2022年度(令和4年度)　修士論文

ニューラル機械翻訳における
類語置換を用いた低頻度語処理

鳥取大学大学院　持続性社会創生科学研究科
工学専攻　情報エレクトロニクスコース

自然言語処理研究室

M21J4062Z 矢野　貴大

概要

機械翻訳の手法の一つに, ニューラル機械翻訳 (以下 NMT)[1] がある. NMT とは, ニューラルネットワークに基づき機械翻訳を行う手法で, 大きく分けて学習と翻訳の 2 つの過程からなる. 学習の過程では, 日本語文と英語文の対訳を学習データとして, 文中に出現する語句の対応関係を学習し, 翻訳モデルを作成する. 翻訳の過程では, 学習の過程で生成した翻訳モデルを用いて, 翻訳を行う. NMT は, 従来の統計翻訳などの手法と比較して流暢性の高い翻訳が可能であると報告され, 現在, 最も主流な機械翻訳手法となっている.

しかし一方で, NMT には翻訳精度の向上を目的とする上で問題となる点が複数存在する. そのうちの一つに, 低頻度語を正しく翻訳できないという問題点がある.

本研究では, この低頻度語に関する問題に着目した. 低頻度語の中でも, 人名は対訳が取ることが比較的容易である点を活かし, 低頻度の人名を含む文に対して, 置換を用いて翻訳する手法を提案する.

実験では, 従来手法と提案手法とで翻訳精度の比較を行った. 実験の結果, 提案手法では従来手法よりも, 低頻度語に対して頑強となり, 翻訳精度も向上することが分かった.

目 次

図 目 次

表 目 次

第1章　はじめに

機械翻訳の手法の一つに, ニューラル機械翻訳 (以下 NMT) がある. NMT とは, ニューラルネットワークに基づき機械翻訳を行う手法で, 大きく分けて学習と翻訳の 2 つの過程からなる. 学習の過程では, 日本語文と英語文の対訳を学習データとして, 文中に出現する語句の対応関係を学習し, 翻訳モデルを作成する. 翻訳の過程では, 学習の過程で生成した翻訳モデルを用いて, 翻訳を行う. NMT は, 従来の統計翻訳などの手法と比較して流暢性の高い翻訳が可能であると報告され, 現在, 最も主流な機械翻訳手法となっている.

しかし一方で, NMT には翻訳精度の向上を目的とする上で問題となる点が複数存在する. そのうちの一つに, 低頻度語を正しく翻訳できないという問題点がある.

低頻度語にはいくつか種類が存在する. 低頻度語のうちの一つに, 人名がある. 入力文に低頻度の人名が含まれていると, 翻訳精度が低下する. しかし一方で, 低頻度の人名を類似の頻出語と置き換えると, 正しく翻訳できることがある.

そこで, 本研究では翻訳時に低頻度の人名を類似の頻出語と置き換えて翻訳し, その後出力文を整形するという手法を提案する. 提案手法を用いることで低頻度の人名を含む文の翻訳精度を向上することができると考える.

第2章　従来研究

本章では, NMT と NMT における低頻度語の問題について説明する. また, これまでの低頻度語の処理に関連した研究について紹介する. また, 提案手法で用いたツールについての紹介を行う.

2.1　NMT

ニューラル機械翻訳 (以下 NMT)[1] は, 機械翻訳の一種である. ニューラルネットワークに基づき機械翻訳を行う手法である. 従来の統計翻訳などの手法と比較して流暢性の高い翻訳が可能であると報告され, 現在, 最も主流な機械翻訳手法となっている.

NMT の動作原理について説明する. 大きく分けて, 学習と翻訳の 2 つの過程からなる.

学習の過程では, 日本語文と英語文の対訳を学習データとして, 文中に出現する語句の対応関係を学習し, 翻訳モデルを作成する. 例を図 2.1 に示す.

翻訳の過程では, 学習の過程で生成した翻訳モデルを用いて, 翻訳を行う. なお, NMT は学習する対訳文の言語によって様々な言語間の翻訳が可能であるが, 本研究では日本語から英語の翻訳を行う. また, NMT の実装には様々なツールがあるが, 本実験では OpenNMT-py[2] を用いる.

