2022年度（令和4年度）　修士論文

単文学習を用いた複文の翻訳実験

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
M21J4031Z 竹本 祐基

概要

従来, ニューラル機械翻訳 (NMT:Neural Machine Translation), フレーズベース機械翻訳 (PBSMT:Phrase Based Statistical Machine Translation), 変換手動型翻訳 (TDSMT;Transfer Driven Machine Translation) などの研究において, 単文を翻訳することを試みてきた. しかし, 私達が普段用いる文は複文が多い. よって, 本研究の目的としては, 学習データを単文のみとした時の複文翻訳を試みる.

本研究の問題点としては, 学習データが単文のみの場合翻訳において精度に不安が生じることである. そこで本研究では単文を 2 文組み合わせることにより, “単文+単文” データを作成し, 複文とみなすことで複文を翻訳する方法を提案する. しかし, 単文 2 文をそのまま組み合わると, 学習データは膨大になり学習しきれなくなってしまうという問題がある. そこで, 学習データと入力文との類似文検索を何度か行い, 学習データを作成する. そして NMT に作成した学習データを学習させることにより, 翻訳する. 実験として, 複文のテスト文 100 文を入力とした翻訳の結果を人手評価した.

その結果, 提案手法と従来手法を比較すると提案手法の翻訳結果の方が優れていた. しかし, 人手評価の結果としては, 従来手法, 提案手法ともに差なしと判断した文が多く出力されたことより, 改善が必要であると考える.

目 次

図 目 次

表 目 次

第1章　はじめに

従来, ニューラル機械翻訳 (NMT:Neural Machine Translation), フレーズベース機械翻訳 (PBSMT:Phrase Based Statistical Machine Translation), 変換手動型翻訳 (TDSMT;Transfer Driven Machine Translation) などの研究において, 単文を翻訳することを試みてきた. しかし, 普段用いる文は複文が多い. そこで本研究の目的としては, 学習データを単文のみとした場合の複文翻訳を試みる.

本研究の問題点としては, 学習データが単文のみの場合翻訳において精度に不安が生じることである. そこで本研究では単文を 2 文組み合わせることにより, “単文+単文” データを作成し複文とみなすことで, 複文を翻訳する方法を提案する. しかし, 単文 2 文をそのまま組み合わると, 学習データは膨大になり学習しきれなくなってしまうという問題がある.

本研究の主な主張点を以下に整理する.

- 学習データを単文のみとした場合の複文翻訳の精度がどのくらいなのかの調査の研究はなく新規である.
- 単文の学習データを用いて, “単文+単文” の学習データを作成し, 複文翻訳において翻訳可能であるかの研究を試みた.

  単文のみの学習データに対して, “単文+単文” の学習データを追加した学習データの翻訳結果の精度が向上した.

本論文の構成は以下の通りである．第 2 章では，ニューラルネットワークについて述べる．第 3 章では, 本研究の類似研究について述べる. 第 4 章では, 文の構造について述べる. 第 5 章では，本研究の手法について述べる．第 6 章では，実験環境について述べる．第 7 章では，実験結果を述べる．第 8 章では，本実験の考察を述べる．第 9 章では，本実験の今後の課題について述べる．第 10 章では, 本実験の簡単なまとめを述べる.

第2章　ニューラルネットワーク (NMT)

2.1　ニューラル機械翻訳 (NMT) の概要

ここでは, 本研究で用いる NMT の概要について下記に示している.

NMT とは, ニューラルネットワークを活用した手法であり, 学習も翻訳も全て同じ枠組みで行うことができる. 対訳文を与えることにより, ニューラルネットワークが翻訳に必要な情報を自動的に学習する. このように入力から出力までが単一のモデルで完結するような枠組みをエンドツーエンド (end-to-end) と呼ぶ.

NMT は大きく分けて 3 つのパーツで構成されている.1 つ目は入力文を実数値の集合であるベクトル表現に符号化 (変換) するエンコーダー (encoder)，2 つ目は出力するべき単語を決定する際に，符号化された入力文のどこに注目すべきかをコントロールするアテンション機構 (attention mechanism)，3 つ目は符号化された入力文とアテンション情報を基に出力文を復号化 (生成) するデコーダー (decoder) である.

図 2.1 にニューラル機械翻訳の流れの図を記載する.

