えで高頻度語 teachers へと言い換えられる。低頻度語 pedagogues は言い換えを 1 度のみ行う場合、低頻度語 educators へは言い換えられない。

原文 :the *pedagogues* had *quarrels*.
1 回目の言い換え:the *educators* had *discussions*.
2 回目の言い換え:the *teachers* had *discussions*.

3.2 言語モデル確率を最大化する言い換え

この手法では、言い換え後の文の流暢性を重視して言い換え候補を選択する。トレーニングデータの目的言語側の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。ただし、複数の言い換え候補が存在する場合、最も言い換え確率が高い候補を選択するのではなく、最も言語モデル確率の高い候補を選択することで言い換え後の文の流暢性を高める。ここで、ある文には複数の低頻度語が存在し得るので、ビタビアルゴリズムによって効率的に 2-gram 言語モデル確率を最大化する言い換え文を選択する。

ビタビアルゴリズムによる言い換えの例を図 1 に示す。原文 “they assert defending the rights.” において、defending が OOV である。defending は高頻度語である guaranteeing への言い換えが可能であり、defending the rights は全て高頻度語である the protection of the rights への言い換えが可能である。この例では、 “assert guaranteeing the rights .” の 2-gram 言語モデル確率、 “assert the”、“rights .” の

表 1: 提案手法の日英翻訳結果（括弧内はテスト文を翻訳した出力文に存在する OOV の数）

手法	言い換え確率	選択方法 LM-Giga	選択方法 LM-ASPEC	トレーニングデータの低頻度語数
Bahdanau+		20.63 (1,489)		474,468
1 回のみ（語）	20.55 (1,240)	19.62 (1,350)	20.49 (1,338)	383,715
2 回まで（語）	20.61 (1,301)	20.24 (1,311)	**20.71** (1,231)	377,369
無制限（語）	20.28 (1,322)	19.21 (1,196)	18.23 (1,229)	377,018
1 回のみ（語＋句）	20.11 (1,274)	19.24 (**1,194**)	17.89 (1,451)	383,618
2 回まで（語＋句）	19.29 (1,408)	18.83 (1,379)	18.38 (1,442)	377,306
無制限（語＋句）	19.61 (1,324)	18.74 (1,331)	18.65 (1,327)	**376,955**

2-gram 言語モデル確率を計算し、最も高い言い換えを選択し “they assert the protection of the rights.” が生成される。

この手法ではフレーズの言い換えの際、フレーズ外の言語モデル確率は計算するが、フレーズ内の言語モデル確率は計算しない。図 1 の例では “assert defending” や “assert the” の言語モデル確率を計算し、フレーズ “the protection of the rights” の言語モデル確率は計算しない。

4 実験

4.1 実験設定

本研究では、アジア学術論文抜粋コーパス (ASPEC) 日英対訳データを使用した。トレーニングにはアライメント確度の高い 100 万文のうち、文長 40 単語以下の文 827,503 文を使用し、チューニングには 1,790 文対すべてを、テストには 1,812 文対すべてを使用した。

提案手法では、言い換え辞書と言語モデルを用いて言い換えを行う。言い換え辞書には PPDB [8] を使用した。言語モデルには KenLM[1] を用いて、2 種類（トレーニングデータの目的言語側および English Gigaword Fifth Edition[2]）の 2-gram 言語モデルを構築した。機械翻訳は、NMTkit[3] を用い、Bahdanau ら [1] のアテンションを用いたニューラル機械翻訳（これを baseline とする）を使用し、入力語彙数、出力語彙数は共に 30,000 とした。翻訳の評価には BLEU を用いた。また、翻訳後に現れる OOV の数の変化による評価を行った。

[1] http://kheafield.com/code/kenlm/
[2] https://catalog.ldc.upenn.edu/LDC2011T07
[3] https://github.com/odashi/nmtkit

4.2 実験結果

各手法ごとの結果を表 1 に示す。最も BLEU スコアが高かったものは、ASPEC を言語モデルとして使用し、低頻度語を高頻度語へ 2 回まで言い換える手法であった。この手法では、baseline と比較して BLEU スコアが 0.08 ポイント向上し、出力文に存在する OOV が 17.3%減少した。