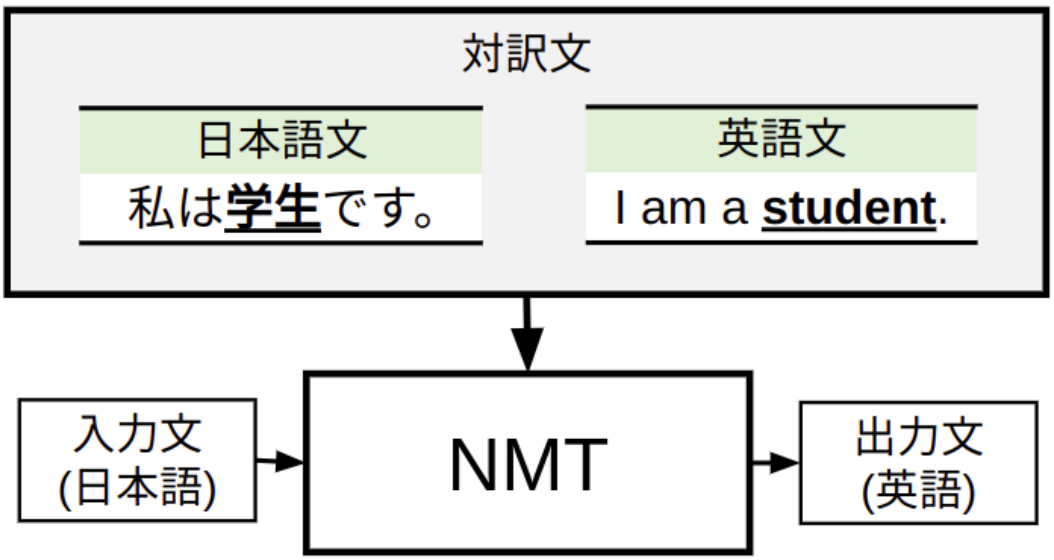

図 2.1: NMT の学習

2.1.1 NMT の問題点

NMT には翻訳精度の向上を目的とする上で問題となる点が複数存在する. そのうちの一つに, 学習文において出現が少ない, 低頻度語を正しく翻訳できないという問題点がある.

NMT は日英の対訳の語句の対応をもとに学習を行う. しかし, 学習文中に一度もでていない, または出現数が少ない語の翻訳精度は大きく減少してしまうことが分かっている. この事例について, 以下の 2 つの具体例をもとに説明する.

- 全く異なる訳語の生成

表 2.1: 低頻度語の翻訳例 1

入力文	法然 はこの地方に仏教を広めた.
翻訳文	Fujii spread Buddhism in this area .

低頻度語の翻訳の際, 低頻度語の訳にあたる部分が全く異なる訳語になることがある.

上記の例では「法然」が学習文に一度も登場していない語である. 出力文と比較すると分かるように, 翻訳文では法然に対応する訳が「Fujii」となっている. 法然と対応するのは正しくは Honen で, 全く異なる訳語が生成されていることがわかる.

- 支離滅裂な文の生成確率の増加

表 2.2: 低頻度語の翻訳例 2

入力文	クラーク は 精神 的 に 参って い た .
翻訳文	The sensation was spiritual .

低頻度語の翻訳の際, 翻訳が全体的に入力文と対応が取れないような, 支離滅裂な文になることがある. 原因として, 学習文に登場する回数の少なさから誤った学習を行い, それが全体の翻訳に悪影響を及ぼしているものと考えられる.

上記の例では, クラークが頻度 1 の低頻度語である. 出力文の The sensation was spiritual . は, 直訳すると「その感覚はスピリチュアルなものだった。」であり, 入力文とは大

きくかけ離れた翻訳となっている. このように, 低頻度語が原因で翻訳精度の低下につながる場合がある.

2.2 低頻度語の対策を目的とした研究

低頻度語の対策を目的とした研究は, 機械翻訳システム自体を改良する手法と, 機械翻訳システムの入出力を編集する手法の 2 つに大別できる. 機械翻訳システムの入出力を編集する手法の場合, 機械翻訳システム自体を改良する手法に比べ, 入出力に対する処理であるため任意のニューラル機械翻訳に適用できるというメリットが存在する.

2.2.1 (今仁, 2018) の研究 [3]

提案手法について説明する. NMT の学習では, 学習文における低頻度語をすべて<unk>と置換して学習を行う. また, 翻訳の過程では 2 つのことを行う. まず出力文に対し, 文中に含まれる<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.

実験の結果, 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.

2.2.2 (岩山,2017) の研究 [4]

低頻度語の一つに, 数詞がある. 数詞は表現が多様であるため, 低頻度語になることが多い. そこで岩山は, 中日翻訳において, 特許文における数詞の翻訳精度の改善を目的として研究を行った. 提案手法では, まず学習文の中の数値表現をすべて<num>に変換してから学習する. また, 翻訳時も, 文中の数値表現をすべて<num>に変換してから翻訳する. 最後に, 翻訳文中の<num>をアテンション機構を用いて元の数値表現に変換する. 実験の結果, 従来法と比べて全体で 1 ポイント, 数詞を含む文で 3.8 ポイントの BLUE 値の向上がみられた.

2.3 固有表現抽出 (NER)

固有表現抽出 (Named Entity Recognition, NER) は, 文書やテキスト内の人名, 組織名, 地名など, 固有の意味を有する単語のグループを抽出するタスクである. これは自然言語処理において重要なタスクの一つであり, 文書の分類, 概念マッピング, イベント抽出などの様々なタスクに利用される. 固有表現抽出は, 手作業で行うことも可能だが, 多くの場合は深層学習による自動的な手法が用いられる.

固有表現抽出を行うツールには複数の種類が存在するが, 本研究では mecab[5] を使用する.

2.3.1 mecab を用いた固有表現抽出

図 2.2 は mecab を用いて, 入力文「法然はこの地域に仏教を広めた。」から固有表現抽出を行った例である. 実行の結果, 固有名詞, 人名として法然が抽出できていることがわかる.