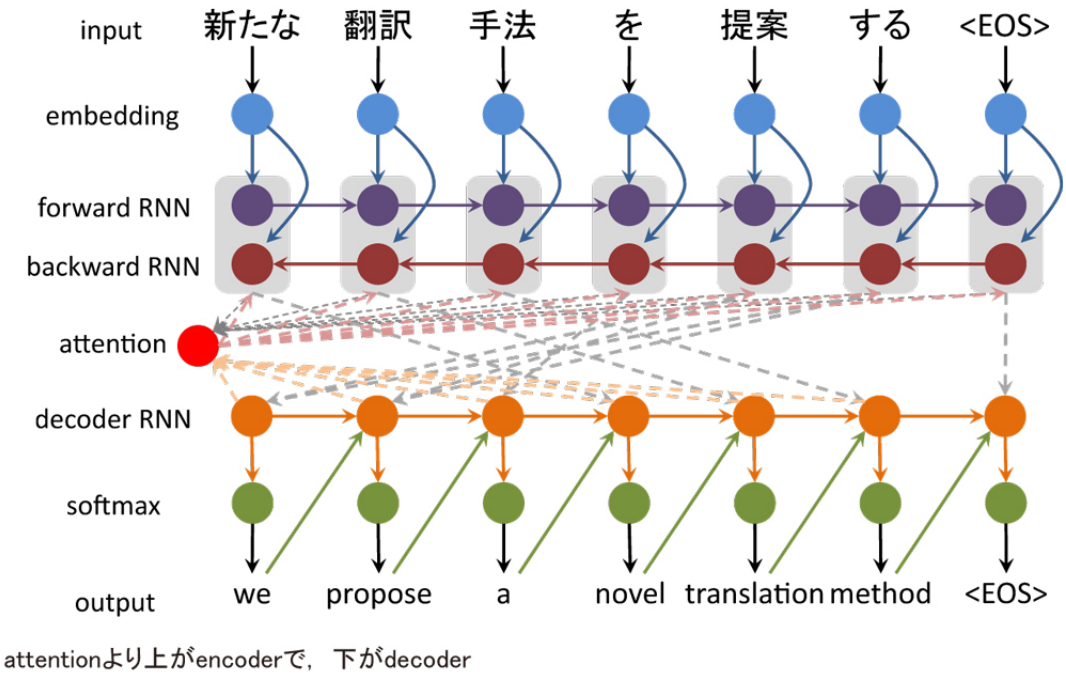

図 2.1: ニューラル機械翻訳の (NMT) の翻訳方式の概要

2.2 NMT の翻訳の流れ

NMT の翻訳時の流れについて以下に記す.

1. 原言語の文章の形態素解析を行い, 単語列に分割する.
2. 分割した単語を数値表現に変換する.
3. 単語の符号化を行い, ベクトルに変換する.
4. エンコーダー（encoder）は原言語の文の符号化をし, 原言語の文章のベクトルを作成する. この層はいままで作成したベクトル群を再帰ニューラルネットワークによって処理する. ここでは, 符号化された入力文のどこに注目すべくアテンション機構（attention mechanism）を用意する.
5. 4で作成された原言語のベクトルとアテンション情報をもとに, デコーダー（decoder）を用いて目的言語のベクトルに変換し, 目的言語の単語を順次に生成し, 新しいベクトルを生成する.
6. 5の同じ処理を繰り返して実行し, 文章の終わりを示す特殊文字 EOS（End Of Sentence）の出力を完成後, 翻訳作業を終了する.

第3章　関連項目

本章では, 本研究に関わる類似研究, 文構造の説明について記載する.

3.1　類似研究

本章は,Seiichiro Kondo らによる類似研究 Sentence Concatenation Approach to Data Augmentation for Neural Machine Translation[3] の抜粋である.

近年 NMT は高い精度で翻訳が行えるため, 広く注目されている. しかしながら, 長文翻訳では低い精度を示し, 低リソース言語での主な問題点となっている. 長文の翻訳品質が低い原因は, 学習データ中の長文が少ないためである.

そこで,Seiichiro Kondo ら [3] は長文処理のための簡単なデータ拡張方法を提案した. これは, 学習データとして指定されたパラレルコーパスを使用し,2 文を連結させる方法である.

この手法により, 短文の翻訳品質は低下するが, 長文の翻訳に有効であることが分かっている.

Kondo ら [3] の研究では, 主に長文の翻訳精度を改善することを試みて研究されている. しかし, 複文や重文の翻訳には触れられていない. また, データ拡張方法でも, テスト文と近くない文同士を連結させていることも考えられる. そこで, 本研究ではこれらの問題を解決する. 本研究では, 主題を複文に置き, 複文テスト文に近い文を連結させる手法を行う. 近い文を連結させるために類似文検索を用い, 類似文を出力させる.

3.2 文構造

本章は, 文構造についてまとめたサイト [4] の抜粋である. 本実験で用いられる単文, 複文, 重文, 重複文の文構造の違いについて一般的な見解を示す.

- 単文とは, 文中に述語が 1 つの文のことを指す.

  例:私は本が好きです。

- 重文とは, 単文が 2 つ以上並列に重なった文のこと. 接続詞で結ばれていることが多いので, 切り離すことができる.

  例:私は本が好きで、弟はすポースが好きです。

- 複文とは,1 つの単文の中に単文が組み込まれている文のことである. 英語では,i think that〜 のように, 文の中に主語+述語が含まれている形である.

