ニューラル機械翻訳における低頻度語処理

今仁優希 [*1]　村上仁一 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
{s142006,murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳の手法として，これまで，ルールベース翻訳，用例翻訳，統計翻訳などが提案されてきた．近年では，新たな機械翻訳の手法としてニューラル機械翻訳 (Neural Machine Translation; NMT) が注目されている．NMT の手法は，これまで提案された他の手法と比較して流暢性の高い翻訳文を生成することができると報告されている．

一方で NMT の手法に関して，語彙数の問題が指摘されている．NMT の手法において，ニューラルネットワークの出力層の次元数は NMT の語彙数に相当し，語彙数が多くなると計算量が膨大になる [1]．このため，計算量を削減する目的で，一部の単語を未知語として特別な記号 (<unk>) に置換することで，語彙数を制限する手法が用いられる．この際，対訳学習文中に出現する頻度の低い単語 (以下，低頻度語) を未知語とするのが一般的である [2]．この手法の問題として，低頻度語は全て未知語となり学習されないため，翻訳時，低頻度語を含む文が入力された際に誤った翻訳が出力される場合がある．

この問題を緩和する一つの方法として，計算量を犠牲にして語彙数を制限せずに学習を行う方法がある．しかし，この方法を用いた場合，計算量が増大する問題に加えて，翻訳精度の問題が生じる可能性がある．低頻度語は統計的信頼性が低く，また対訳学習文中に出現する語彙の多くを占めている. このため, 低頻度語を含む文の学習が NMT のシステム全体の翻訳精度の低下を招く可能性がある．

本研究では，まず学習時に，語彙数制限処理として対訳学習文中に 1 回のみ出現する単語 (以下，頻度 1 単語) を全て未知語として <unk> に置換し，学習を行う．これは，頻度 1 単語が低頻度語の中でも特に統計的信頼性が低いと考えられるためである．次に翻訳時，Jean ら [3] による <unk> の置換処理を行い，出力文中の <unk> を Attention 確率が最も高い原言語単語に置き換える．最後に未知語処理として出力文に含まれる未知語を対訳学習文と IBM Model 1 により作成した対訳単語辞書を用いて置換する [4] 手法を提案する．

結果として，提案手法を用いることで，語彙数を制限しない未知語処理の方法と比較して自動評価及び人手評価共に翻訳精度の向上が確認できた．

2 ニューラル機械翻訳

ニューラル機械翻訳 (NMT) の手法では，まず対訳学習文を用いてニューラルネットワークによるモデル (以下，NN モデル) の学習を行う．その後，学習した NN モデルを用いて入力文を翻訳し，出力文を生成する．

2.1 学習

NMT のシステムでは入力文中の原言語単語と出力文中の目的言語単語との対応が Attention 確率 [2][5] を用いて学習される．図 1 に NMT における学習の枠組みを示す．NMT では学習時，対訳学習文の原言語文及び目的言語文を用いて Attention 確率を求め，求めた確率の重みが NN モデルに保存される．一般的な NMT の手法では，語彙数を制限するため，学習時に低頻度語は未知語として <unk> に置換され，システムに入力される．

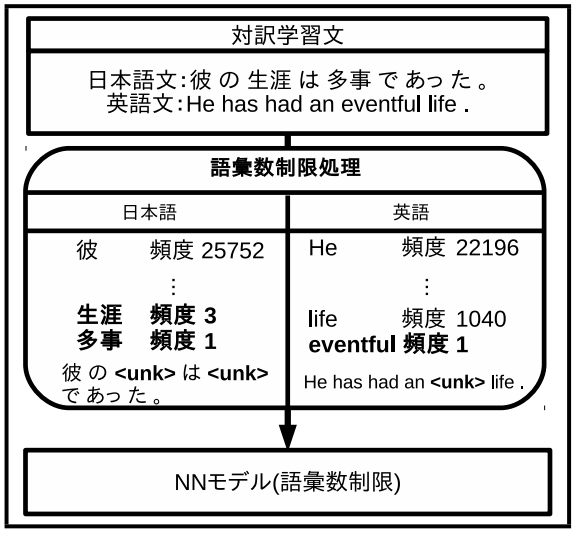

図 1　NMT の学習

2.2 翻訳

NMT の翻訳では，学習によって得られた NN モデルを用いて入力文から出力文を生成する．入力文のそれぞれの単語について Attention 確率が計算され，最終的な出力文が生成される．この際，入力文に低頻度語や未知語が含まれる場合などに，<unk> を含む出力文が生成される．