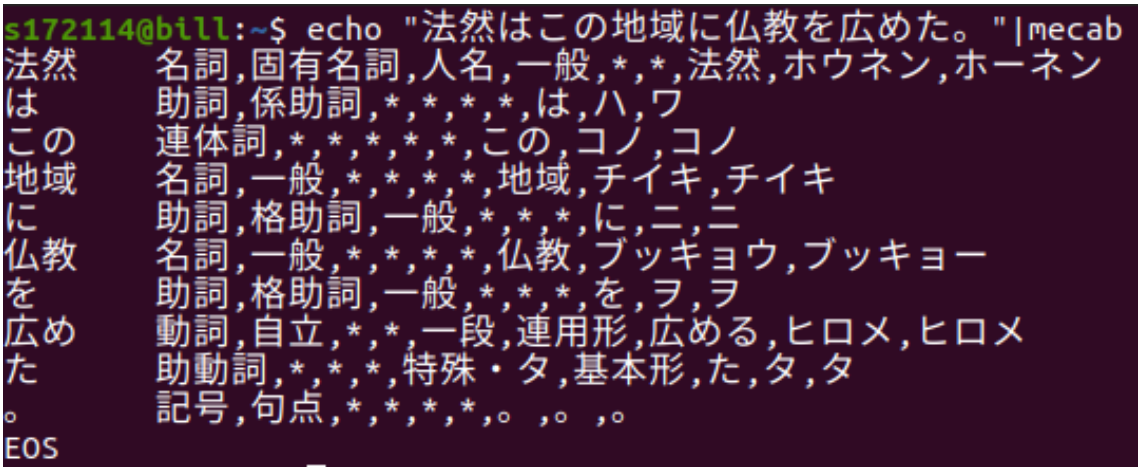

図 2.2: mecab を用いた固有表現抽出の例

2.4　対訳辞書抽出 (BLE)

BLE (Bilingual Lexicon Extraction) は, 二つの言語間の対応する語彙のペアを抽出するタスクである. このタスクは翻訳システムやマルチリンガルな文書処理システムの構築に必要な語彙リソースの構築に役立つ.

対訳辞書抽出を行うツールとして, 本研究では word2word[7] を用いる.

2.4.1　word2word

word2word[7] は, 対訳から対訳単語の対を抽出するための python 製のツールである. 62 言語対に対応し, 3564 の言語ペアにおいて言語間の対訳を取得することが可能である. 本研究では, 日英における対訳辞書の取得に活用した.

2.4.2　word2word を用いた対訳単語の取得

図 2.3 は word2word を用いて, 入力「クラーク」に対して対訳辞書抽出を行った例である. 実行の結果, クラークの対訳辞書として, 「Clark」が得られた.

```
word2words172114@bill:~/experiment/word2word$ python
Python 3.8.10 (default, Nov 14 2022, 12:59:47)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more informa
tion.
>>> from word2word import Word2word
>>> ja2en = Word2word("ja", "en")
>>> s = input()
クラーク
>>> print(ja2en(s))
['Clark', 'Clarke', 'Kent', 'Inspector', 'Heather']
>>>
```

図 2.3: word2word を用いた対訳単語の取得の例

2.4.3 uconv

uconv[6] は異なる文字エンコーディング間でテキストファイルを変換するコマンドラインツールである. このツールは, 対訳辞書抽出に用いられるものではない. しかし, 人名など, ローマ字変換することで対訳となる語彙に限れば擬似的に対訳辞書抽出が可能であると考えた.

2.4.4 mecab と uconv を用いた対訳単語の取得

uconv を用いることで, ひらがなからローマ字への変換が可能である. しかし, 漢字からローマ字への直接変換は不可能である. そこで, mecab で漢字からひらがな変換ができる機能を活用する. mecab と uconv を組み合わせることで漢字からもローマ字への変換を可能とする.

図 2.4 は word2word を用いて, 入力「徳川 家光」に対してローマ字変換を行うことで擬似的に対訳単語の取得を行った例である. 実行の結果,「徳川家光」の対訳単語として,「Tokugawa Iemitsu」が得られた.

```
s172114@bill:~/syuuron/v1_3$ echo "徳川 家光"|\
> mecab -Oyomi|\
> uconv -x latin|\
> sed 's/\b\(.\)/\u\1/g'
Tokugawa Iemitsu
```

図 2.4: mecab と uconv を用いた対訳単語の取得の例

第3章　提案手法

3.1　提案手法の概要と手順

本研究では, 低頻度語を含む文での翻訳精度の改善を目的とする. そこで, 提案手法では, 低頻度語の中でも人名に着目し, それを類似の頻出語と置き換えて翻訳する.

提案手法では, NMT の翻訳の過程に対して変更を加える. 通常翻訳のみを行う部分を, 変換, 翻訳, 再変換の 3 つの過程を踏んで翻訳する.

具体的な提案手法の手順について, 次のページで図 3.1 を用いて説明する.