  例:私は父が買ってくれた絵を大切にしている。

- 重複文とは, 重文と複文が結合された文のこと.

  例:私は毎晩仕事終わりにのんびりしながら飲むお酒が大好きだ。

また, 複文にはいくつかのタイプがある.

1. 連体節:名詞を修飾しているタイプの複文

   例:彼女が作ったケーキは、おいしかった。

2. 補足説 (名詞句化):「こと」や「の」を伴って名詞になったタイプの複文

   例:テーブルの上のケーキを食べたのは、私です。

3. 補足説 (引用節):文中に「」があるタイプの複文

   例:彼は、僕はケーキを食べていないと言った。

4. 補足説 (疑問表現):「か」や「かどうか」で前の文を受けている疑問表現タイプの複文

   例:何をしていたかを説明しなさい。

5. 述語を修飾して, 原因・理由, 目的, 条件などを表すタイプの複文

   例:おなかが空いていたので、ケーキを食べた。

第4章　提案手法

本章では, 本研究で扱う提案手法の説明を行う.

4.1　提案手法の概要

複文の翻訳において, 学習データが単文のみしか存在しない場合において複文の翻訳はほぼすべての文において翻訳精度が不安定になってしまうという問題がある.

そこでその問題を解決するため単文の学習データを組み合わせることにより,“単文+単文”のデータを作成する. しかし, 単文のデータ全てをかけ合わせてしまうと学習データ量が膨大になってしまい, 学習不可能となってしまう. そのため類似文検索を行い, 掛け合わせる単文のデータを出力させる. 出力させた類似文同士を組み合わせ,“単文+単文”の学習データを作成する. その後作成した “単文+単文”の学習データとオリジナルの単文学習データと足し合わせることによって, 最終的に NMT に学習させる学習データを作成する.

4.2　“単文+単文”データ

以下に, 作成する “単文+単文”のデータの説明を行う. 本来複文は,2 つ以上の単文に分けることができる文のことである. そこで, 本実験では単文同士を組み合わせることによって “単文+単文”を作成する.“単文+単文”とは, 単文 2 文が繋がっている文のことである.

単文+単文データの作成について表 4.2.1 に示す.

表 4.2.1: “単文+単文”データの作成

単文 1	私は学校に行きます。I go to school.
単文 2	私はテニスをします。I play tennis.
“単文+単文”	私は学校に行きます。私はテニスをします。I go to school. I play tennis.

表 4.2.1 において, 単文 1 と単文 2 を文同士を結合させるプログラムを用いて連結させ, “ 単文+単文 ”のデータを作成する. この時, 日本語文, 英語文ともに文連結を行う.

4.3 類似文検索

“単文+単文”のデータを作成するために複文のテスト文と単文の学習データの間で類似文検索を行う. この時類似文検索を行うのは, 学習データの量の削減を行わなければ ,NMT が学習不可能となるためである. 本研究で行っている類似文検索は,TF を用いて行っており, 複文のテスト文と単文の学習データの間に同じ文字が含まれている割合を算出し, その値が高い順に並べ, その上位 100 個を出力としている.

上記のようにして, 類似文検索を行い, 出力された類似文を組み合わせて学習データとみなす. この時, 組み合わせた学習データが膨大であると出力結果の翻訳精度が著しく低下する. そこで, 精度低下を防ぐため, 本実験では類似文検索を 2 度行った.

類似文の出力例を表 4.3.1 に示す. この例では, 日本文「私は学校に行き、テニスをします。」に対して, 類似文 1「私は学校に行きます。」, 類似文 2「私はテニスをします。」が出力されている.

表 4.3.1: 類似文の出力例

日本文	私は学校に行き、テニスをします。
参照文	I go to school and play tennis.
類似文 1	私は学校に行きます。I go to school.
類似文 2	私はテニスをします。I play tennis.

類似文検索の手順について示す.

1. まず初めに, 複文のテスト文があるデータからランダムに 100 文を抽出する.
2. 抽出した複文 100 文とオリジナルの単文学習データとの間で類似文検索を行う. 複文 1 文に付き各 100 文の類似文を出力される.
3. 1 度目の類似文検索によって出力された類似文をかけ合わせることにより,“単文+単文”のデータを作成する. 複文 1 文に付き 10,000 文の “単文+単文”データを作成する.100 文合計で 100 万文の “単文+単文”データとなる.

4. 100 万文の “単文+単文” データを学習させると翻訳精度が低下するので, 複文のテスト文と各 “単文+単文” データの間で 2 度目の類似文検索を行う. 出力は複文 1 文に付き上位 100 文出力させる.

図 4.3 に類似文検索の手順を記載する.