2.3 出力文中の未知語の原言語単語への置換

Jean ら [3] は Attention 確率を応用して，出力文中の <unk> 記号を Attention 確率が最も高い入力文中の原言語単語に置換する手法を提案した．この手法により，固有名詞等の原言語と目的言語の両方に共有される単語を未知語とすることなく翻訳することが可能となり，翻訳精度の向上が報告されている．この手法は出力の <unk> 記号を全て Attention 確率の最も高い原言語単語に置き換えて文を生成するため，表 1 に示すように出力文に両言語に共有される単語以外の未知語も含まれる．

表 1　Attention 確率を利用した <unk> 記号の原言語単語への置換処理結果の例

入力文	空 に 入道雲 が むくむく 出 て き た 。
出力文	A <unk> is rising in the sky .
置換	A 入道雲 is rising in the sky .

3　従来手法:語彙数無制限

入力の低頻度語を扱う従来の手法として，学習時に低頻度語の <unk> への置換処理を行わず，学習の語彙数を無制限とする方法がある．語彙数を無制限とし，低頻度語を含む対訳学習文中の全語彙を学習することで，翻訳時に入力の低頻度語に対して正しい出力が得られる可能性がある．一方で，低頻度語は統計的信頼性が低く，対訳学習文中の多くの語彙を占めるために，学習することで NMT のシステム全体の翻訳精度の低下を招くおそれがある．また，語彙数はニューラルネットワークの出力層の次元数に相当し，語彙数が多くなることで計算量が多くなる問題も生じる．

本研究では，語彙数を無制限とする手法を従来手法としてベースラインの低頻度語処理の手法とする．

4　提案手法

従来手法について，低頻度語を含む全語彙を学習するために NMT のシステム全体の翻訳精度の低下を招く可能性を指摘した．提案手法では，統計的信頼性の低い低頻度語を学習せず，出力文中の未知語を対訳単語辞書を用いて翻訳する方法により翻訳精度の向上を試みる．

提案手法では，図 2 に示す学習の過程の後，図 3 に示す翻訳及び未知語処理の過程を行う．

4.1　学習

学習の過程では対訳学習文から NN モデル及び対訳単語辞書を学習する．第 2.1 節と同様に，対訳学習文から確率の重みを求めて，NN モデルに保存する．この際，語彙数制限処理として，対訳学習文における日本語文及び英語文に含まれる頻度 1 単語を <unk> に置換し，システムに入力する．これにより，モデルの学習する語彙数を頻度 1 単語を除いた語彙数に制限する．

対訳単語辞書は対訳学習文と IBM Model 1 を用いて学習する．

4.2　翻訳及び未知語処理

翻訳及び未知語処理の過程ではまず，学習の過程で得られた NN モデルを用いて入力文から出力文を生成する．NN モデルは語彙数が制限されているため，翻訳時，入力文に低頻度語や未知語が含まれる場合などに，出力文中に <unk> が生成される．<unk> を含む出力文には，第 2.3 節で説明した <unk> の原言語単語への置換処理を行う．

最後に未知語処理として，入力文中の原言語単語 (未知語) を含む出力文に対して，未知語を対訳単語辞書から検索し，翻訳確率最大となる訳語を選択する [4]．訳語の選択が成功した場合は，その訳語を元の出力文の未知語部分に置換し，最終的な出力文を生成する．