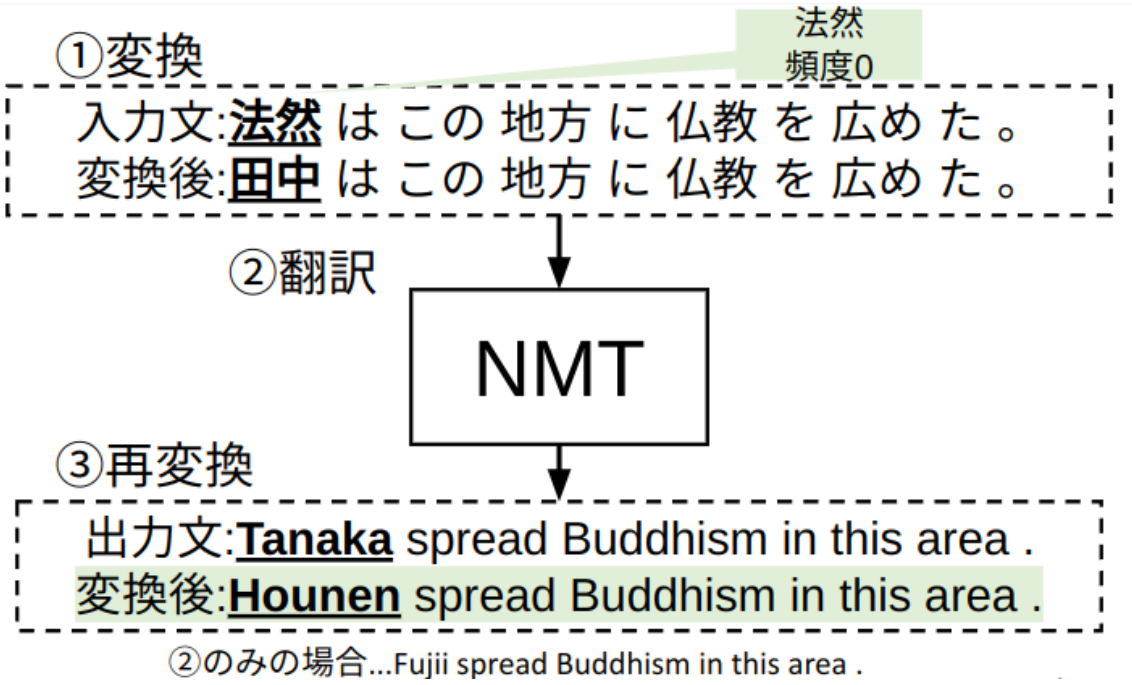

図 3.1: 提案手法の手順

変換 変換では, 入力文に対して mecab を用いて低頻度語の検索を行う. 例では, 法然が学習文中に一度も出ていない低頻度語である. 検索でヒットした語を, 類似の語と変換する. 図 3.1 の例では, 法然を, 同じ人名でかつ頻出語である田中と変換する.

翻訳 翻訳では, 変換した文を翻訳する.

再変換 再変換では, 出力文に出た田中の対訳である Tanaka を, 法然の対訳である Hounen と変換する.

以上の過程を経ることで, 名詞の低頻度語の翻訳が可能であると考える. 提案手法では翻訳の過程においてのみ変更を加えるため, 他の学習の過程を変更する手法と比べ, 任意の NMT モデルを用いることができるという利点が存在する.

3.1.1 対訳の取得方法

対訳の取得には, word2word と uconv を用いる. word2word はカタカナの対訳の取得には有効であるが, 漢字の人名の対訳の取得はデータベースに存在しない, または存在したとして誤った対訳であることが多い. そこでカタカナのみで構成された人名は word2word, それ以外は uconv を用いてローマ字化したものを対訳として使用する. また, word2word で対訳が得られなかった場合も uconv を用いる.

先述の対訳の取得方法について, 図示したものを図 3.2 に載せる. また対訳の例を表 3.1 に載せる.

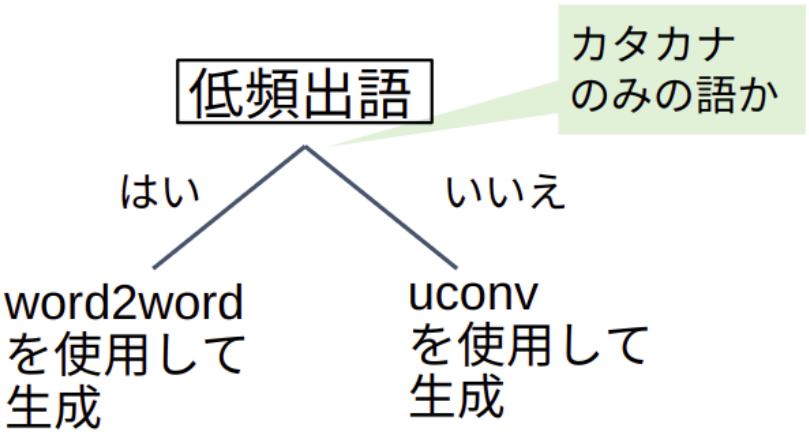

図 3.2: 対訳の取得方法

表 3.1: 対訳の例

アルキメデス	Archimedes
福田	Fukuda
カストロ	Castro
秋篠宮	Akishinonomiya
マラドーナ	Marado-na

第4章　実験

4.1　実験目的と方法

本研究の目的は, 低頻度語を含む文での翻訳精度の改善である. 提案手法として, 翻訳時に低頻度語を頻出語に変換する処理を加える. 評価は図 3.1 において, ②のみを行ったものをベースライン, ①から③まで通して行ったものを提案手法する. ただし, NMT における乱数の影響を抑えるため, 提案手法の ②の翻訳の過程では (矢野,2021)[9] で提案された手法を適用する. また, ベースラインの翻訳の過程では, 8 個のモデルの翻訳精度の平均を用いる. 評価は, ベースラインと提案手法の比較することで行う. 評価方法には自動評価と人手評価を用いる.