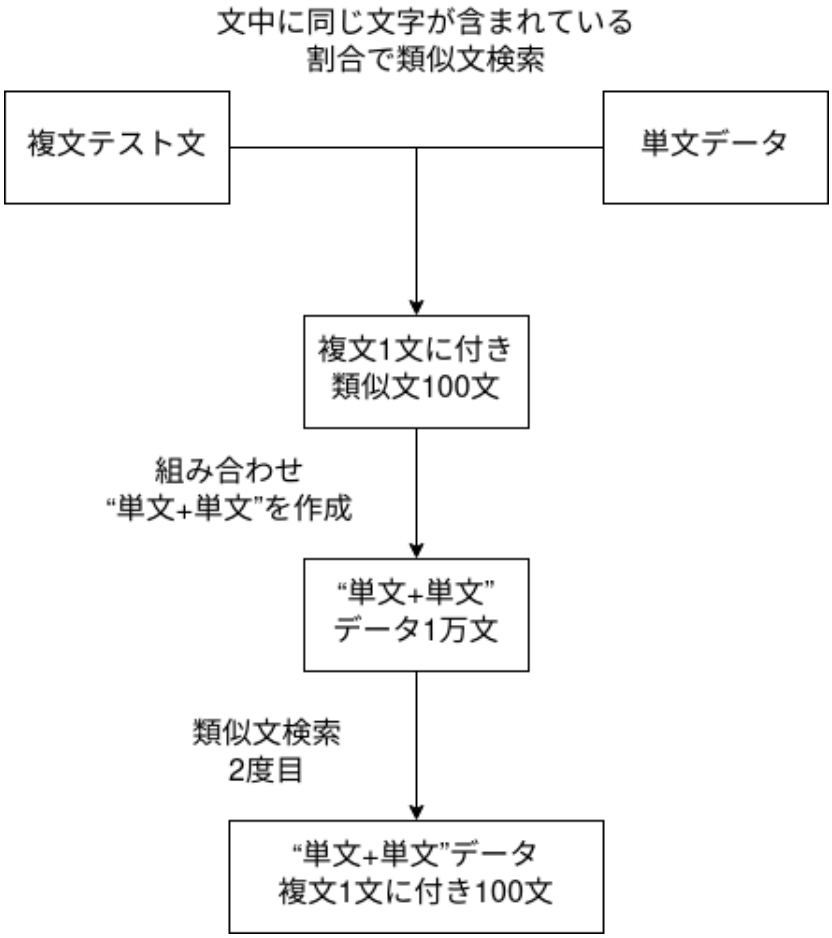

図 4.1: 類似文検索の手順

4.4 学習と翻訳

本研究における学習の過程では, 上記の学習データ作成の過程において類似文検索を行い出力された “単文+単文” のデータをオリジナルの単文学習データと共に NMT に学習させる. この時, テスト文 100 文の “単文+単文” データは合計 10,000 文となる. その後, 複文のテスト文 100 文を翻訳にかける.

第5章　実験環境

5.1　使用NMT

本実験には,Open-NMT[5] を使用した.

5.2　使用学習データ

本実験では,“単文+単文”の学習データを作成, 及び翻訳実験に用いる入力文のため, 電子辞書などの例文より抽出した研究室の単文, 複文コーパスを用いる. データの内訳を表5.2.1 に示す.

表 5.2.1: 使用データ

単文データ	163,188
“単文+単文”データ	10,000
複文テストデータ	9243

5.3 実験方法と目的

5.3.1 実験方法

NMT を用いて, 複文の翻訳を行う.

実験に用いる学習データは 2 つで, ベースラインはオリジナルの単文データ 16 万文, 提案手法はオリジナルの単文データ 16 万文+“単文+単文”データ 10,000 文として学習, 複文のテスト文 100 文の翻訳を行う. その後, それぞれの翻訳結果を人手による対比較評価で行う. また, 自動評価の結果も示す.

表 5.3.1: それぞれの学習データ数

ベースライン	163,188
提案手法	173,188

5.3.2 目的

本実験の目的としては, 単文の学習データのみを使用して, 複文の翻訳精度をどれほど向上させることができるのかを調査することである. そのために, 複文のテスト文と単文の学習データから類似文を出力させ,“単文+単文”の学習データを作成し, 本実験に用いている.

第6章　実験結果

6.1　人手評価結果

複文のテスト文を翻訳させた 100 文を用いて, ベースライン (単文データのみ) と提案手法 (単文データと “単文+単文”データ) にお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. 表 6.1.1 に人手評価で対比較評価を行った結果を示す.