5 考察

まず、単語単位の言い換えにおいて、BLEU スコアに注目すると、1 回のみの言い換えよりも 2 回までの言い換えを行った方が BLEU スコアが向上した。しかし、無制限に言い換えを繰り返しても、BLEU スコアがさらに改善されるわけではなかった。言い換えとはいえ、元の表現の意味を完全に保持できるとは限らないので、複数回言い換えを繰り返すことで意味の異なる表現に変換される可能性がある。そのため、言い換えによる意味のずれと、高頻度語への言い換えによる OOV 削減のバランスのとれた 2 回までの言い換えが最も BLEU スコアを改善したと考えられる。

次に、OOV の数に注目すると、言い換え回数を増やすほど、トレーニングデータ中の OOV は削減されている。しかし、トレーニングデータ中の OOV の減少に伴って、翻訳結果の OOV も削減されるわけではなかった。これは、言い換えを繰り返した結果、意味が保持されない変換や品詞が異なる変換を行った場合、ニューラル機械翻訳が出力文の流暢性を担保するために言い換え結果を出力しないためだと考えられる。

また、単語のみの言い換えが、句の言い換えを含めた場合よりも BLEU スコアが高い傾向がある。これは句の内部の言語モデル確率を考慮していないため、流暢性を損なう言い換えが行われた可能性がある。

表 2: 翻訳例（提案手法は ASPEC 言語モデルを使用）

手法	翻訳
reference	ozone formation increased about 2mg / h .
baseline	the amount of ozone generation increased by about “OOV” / h .
2 回まで（語）	the ozone generation increased by about 2 mg / h .
2 回まで（語＋句）	the amount of ozone generation was about 2 mg / h .
reference	the optical switching of the title and its optical recording image were formed , and the stability was examined .
baseline	the “OOV” and “OOV” images were formed , and their stability was investigated .
2 回まで（語）	the optical switching and optical recording images were formed , and the stability was examined .
2 回まで（語＋句）	the “OOV” optical switching and optical recording images were formed and their stability was examined .
reference	modeling a dentin resin impregnated layer structure showed the relation between hardness and elastic modulus .
baseline	the “OOV” resin agglomerate layer was modeled and the relationship between the hardness and the elastic modulus was found .
1 回まで（語）	the “OOV” resin impregnated layer structure was modeled and the relationship between hardness and modulus was found .
2 回まで（語）	the model for the dentin resin was used to model the structure of the dentin resin , and the relationship between the hardness and the elastic modulus was found .
無制限（語）	the authors have modeled the cross-sectional structure of the resin-impregnated resin layer and the relationship between hardness and elastic modulus was found .

表 2 は実際の翻訳例である。一つ目の例は、baseline が低頻度語 2mg を OOV として出力しているが、提案手法によってトレーニングデータ中で 2mg がそれぞれ高頻度な 2 と mg に言い換えられた結果、OOV ではなく妥当性の高い出力が得られた。二つ目の例は、単語のみの言い換えが、句の言い換えを含めた場合よりも翻訳が良い例である。三つ目の例は、単語単位の言い換えにおいて言い換え回数を増やすことで OOV が削減される例である。

6 おわりに

本研究では、ニューラル機械翻訳の OOV を減らすために、あらかじめトレーニングデータの目的言語側に存在する低頻度語を高頻度語に言い換えた。ASPEC の日英翻訳コーパスを用いた評価によって、翻訳結果の OOV の数が減少し、BLEU スコアが向上することが確認できた。この手法はニューラル機械翻訳に限定されず、語彙次元の分類問題を解く文圧縮や対話など多くのニューラルネットワークを用いる生成タスクにおいて有効であると考えられる。

今後は、言い換え確率と言語モデル確率の両方を組み合わせて適切な言い換えを選択したい。また、参照訳に低頻度語が含まれるとき、本研究で出力する高頻度な同義表現は表層では一致せず、BLEU では適切に評価できない場合があるため、人手評価によって妥当性と流暢性を評価したい。

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*, 2015.

[2] Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. Vocabulary manipulation for neural machine translation. In *Proc. of ACL*, pp. 124–129, 2016.

[3] Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proc. of ACL-IJCNLP*, pp. 11–19, 2015.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proc. of ACL*, pp. 1715–1725, 2016.

[5] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Proc. of ACL-IJCNLP*, pp. 1–10, 2015.

[6] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proc. of ACL*, pp. 1054–1063, 2016.

[7] Sanja Štajner and Maja Popovic. Can text simplification help machine translation? *Baltic Journal of Modern Computing*, Vol. 4, No. 2, pp. 230–242, 2016.

[8] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. of ACL*, pp. 425–430, 2015.