4.2 実験データ

4.2.1 学習文

表 4.3 に NMT の学習に用いた学習文対の総数を示す.

表 4.1: 学習文対の総数

学習文対	159, 998 対

本実験で使用する学習文対は日本語文と英語文の対である. 使用する学習文対は電子辞書などの例文より抽出した単文データ [8] である. 学習文対の例を表 4.2 に示す.

表 4.2: 学習文対の例

学習文対例 (1)	
日本語文	ピアノ の 勉強 に ヨーロッパ に 行く .
英語文	Go to Europe to study the piano .
学習文対例 (2)	
日本語文	公園 は 川 まで 広がっ て いる .
英語文	The park reaches to the river .
学習文対例 (3)	
日本語文	きょう は 時折 小雪 の ちらつく 寒い 一 日 だっ た .
英語文	It was a cold day today with occasional light snowfall .

4.2.2 テスト文

表 4.3 に NMT の学習に用いた学習文対の総数を示す.

表 4.3: テスト文の総数

テスト文　197 文

本実験で使用するテスト文は, 提案手法の有効性の確認のため, すべて学習文において低頻度の人名を含むものである. テスト文の例を表 4.17 に示す.

表 4.4: テスト文の例

テスト文の例 (1)	
日本語文	良子 は 東京大学 に 決め た .
参照文	Yoshiko decided on Tokyo University .
テスト文の例 (2)	
日本語文	最近 , しのぶ は ずいぶん お しとやか に なっ た .
参照文	Shinobu has become quite ladylike recently .
テスト文の例 (3)	
日本語文	ロナルド ・ レーガン の まね を 実に うまく 演じ た .
参照文	He impersonated Ronald Reagan with great skill .

なお, 表 4.17 では, 良子としのぶとロナルドが低頻度語である. 良子としのぶは 2 回, ロナルドは 1 回学習文中に登場している.

また, 本研究における低頻度語の定義は,「学習文中での頻度が 5 回未満の単語」とした.

4.3 OpenNMT のハイパーパラメータ

OpenNMT のハイパーパラメータを表 4.5 に載せる.

表 4.5: OpenNMT のパラメータ

n_sample	-1
train_steps	100,000
seed	1-8
n_best	1

語彙生成 (onmt_build_vocab)

NMT では, 学習と翻訳の過程の前に語彙生成の過程がある. ここでは, NMT の学習と翻訳の際に用いる語彙リストの作成を行う. OpenNMT では, onmt_build_vocab のコマンドで実行できる. 語彙数を指定する変数として,n_sample があるが, これを本研究では-1 とした. -1 とすることで, 学習文に出現する全ての語彙を使用することができる.

学習 (onmt_train)

OpenNMT では, 学習は onmt_train というコマンドによって行う. 学習ステップ数を指定する変数として,train_steps があるが, これを本研究では 100000 とした. また, 本研究では学習済みモデルを 8 個作成するため, 乱数を変更しながら学習を 8 回行った. 乱数を指定する変数に seed があるが, これを学習回数に応じて変更した. このことについては, 4.3.1 節にて説明する.

翻訳 (onmt_translate)

OpenNMT では, 翻訳は onmt_translate というコマンドによって行う. 翻訳の出力数を指定する変数として,n_best があるが, これを本研究では 1 とした.

なお, その他の値はデフォルトのものを用いた. また, 以後の考察においても特に断りのない限りはここで示した値を用いる.

4.3.1 (矢野,2021)[9] の研究

概要

NMT は学習の過程で乱数を用いている. その影響で同じ学習文を用いたとしても, 翻訳文が同じにならないことが多い. そこで,pred 値を用いた翻訳選択を行う. 複数のモデルを用いて翻訳選択を行うことで, 乱数の影響を抑えることが期待できる.

pred 値

pred 値は, NMT において翻訳時に出力される値で, 翻訳の尤もらしさを表す. 0 に近いほど NMT モデルにとって自信のある翻訳である. pred 値は他の自動評価と異なり, 参照文なしで値を算出できる. そのためクローズドデータの心配をすることなく利用が可能である.

翻訳選択の手順

a. 乱数のみを変更して学習済みモデルを 8 個生成する
b. 学習済みモデル 8 個で翻訳文を pred 値も含めて生成する
c. 各翻訳の pred 値を参照し、最も pred 値のよいものを出力文とする

実際の出力例を, 図 4.1 に載せる.NMT1 から NMT8 を用いて「彼はまだ歩けない赤ん坊だ。」の翻訳を行う。の 3 列目の pred 値が、それぞれの翻訳に対応する pred の値である. c の手順において, この例では, NMT3 の出力「He is still a little baby .」が選択される.

提案手法において, この pred 値を用いた翻訳選択による翻訳を, 図 3.1 の②と置き換える形で適用する.

入力文	彼 は まだ 歩け ない 赤ん坊 だ 。	
参照文	He is still a baby in arms.	PRED
NMT1	He is still a short baby.	-1.4014
NMT2	He is still a baby.	-2.2151
NMT3	He is still a little baby.	-0.4924
NMT4	He is still a baby boy.	-0.9848
NMT5	He is still a walking boy.	-1.0558
NMT6	He is an baby’s baby.	-1.9851
NMT7	He is still a little baby.	-0.9442
NMT8	He is still hard.	-3.3828

図 4.1: pred 値を用いた翻訳選択の例

4.4 評価方法

自動評価と人手評価を行う. 評価は, テスト文 197 文をベースラインと提案手法を用いて翻訳したものを用いて行う.