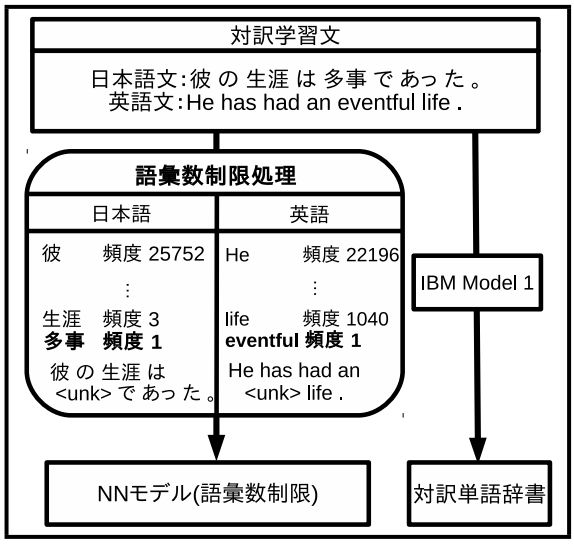

図 2　提案手法における NN モデル及び対訳単語辞書の学習

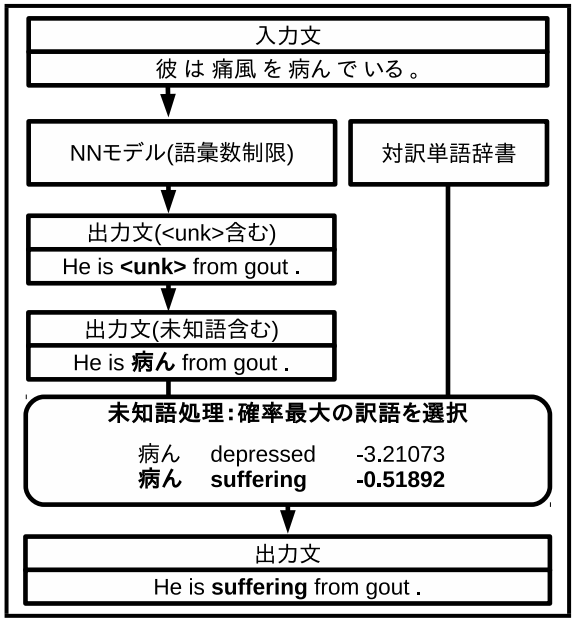

図 3　提案手法における翻訳及び未知語処理

5　実験環境

実験では日英ニューラル機械翻訳の学習及び翻訳を行う．従来手法として語彙数を無制限とする方法を，提案手法と比較する．

5.1　実験データ

本実験には，電子辞書などの例文より抽出した単文コーパス [6] を用いる．使用するデータの内訳を表 2 に示す．

表 2　実験データ

対訳学習文	160,000 文
ディベロップメント文	1,000 文
入力文	1,000 文

学習は対訳学習文を用いて行い，翻訳は入力文に対して行う．ディベロップメント文は BLEU スコアが高いモデルを選択するために用いる．

5.2 実験設定

NMT のツールキットには OpenNMT [7] を用い，モデルは Luong らにより提案された Global Attention を用いる. Encoder，Decoder の LSTM は 2 層とし, ユニット数は 500，単語の分散表現のベクトルサイズは 500 を設定する．ミニバッチサイズは 40 とし，モデルの訓練は最大 32 エポック行う．

5.3 評価方法

本研究では，従来手法及び提案手法の出力文の翻訳精度の比較評価を行う．翻訳精度の評価には人手評価と自動評価の 2 種類の評価を行う. 人手評価では, 従来手法の出力文と提案手法の出力文について, 正確性 (adequacy: 入力文の意味をどれだけ正確に翻訳英文から読み取れるか) に基づいて対比較評価を行う. 自動評価には BLEU, METEOR，RIBES，及び TER の指標を用いる．

6 実験結果

6.1 学習時の語彙数

従来手法と提案手法の学習時の異なり語彙数の比較を表 3 に示す．提案手法では，頻度 1 単語処理として対訳学習文における日本語文及び英語文のそれぞれについて，頻度 1 単語を <unk> 記号に置換する処理を行ったため，学習の語彙数が減少している．

表 3　従来手法と提案手法の学習時の語彙数

	語彙数 (日)	語彙数 (英)
従来手法	42,760	45,637
提案手法	28,412	25,839

6.2 未知語数

従来手法及び提案手法の出力について, 未知語を含む文数の調査を行った. 調査結果を表 4 に示す.