- ベースライン○　　：ベースラインの方が良い
- 提案手法○　　：提案手法の方が良い
- 両方良い　　：翻訳精度に明確な差がなく翻訳が正しい
- 両方よくない　　:翻訳精度に明確な差がなく翻訳が正しくない

表 6.1.1: ベースラインと提案手法の人手評価結果 (100 文中)

提案手法○	ベースライン○	両方良い	両方よくない
18 文	8 文	5 文	69 文

表 6.1.1 の結果より, ベースラインと提案手法の翻訳結果を比較すると, 提案手法の翻訳結果がよく, 精度の向上が確認できる．しかし, ベースラインと提案手法において両方良い文は 5 文, 両方よくない文は 69 であった. これより, 全体の翻訳精度としては高くない.

6.2　自動評価結果

複文のテスト文 100 文を入力文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. 自動評価は,BLEU[6],METEOR[7],RIBES[8],TER[9] を使用した. 表 6.2.1 に, それぞれの手法における自動評価の結果を示す.

表 6.2.1: 自動評価結果 (精度が高い方を太字で示す)

自動評価	BLEU	METEOR	RIBES	TER
提案手法	0.096	**0.353**	**0.685**	**0.728**
ベースライン	**0.104**	0.351	0.671	0.746

自動評価の結果としては,“単文+単文”の学習データを混ぜた提案手法のの結果がよくなっている.

6.3 出力類似文

類似文検索によって, 出力された類似文の例を下記の表へ各評価ごとに示す. 類似文の例は, 日本文, 英文どちらも示す.

表 6.3.1: 提案手法○の時の類似文出力 k 結果

入力文	馬 に むち を 当て て 走ら せ 続け た 。
参照文	He whipped his horse along .
類似日本文 1	彼 は 馬 の しり に ピシッ と むち を 当てた 。
類似英文 1	He cracked the whip on the horse’s rear .
類似日本文 2	馬 に 拍車 を 当てた 。
類似英文 2	He dug his spurs .
類似日本文 3	君 は 現在 の 仕事 を 続け た 方 が いい 。
類似英文 3	You had better stick to your present job .
類似日本文 4	冬 服 は 風 を 当て て しまった 。
類似英文 4	I aired my winter clothes before putting them away .

提案手法○の類似文出力例 6.3.1 では,「馬」,「むち」,「当てて」,「続け」, の入力文中の主要な単語が類似文として出力されている.

表 6.3.2: ベースライン○のときの類似文出力結果

入力文	春 が 来 て 牧場 は 一 面 緑 で 覆わ れ た 。
参照文	When spring came , the pasture became covered with green .
類似日本文 1 類似英文 1	早く 春 が 来 て 欲しい 。 I wish spring would come soon .
類似日本文 2 類似英文 2	野原 は 一 面 に 白い 雪 で 覆わ れた 。 The field was covered with white snow .
類似日本文 3 類似英文 3	木々 は まもなく 緑 で 覆わ れる だろ う 。 The trees will soon be green .
類似日本文 4 類似英文 4	電流 が 来 て いる 。 The current is on .

ベースライン○の類似文出力例 6.3.2 では,「春」,「一面」,「緑」,「覆われた」, 等の主要な単語は出力されたが,「牧場」という単語は出力されていないことが分かる.

表 6.3.3: 両方良いのときの類似文出力結果

入力文	彼 は バナナ の 皮 を 踏ん で すべっ た 。
参照文	He slipped on a banana skin .
類似日本文 1 類似英文 1	私 は バナナ の 皮 を 剥いた 。 I peeled the banana .
類似日本文 2 類似英文 2	彼 は バナナ の 皮 で 足 を 滑ら せた 。 He slipped on a banana peel .
類似日本文 3 類似英文 3	彼 は アクセル を 踏ん だ 。 He stepped on the accelerator .
類似日本文 4 類似英文 4	彼 は ひざまずいた 。 He fell on his knees .

両方良いときの類似文出力例 6.3.3 では,「バナナ」,「皮」,「踏ん」,「滑ら」等の入力文を構成する単語が全て出力されている.

両方よくないのときの類似文出力例 6.3.4 では,「特赦」,「出獄」等の単語が出力されていない.

表 6.3.4: 両方良くないのときの類似文出力結果

入力文	彼 は 特赦 に あっ て 出獄 し た 。
参照文	He was let out of prison under an amnesty .
類似日本文 1	子細 あっ て 彼 は 姿 を 見せ なかった 。
類似英文 1	He was absent for a certain reason .
類似日本文 2	彼 は むっつり し た 人 だ 。
類似英文 2	He is a morose man .
類似日本文 3	彼 は 自力 で 出世 し た 人 だ 。
類似英文 3	He is a self-made man .
類似日本文 4	彼 は 南極 に 到達 し た 最初 の 人 でした 。
類似英文 4	He was the first man to reach the South Pole .

6.4 翻訳結果

実験の結果, 出力された複文翻訳の結果例を示す. また, それぞれの結果の時に学習させた “単文+単文”データも示す.