自動評価では, 自動評価指標である BLEU[10], NIST[11], METEOR[12], RIBES[13], TER[14], WER[15] にて評価を算出する.

人手評価では, 対比較評価を行う. ベースラインと提案手法の翻訳を比較し, どちらがよりいい翻訳であるかという観点から評価を行う.

4.5 実験結果

4.5.1 各評価結果

自動評価結果

自動評価結果を表 4.6 に示す.

表 4.6: 自動評価結果 (197 文)

	BLEU	NIST	METEOR	RIBES	TER	WER
ベースライン	0.112	3.298	0.315	0.672	0.782	0.800
提案手法	0.124	3.328	0.353	0.721	0.700	0.715

表 4.6 について説明する. BLEU, NIST, METEOR, RIBES は, 値が大きいほど良い翻訳であることを示す. また, TER, WER は値が小さいほど良い翻訳であることを示す. ベースラインと提案手法の値を比較すると, 全ての指標において提案手法がベースラインの値を上回る結果となった.

人手評価結果

人手評価結果を表 4.7 に示す.

表 4.7: 対比較評価 (100 文)

提案手法○	ベースライン○	互角
42	8	50

表 4.7 について説明する. 人手評価では, 対比較評価を行う. ベースラインと提案手法の翻訳を比較し, どちらがより良い翻訳であるかという観点から評価を行う. 提案手法がベースラインよりも良いと判断したものを提案手法○, ベースラインが提案手法よりも良いと判断したものをベースライン○, どちらがよいかの判断がつかなかったものを互角とする. 人手評価の結果, ベースラインよりも提案手法のほうが良いことが分かった.

以上の結果より, 提案手法を用いることで低頻度の人名の翻訳精度の改善が可能であることが確認できた.

4.5.2 改善例

人手評価にて, 提案手法の翻訳がベースラインよりもよいと判断した例について載せる.

表 4.8: 提案手法◯の例 (1)

日本語文	ジョエル から の その 手紙 は 5月 5 日 付け で ある .
参照文	The letter from Joel is dated May 5 .
ベースライン	Out of Joel , the letter is dated May 5 .
提案手法	The letter from Joel is dated May 5 .

表 4.9: 提案手法◯の例 (2)

日本語文	アーサー ・ ウェイリー は 東洋 について の 直接的 な 知識 に 欠け て い た .
参照文	Arthur Waley lacked a first-hand acquaintance with the Orient .
ベースライン	Clarke lacked the knowledge of the Orient on the Orient .
提案手法	Arthur-san had no direct knowledge of the Orient .

表 4.10: 提案手法◯の例 (3)

日本語文	ぼく は 先日 ジェリー 君 に 会っ た .
参照文	I met Jerry the other day .
ベースライン	I met you the other day .
提案手法	I met Jerry the other day .

4.5.3 互角の例

人手評価にて, 提案手法の翻訳がベースラインの翻訳と比較して優劣がつかないと判断した例について載せる.

表 4.11: 互角の例 (1)

日本語文	虫 が 庭 で チー チー と 鳴い て いる .
参照文	Insects are chirping in the garden .
ベースライン	The insect is quacking in the garden .
提案手法	Insects are clucking in the garden .

表 4.12: 互角の例 (2)

日本語文	私 は 川田 と 申し ます .
参照文	My name is Kawada .
ベースライン	I tell goodbye .
提案手法	My name is Yamada .

表 4.13: 互角の例 (3)

日本語文	王 によって 議会 の 開会 が 宣言 さ れ た .
参照文	Parliament was opened by the King .
ベースライン	The Congressional assembly opened the assembly opening .
提案手法	The meeting opened by Ou .

4.5.4 改悪例

人手評価にて, 提案手法の翻訳がベースラインよりも悪いと判断した例について載せる.

表 4.14: ベースライン◯の例 (1)

日本語文	ロナルド ・ レーガン の まね を 実に うまく 演じ た .
参照文	He impersonated Ronald Reagan with great skill .
ベースライン	He was most successful in Ronald Reagan .
提案手法	She did the most of Ronald .

表 4.15: ベースライン◯の例 (2)

日本語文	この コンテスト に 入賞 し た 人 は , 土田 さん , それ から 小島 さん です .
参照文	Persons receiving awards in this contest were Tsuchida-san and also Kojima-san .
ベースライン	Persons receiving awards in this contest were Tsuchida-san and Kojima-san .
提案手法	People who won this contest are Tsuchita-san , Kojima-san , Tanaka-san .

表 4.16: ベースライン◯の例 (3)

日本語文	この 山 に は 高山 植物 が 群生 し て いる .
参照文	Many alpine plants grow in crowds in this mountain .
ベースライン	Alpine plants grow thickly in this mountain .
提案手法	The plants grow thickly in this mountain .

4.5.5 方式限界の例

正しく置換が成功しているにもかかわらず翻訳が改善しなかった例がいくつか存在した. 例を以下に載せる. 分析を行ったところ, 提案手法で精度が改善しなかったものとしては, 長文翻訳の問題, 置換しても出力されない問題の 2 つが考えられる.