表 4　従来手法と提案手法の未知語を含む文数

従来手法	0 文
提案手法 (未知語処理前)	213 文
提案手法 (未知語処理後)	82 文

表 4 から, 提案手法により未知語処理に成功した文数は 131 文であったことが確認できる.

6.3 出力文の翻訳結果

提案手法における出力文の翻訳精度を従来手法の出力文と比較した．比較方法として，人手評価と自動評価を行った．以下に評価結果を示す．

6.3.1 提案手法における人手評価結果

従来手法と提案手法の出力文から，それぞれランダムに抽出した 100 文を用いて，人手による対比較評価を行った．評価の基準を以下に示す．評価結果を表 5 に示す．また，評価及び出力例は第 7 章において示す．

- 提案手法○　：提案手法の方が正確性が高い
- 提案手法×　：提案手法の方が正確性が低い
- 差なし　　：翻訳精度に明確な差がない

表 5　従来手法と提案手法の評価結果 (100 文中)

提案手法○	提案手法×	差なし
36 文	18 文	46 文

6.3.2 自動評価結果

単文 1,000 文を入力として翻訳実験を行い，出力文に対して自動評価を行った．表 6 に，それぞれの手法における自動評価の結果を示す．

表 6　自動評価結果. 精度が高い方を太字で示す

翻訳手法	BLEU	METEOR	RIBES	TER
従来手法	0.1937	0.4756	0.7813	0.5953
提案手法	**0.2018**	**0.4894**	**0.7866**	**0.5793**

表 6 より，実験に用いた全ての自動評価指標において提案手法が従来手法より高い翻訳精度となった．

6.4 実験結果のまとめ

表 5 と表 6 より，提案手法は従来手法と比較して翻訳精度の向上が確認できる．

7 考察

7.1 未知語を含まない文

提案手法の未知語処理前の出力文において，未知語を含まない文は 787 文あった. これらの文について，提案手法○及び提案手法×の場合の例を示す .

表 7 において，入力文の “考察している” に対する出力が提案手法では “is considering” と比較的正しい訳になっている．しかし従来手法では “has revolutionized” となっており，誤った訳が出力されている．この原因として，従来手法では NN モデルが学習する語彙数が大きすぎるために，出力単語の確率的選定の精度が低下していることが推察される．

表 7　未知語を含まない文:提案手法○の出力例

入力文	本書 は きわめて 重大 な いくつ か の 問題 を 考察 し て いる 。
参照文	The book discusses some vital issues .
従来手法	This book has revolutionized several important problems .
提案手法○	This book is considering several serious questions .

表 8 において，提案手法の出力文では入力文の意味を大きく損失しているのに対し，従来手法の出力文では比較的正確な訳となっている．この例では入力文の “状勢” が頻度 1 単語であり，頻度 1 単語を学習している従来手法では正しい訳が出力されたと考えられる．提案手法の出力文に入力文中の原言語単語 (未知語) が出力されず，対訳単語辞書を用いた翻訳が行われなかったことも誤りの一因と考えられる．

表 8　未知語を含まない文:提案手法×の出力例

入力文	これ で 状勢 は すっかり 変わっ た 。
参照文	This has completely altered the situation .
従来手法	This transformed the state of affairs .
提案手法×	This is quite changed .