表 6.4.1: 提案手法○の出力結果

入力文 1	馬 に むち を 当て て 走ら せ 続け た 。
参照文 1	He whipped his horse along .
提案手法 1	He ran the whip on his horse .
ベースライン 1	He drove the horse to the bridle .
“単文+単文”日本語データ 1	冬 服 は 風 を 当て て しまった 。 馬 に 飛び乗った 。
“単文+単文”英語データ 1	I aired my winter clothes before putting them away . He leapt onto his horse .
“単文+単文”日本語データ 2	馬 に 拍車 を 当てた 。 冬 服 は 風 を 当て て しまった 。
“単文+単文”英語データ 2	He dug his spurs . I aired my winter clothes before putting them away .
入力文 2	二 人 の 哀れ な 恋 は 村 の 噂 に 残っ て いる 。
参照文 2	Their sad loves live in village gossip .
提案手法 2	Their love is still afloat of the village .
ベースライン 2	Their love is inhabited by gossip in the village .
“単文+単文”日本語データ 1	彼女 は 二 人 の 子ども が いる 。 雪 が 所所 に 残って いる 。
“単文+単文”英語データ 1	She has two children . The snow remains here and there .
“単文+単文”日本語データ 2	この 習慣 は まだ 地方 に 残っ て いる 。 彼女 は 二 人 の 子ども が いる 。
“単文+単文”英語データ 2	This custom is still lingering on in the rural districts . She has two children .

表 6.4.1 より, 複文テスト文を入力文として, 提案手法とベースラインをそれぞれ学習させたものをそれぞれ翻訳にかけた. 提案手法とベースラインをそれぞれ比較すると, 入力文 1 の出力結果では, 提案手法 1 の出力がベースライン 1 の出力よりも参照文 1 に近い出力担っていることが分かる. 入力文 2 では, どちらの出力も哀れなという意味の sad は出力されていないが, 提案手法 2 では,afloat((うわさが) 広まる) が出力されているのに対し, ベースライン 2 では,inhabit(習慣) が出力されており, 提案手法の出力が近い意味になっていると分かる. また, 提案手法○の出力例では, 入力文に含まれる単語が“ 単文+単文 ”

データに含まれていることも分かる.

表 6.4.2: ベースライン○の出力結果

入力文	春 が 来 て 牧場 は 一 面 緑 で 覆わ れ た 。 。
参照文	When spring came , the pasture became covered with green .
提案手法	The spring was covered with verdure .
ベースライン	The pasture was green with spring .
“単文+単文”日本語データ 1	早く 春 が 来 て 欲しい 。 野原 は 一 面 に 白い 雪 で 覆わ れた 。
“単文+単文”英語データ 1	I wish spring would come soon . The field was covered with white snow .
“単文+単文”日本語データ 2	早く 春 が 来 て 欲しい 。 空 は 一 面 に かき曇った 。
“単文+単文”英語データ 2	I wish spring would come soon . The sky clouded over .

表 6.4.2 において, ベースライン○となったのは, 提案手法の出力において,pasture(牧場) が出力されず, 代わりに verdure(新緑) が出力されたため.“単文+単文”データにおいても,pasture が含まれる文は存在しなかった.

表 6.4.3: 両方良いの出力結果

入力文	彼 は バナナ の 皮 を 踏ん で すべっ た 。
参照文	He slipped on a banana skin .
提案手法	He slipped on the banana peel .
ベースライン	He slipped on the banana peel .
“単文+単文”日本語データ 1	私 は バナナ の 皮 を 剥いた 。 彼 は アクセル を 踏んだ 。
“単文+単文”英語データ 1	I peeled the banana . He stepped on the accelerator .
“単文+単文”日本語データ 2	彼 は 勇躍 ホーム を 踏ん だ 。 私 は バナナ の 皮 を 剥いた 。
“単文+単文”英語データ 2	He crossed the home plate in high spirits . I peeled the banana .

表 6.4.4: 両方良くないの出力結果

入力文	彼 は 特赦 に あっ て 出獄 し た 。
参照文	He was let out of prison under an amnesty .
提案手法	He was in a league .
ベースライン	He shoved himself in a roundabout way .
“単文+単文”日本語データ 1	その 大 寺院 は 幾多 の 災害 に あっ て きた 。 彼 は こせこせ し た 男 だ 。
“単文+単文”英語データ 1	The cathedral has had many disasters . He is a fussy man .
“単文+単文”日本語データ 2	お前 の おかげ で また こんな 目 に あっ て しまった 。 彼 は のっそり し た 男 だ 。
“単文+単文”英語データ 2	Here’s another fine mess you’ve gotten me into . He is very slow in moving .

6.5 学習データの分割

本実験では, 提案手法の学習データは, テスト文 100 文の各類似文をまとめ,1 度の NMT 学習で翻訳実験を行った. しかし, ここでテスト文 100 文の類似文をまとめて学習させるのではなく, テスト文を 10 文ずつに分割し,NMT に学習させる翻訳実験を行った.10 文づつに分割しているため,NMT 学習は 10 回行われている.