長文翻訳の問題

表 4.17: 方式限界の例 (1)

日本語文	越智 氏 の 更迭 劇 が 今後 の 国会 運営 , 政局 に 与える 影響 も 小さく ない .
参照文	The drama over his sacking is expected to influence the nation's political situation and the ruling camp's management of Diet affairs .
ベースライン	Eda said he has a negative impact on the agenda for his recuperation and the political situation .
提案手法	Ochi has also had a change in the current Diet session , which has no effect in the current Diet session .

長文が翻訳できないという問題である. ベースラインでは越智 (読み:おち) の訳が Eda になっているのに対し, 提案手法では Ochi と正しく翻訳できているように見られる. しかし, 文全体の意味を考えると, ベースラインも提案手法も正しい翻訳ができているとは思えない.

置換しても出力されない問題

表 4.18: 方式限界の例 (2)

日本語文	最上 一族 は 伊達 の 軍勢 に 滅ぼさ れ た .
参照文	The entire Mogami family was destroyed by the Date soldiers .
ベースライン	The top class was engulfed by for for show .
提案手法	The Imperial family was caught in the military forces .

表 4.18 は, 日本語文において最上が低頻度語である. しかし, ベースラインにおいても提案手法においても対応する単語が見られない. これは, 低頻度語を置換しても, NMT モデルにとって文章の解釈が困難であったためと考えられる.

第5章　考察

5.1　提案手法の翻訳精度の理論値の調査

提案手法では, 行程は全て自動で行うようにした. しかし, 固有表現抽出と対訳単語の取得においては, 精度が 100%ではない. そこで, 仮にこれらを完璧に行えたとして, どこまで精度向上が見込めるのかについて検証を行った.

5.1.1　固有表現抽出の精度

テスト文 197 文に対して正しく固有表現抽出が出来ているかどうか調査を行った. 調査の結果, 誤った固有表現抽出があったのは 42 文で, 精度としては 79%だった.

固有表現抽出の誤りの例を表 5.1 に載せる.

表 5.1: 誤った固有表現抽出の例

誤った固有表現抽出の例 (1)	
日本語文	彼 は 今金 に 困っ て いる .
参照文	He's having financial difficulties .
誤った固有表現抽出の例 (2)	
日本語文	モー ガン から 東 に 行き 大 ヴィクトリア 砂漠 に 入っ た .
参照文	We traveled east from Morgan into the Great Victoria Desert .
誤った固有表現抽出の例 (3)	
日本語文	北海道 ・ 日高 地方 に 点在 する 競走 馬 の 育成 牧場 は 約 千 五 百 .
参照文	There are about 1, 500 ranches raising racehorses in the Hidaka district of Hokkaido .

表 5.1 において, 人名と判断したのは日本語において下線が引かれた語である. それぞれ人名ではなく, 期待した抽出ができていない.

5.1.2 対訳単語の取得の精度

テスト文から固有表現抽出の誤った文 42 文を抜き出したあと, 対訳単語の取得の精度について調査を行った. つまり対象のテスト文は 155 文である. 対訳単語の精度は 91%(155 文中 141 文正解) だった.

誤りの例について, 表 5.2 に載せる.

表 5.2: 誤った対訳の例

マラドーナ	Marado-na
福沢	Fukusawa
メージャー	major-league
春日	Syunjitsu
ラビン	Fireman

間違いのパターンは 2 つに大別される.

mecab の誤り

正しい漢字の読みを取得出来ていないことがある. 表 5.2 では, 「福沢」「春日」がこれに該当する.

word2word の誤り

正しい対訳を word2word で得られていないことがある. これは word2word のデータベースが誤っていることが原因である. 表 5.2 では, 「マラドーナ」「メージャー」「ラビン」がこれに該当する.

5.1.3 実験環境

基本の実験環境は提案手法と同じである. ただしテスト文と対訳単語に対して, 以下の2つの処理を手作業で行った.

- テスト文から固有名詞の抽出を誤っている文の削除
- 対訳単語を参照し, 誤っているものを正しく修正

比較は, 提案手法の翻訳精度, 提案手法からテスト文の修正を行った際の翻訳精度, 提案手法からテスト文と対訳単語の修正を行った際の翻訳精度, 以上3つの比較を行う. 比較を行うことで, 固有表現抽出と対訳単語の取得の精度がそれぞれ翻訳精度にどれだけ影響を与えているかの推測が可能であると考える.

5.1.4 実験結果

検証の結果を表5.3に載せる.

表5.3: 自動評価結果

	BLEU	NIST	METEOR	RIBES	TER	WER
提案手法	0.124	3.328	0.353	0.721	0.700	0.715
+テスト文修正	0.130	3.355	0.364	0.740	0.688	0.703
+対訳単語修正	0.131	3.375	0.366	0.741	0.687	0.702

表5.3において, 全ての自動評価指標において, 最後のテスト文修正と対訳単語修正を行った場合の値が最もよいものとなった. よって, 手動で誤りを修正することで, 翻訳精度が更に向上することがわかった.

しかし, 対訳単語の修正を行ったとしても, 上昇値はわずかである. そのため, 提案手法の精度向上に向けては, 固有表現抽出の精度を高めることを優先すべきであると考える.