7.2 未知語処理に成功した文

提案手法の未知語処理前の出力文において，未知語を含む文は 213 文あった．そのうち，未知語処理に成功した文は 131 文あった．これら 131 文と従来手法の出力文との人手による対比較評価を行った．評価基準は第 6.3.1 項に示したものと同様である．表 9 にその結果を示す．

表 9　従来手法と提案手法の評価結果 (131 文中)

提案手法○	提案手法×	差なし
56 文	30 文	45 文

表 9 より，未知語処理に成功した文においても，従来手法と比較して提案手法の方が翻訳精度が高いという結果になった．これより，提案手法の未知語処理が，翻訳精度の向上に比較的有効であることがわかる．

また，提案手法○の場合の出力結果を表 10 に示す．提案手法の未知語処理前の出力文の未知語 “香辛料” が対訳単語辞書を用いた未知語処理によって “ spices” と置換されている．未知語処理後の出力文は，文法的な正しさを多少欠いているが入力文の意味を概ね推測できる文と考えられる．従来手法の出力文では “香辛料のきいた食べ物” が “sacred food” と訳されている．文法的には提案手法よりも正しいが，入力文の意味と大きく異なっている．

この例のように，提案手法の未知語処理を用いると，出力文の文法性 (あるいは流暢性) が従来手法より多少低いものの，正確性が高い出力文が得られる傾向にあった．

表 10　未知語処理に成功した文:提案手法○の出力例

入力文	香辛料 の きい た 食べ物 が 好き だ 。
参照文	I love spicy food .
従来手法	I like the sacred food .
提案手法 (未知語処理前)	I like the 香辛料 food .
提案手法○ (未知語処理後)	I like the spices food .

7.3 対訳単語辞書の問題

表 9 において，対訳単語辞書の問題により提案手法×となった文が存在した．その例を表 11 に示す．従来手法は “歌謡コンテスト” を “music contest”，提案手法は “entry contest” と訳しており，従来手法の方が入力の意味により近いと評価できる．

この例から，対訳単語辞書の精度の問題が指摘できる．提案手法で用いている対訳単語辞書において，“歌謡” に対する訳語が “entry” と誤った対応になっている．提案手法において，このような対訳単語辞書の誤りが出力文全体の翻訳精度の低下を招いている例は 131 文中 16 文存在する．これら 16 文のうち 15 文は表 9 において提案手法×もしくは差なしと評価されている．

このことから，提案手法において対訳単語辞書の精度が一つの課題であると考察できる．

8 おわりに

本研究では，NMT の翻訳精度の向上を目的とする手法を提案した. 提案手法は, 学習時に頻度 1 単語を <unk> に置換して学習し，翻訳時に Jean らによる <unk> の置換処理を行った後，出力文に含まれる未知語を IBM Model 1 により学習した対訳単語辞書を用いて処理する手法である．結果として，出力文 100 文における人手評価では提案手法○が 36 文，提案手法×が 18 文となった．これより，提案手法を用いた場合，従来手法と比較して翻訳精度が向上することが確認できる．また，表 5，表 6 より，本研究の実験結果は，低頻度語の学習が NMT のシステム全体の翻訳精度の低下を招くという仮説を裏付ける結果となっていることがわかる．

表 11　提案手法において対訳単語辞書の問題を含む文

入力文	歌謡 コンテスト で わたし は １０ 万 円 の 賞金 を もらっ た 。
参照文	I won one hundred thousand yen in prize money in the singing contest .
従来手法	In the music contest I won a prize of 100,000 yen .
提案手法 (未知語処理前)	I won a prize of 100,000 yen in the 歌謡 contest .
提案手法× (未知語処理後)	I won a prize of 100,000 yen in the entry contest .

今回の実験から提案手法において対訳単語辞書の翻訳精度に課題があることがわかった．また，未知語処理が成功しなかった文が 82 文存在した．今後は対訳単語辞書の翻訳精度を向上させる手法や，より多くの未知語を処理する手法を検討したい．

9 謝辞

本研究における提案手法の未知語処理システムは同研究室の川原氏 [4] の尽力により実現されたものです．本研究で用いている NMT のシステムへの適応にも多大なお力添えをいただきました．心より感謝申し上げます．

参考文献

[1] 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, pp. 982–985, 2017.

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

[3] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. *In ICML*, 2014.

[4] 川原宰, 村上仁一. 日英翻訳における IBM Model 1 を用いた未知語処理. 言語処理学会 第 24 回年次大会, 2018.

[5] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *In Proceedings of EMNLP*, pp. 1412–1421, 2015.

[6] 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.

[7] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation, 2017.