6.5.1 評価結果

この時の, 対比較評価を表 6.5.1, 自動評価の結果を表 6.5.2 に示す.

表 6.5.1: 学習データ分割:ベースラインと提案手法の人手評価結果 (100 文中)

提案手法〇	ベースライン〇	両方良い	両方よくない
22 文	6 文	4 文	68 文

表 6.5.2: 学習データ分割:自動評価結果 (精度が高い方を太字で示す)

自動評価	BLEU	METEOR	RIBES	TER
提案手法	**0.109**	0.344	**0.707**	**0.739**
ベースライン	0.104	**0.351**	0.671	0.746

人手評価の結果 6.5.1 より,6.1.1 と比べると, ベースライン〇が 8 文から 6 文, 提案手法〇が 18 文から 22 文となり, 類似文を 100 文まとめて学習させたときよりも 10 文ずつに 10 回学習を行った時の結果の方が, 翻訳精度が改善されている. また, 自動評価の結果でも, 評価値がよくなっている.

6.5.2 出力結果

10 文を 10 分割した時の出力結果について下記の表に示す.

表 6.5.3: 学習データ分割:提案手法○の出力結果

入力文	こんな 軽はずみ な 計画 に 名前 を 貸す つもり は ない 。
参照文	I am not going to lend my name to a harebrained scheme like this .
提案手法	I don't intend to lend my name to this worthless plan .
ベースライン	No name is negligent of this worthless scheme .
“単文+単文”日本語データ 1	考え直す 余地 は ない 。 黒板 に 名前 を 書き なさい 。
“単文+単文”英語データ 1	There is no room for reconsideration . Write your name on the board .
“単文+単文”日本語データ 2	疑う 余地 は ない 。 空欄 に 名前 を 書き なさい 。
“単文+単文”英語データ 2	There is no room for doubt . Write your name in the blank .

表 6.5.4: 学習データ分割:ベースライン○の出力結果

入力文	是 が 非 でも 行く と 言っ て 脅し て いる 。
参照文	He threatens to go whether or no .
提案手法	I am afraid he is not getting by .
ベースライン	I am threatened to go by all means .
“単文+単文”日本語データ 1	皆 そう 言っ て いる 。 是 が 非 でも 勝ち たい 。
“単文+単文”英語データ 1	Everybody says so . I hope to win by all means .
“単文+単文”日本語データ 2	死にかけ て いる 。 是 が 非 でも 勝ち たい 。
“単文+単文”英語データ 2	He is near death . I hope to win by all means .

6.6 類似文検索 1 度の結果

本研究では, 類似文検索は 2 度行った. ここでは類似文検索を 1 度しか行わなかった時の結果を示す. 類似文検索を 1 度しか行わなかったため,NMT に学習させた類似文は 1 文につき 1 万文であり, 合計は 100 万文となる. この時, 使用した学習データは表 6.6.1 である.

表 6.6.1: 類似文検索 1 度:使用データ

“単文+単文”データ	1,000,000
提案手法	1,148,958
ベースライン	163,188

6.6.1 人手評価と自動評価結果

人手評価結果を表 6.6.2 に, 自動評価結果を 6.6.3 に示す.

表 6.6.2: 類似文検索 1 度:ベースラインと提案手法の人手評価結果 (100 文中)

提案手法〇	ベースライン〇	両方良い	両方よくない
5 文	54 文	1 文	40 文

表 6.6.3: 類似文検索 1 度:自動評価結果 (精度が高い方を太字で示す)

自動評価	BLEU	METEOR	RIBES	TER
提案手法	0.052	0.063	0.361	1.114
ベースライン	**0.104**	**0.351**	**0.671**	**0.746**

6.6.2 出力結果

類似文検索 1 どの時の翻訳結果を表 6.6.4,6.6.5 に示す.

表 6.6.4: 類似文検索 1 度:提案手法〇出力結果

入力文	きみ なんか い ない も 同じ だ 。
参照文	You are as good as absent .
提案手法	You are as good as absent . You are as good as absent .
ベースライン	You are as good as good .

表 6.6.5: 類似文検索 1 度:ベースライン○出力結果

入力文	春 が 来 て 牧場 は 一 面 緑 で 覆わ れ た 。
参照文	When spring came , the pasture became covered with green .
提案手法	The pasture was green with spring .
ベースライン	The spring was aglow with spring . The spring was come to spring with a spring in spring .

6.7 両手法に複文を追加した結果

ベースライン, 提案手法に複文の学習データを追加した実験を行った. 追加した複文学習データと, 追加後のベースライン, 提案手法の学習データ数について表 6.7.1 に示す.

表 6.7.1: 使用データ

複文データ	92,427
提案手法	265,615
ベースライン	255,615

6.7.1 対比較評価と自動評価

人手評価結果を表 6.7.2 に, 自動評価結果を表 6.7.3 に示す.