5.2　置換に用いる人名の変更

提案手法では, 低頻度語を類似の頻出語と置き換えて翻訳を行う. ここで用いる頻出語には, 実験では変更したところで大きな差はないと考え, 全て「田中」を用いていた. そこで本項では, 置換する単語を変更してそれぞれの翻訳精度を求め, よりよい置換に用いる頻出語がないかを探求する.

5.2.1　実験環境

基本の実験環境は提案手法と同じである. 置換に用いる頻出語を適宜変更し, それぞれを用いた際の翻訳精度を比較する. 置換に用いる頻出語には,「人名ではないが頻度が大きい語」,「人名であるが語義が文脈によって複数ある語」などを含め, 精度向上のための知見を得られないか検討する.

5.2.2 実験結果

検証の結果を表 5.4 に載せる.

表 5.4: 自動評価結果

	頻度	BLEU	NIST	METEOR	RIBES	TER	WER
ベースライン		0.112	3.154	0.298	0.649	0.788	0.810
提案手法 (田中)	23	0.124	3.328	0.353	0.721	0.700	0.715
花子	16	0.121	3.353	0.357	0.723	0.703	0.722
山田	17	0.123	3.227	0.346	0.722	0.707	0.718
ジョン	70	0.119	3.200	0.352	0.711	0.697	0.711
日本	1566	0.117	3.373	0.341	0.719	0.724	0.740
森	34	0.110	3.081	0.319	0.663	0.749	0.766

表 5.3 について考察する. 各評価指標において最良の値は下線を引いている. それぞれ僅かな差ではあるが, 最良の値はばらける結果となった.

「日本」は, 頻度こそ多いものの, 一般には人名ではない語である. そのためか, 他の多くの人名の翻訳精度には劣る結果となった. しかし, ベースラインの性能を上回ることには成功した. このことより, 人名の低頻度語を文から削除するだけでも, 翻訳精度の向上に効果があるのではないかと考える.

置換を用いた手法としては,「森」を置換に用いた際の翻訳精度が最も悪かった. これは, 森が人名 (=mori) であったり, また森林の意味での森 (=forest) として使われるなど, 文脈によって訳語が異なることが原因であると推測する.

実験の結果より, 精度向上のために用いる頻出語の条件について, 以下のことが言える.

- 人名の語であること
- 文脈による多義性のない語であること

第6章　おわりに

機械翻訳の手法の一つに, ニューラル機械翻訳がある. 流暢性の高い翻訳が可能であるとされ, 現在機械翻訳において主流の方式である. しかし一方で, NMT には翻訳精度の向上を目的とする上で問題となる点が複数存在する. そのうちの一つに, 低頻度語を正しく翻訳できないという問題点がある.

本研究では, この低頻度語に関する問題に着目した. 低頻度語の中でも, 人名は対訳が取ることが比較的容易である点を活かし, 低頻度の人名を含む文に対して, 置換を用いて翻訳する手法を提案した.

実験では, 従来手法と提案手法とで翻訳精度の比較を行った. 実験の結果, 提案手法では従来手法よりも, 低頻度語に対して頑強となり, 翻訳精度も向上することが分かった.

今後の課題を以下にまとめる.

- 人名のみならず, 地名などの他の名詞においても同様の手法が取れないかについて検討を行う
- 固有表現抽出の精度向上を図る

謝辞

本研究を進めるにあたり, 研究の説明や論文の書き方など様々なご指導を頂きました鳥取大学工学部電気情報系学科自然言語処理研究室の村上仁一准教授に心から御礼申し上げます. また, 本研究を進めるにあたり, 御指導, 御助言を頂きました, 村田真樹教授に心から御礼申し上げます. また, 同じ班に所属されていた自然言語処理研究室の皆様へ心から感謝の気持ちと御礼を申し上げたく謝辞にかえさせて頂きます.

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

[2] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-source toolkit for neural machine translation. ArXiv e-prints, 2017.

[3] 今仁優希. ニューラル機械翻訳における低頻度語処理. 言語処理学会第 24 回年次大会. 2018.

[4] 岩山 真. 深層学習に基づく特許翻訳における数値表現の扱い. Japio year book 308-313, 2017.

[5] https://taku910.github.io/mecab/

[6] https://linux.die.net/man/1/uconv

[7] Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. word2word: A Collection of Bilingual Lexicons for 3, 564 Language Pairs. arXiv:1911.12019, 2019.

[8] 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, pages119-130, 2012.

[9] 矢野貴大, 村上仁一. ニューラル機械翻訳に乱数が与える影響. 言語処理学会第 27 回年次大会. 2021.

[10] K. Papineni, S.Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In ACL, 2002.

[11] G. Doddington. Automatic evaluation of machine translation quality using N-gram CoOccurrence statistics. Proc. of Second International Conference on Human Language Technology (HLT), San Diego, pages 138-145, 2002.

[12] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005.

[13] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. Conference on Empirical Methods on Natural Language Processing (EMNLP), 2010.

[14] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. Proceedings of Association for Machine Translation in the Americas, 2006.

[15] Klakow, Dietrich, and Jochen Peters. Testing the correlation of word error rate and perplexity. Speech Communication,pages 19–28, 2002.