表 6.7.2: 複文追加:ベースラインと提案手法の人手評価結果 (100 文中)

提案手法○	ベースライン○	両方良い	両方よくない
17 文	14 文	22 文	47 文

表 6.7.3: 複文追加:自動評価結果 (精度が高い方を太字で示す)

自動評価	BLEU	METEOR	RIBES	TER
提案手法	**0.187**	**0.471**	0.744	0.635
ベースライン	0.185	0.464	**0.747**	**0.629**

6.7.2 出力結果

出力結果を下記の表に示す.

表 6.7.4: 複文追加:提案手法○出力結果

入力文 1	この 迷路 は 1 時間 以内 に 抜け られ なけれ ば 失格 です 。
参照文 1	You will be disqualified if you fail to go through the maze within one hour .
提案手法 1	You will be disqualified if you can't get out of this maze within an hour .
ベースライン 1	This maze is a failure to be lost within an hour .
入力文 2	この 川 に 鉄橋 を 掛ける の に 何 年 も かかっ た 。
参照文 2	It took years to construct the iron bridge across the river .
提案手法 2	It took years to catch a bridge across this river .
ベースライン 2	It took many years to bridge this river over a bridge .

表 6.7.5: 複文追加:ベースライン○出力結果

入力文 1	彼ら の コンピューター に アクセス する 方法 を 偶然 発見 し た 。
参照文 1	I accidentally discovered how to gain access to their computer .
提案手法 1	I discovered how access they were access to their computers .
ベースライン 1	We accidentally discovered the method of access to their computers .
入力文 2	その 広告 に 釣ら れ て 多く の 人々 が その 品物 を 買っ た 。
参照文 2	Many people were allured by the advertisement to buy the goods .
提案手法 2	The ad was bought by the advertisement .
ベースライン 2	The advertisement allured many people into buying the goods .

第7章　考察

本章では提案手法の考察を行う.

7.1　テスト文の類似文の分割

6.5 節の学習データの分割において, 対比較評価結果の表 6.5.1 と自動評価結果の表 6.5.2 は, 学習データを分割せずに翻訳実験を行った結果の表 6.1.1, 表 6.5.2 と比べると, どちらも値が改善されている.

これより, テスト文の類似文を学習させる時, 一括でまとめて学習をさせるよりも, 類似文を分割し学習をさせる方が良い結果になることが分かった. これは, 類似文をまとめて学習させると, 翻訳するテスト文以外の類似文が多く学習されているため, 翻訳時にノイズが含まれるためであると考える.

上記より, テスト文の類似文を学習する際,1 文ごとに学習を行うことが最も翻訳精度を高める方法であると言える. しかし,1 文ごとに学習を行うのは,NMT の学習時間が 10 倍,100 倍にも膨れ上がり, 多くの時間を費やすデメリットもある.

7.2　“単文+単文”データの数

本実験では,1 文につき類似文を 100 文出力させ,“単文+単文” とした学習データをもう 1 度類似文検索にかけ,1 文につき 100 文づつ学習させている.6.6 節では, 類似文検索を 1 度しか行わなかった時の結果を示した.6.6 節では, 学習させる “単文+単文” の学習データは,1 文に付き 1 万文であり, 合計で 100 万文学習させている.

6.6 節の結果の表 6.6.2, 表 6.6.3 より, 提案手法の結果はどちらも精度が悪くなってしまった. これは, オリジナルの単文データよりもはるかに多くの “単文+単文” データを学習させたためである. 翻訳結果が “単文+単文” の形に引っ張られ,“単文+単文” の形の出力結果が多く出力された. この時, 学習させた “単文+単文” 学習データが正確に複文テスト文を “単文+単文” の形にされているのであれば, 正確な出力結果になると考える.

本実験では, テスト文に近い類似文が少なく, 多くの “単文+単文” データを学習させすぎると還って翻訳精度が低くなってしまう. そのため, 多くの “単文+単文” データを学習させた場合,“単文+単文” データを減らさなければならない. これより,6.6 節の結果は,100 万文ほどの “単文+単文” データを学習させる場合, 類似文検索を複数回行う必要性があることを示している.

7.3 精度

自動評価の結果から分かるとおり, 提案手法のほうが良いと判断した文は 18 文, ベースラインが寄りと判断した文は 8 文, 両方良いが 5 文, 両方良くないと判断した文は 69 文出力された. これより, 本実験の翻訳精度は高いとは言えず, 精度の改善が必要である.

章 6.7.2 では, 複文の学習データ約 9 万文を追加して, 翻訳実験を行った. その結果, 大幅に精度が向上した. これより, 複文の類似文を数文づつ追加するだけでも, 精度が向上するのではないかと思われる.

第8章　今後の課題

本研究では, 単文の学習データを用いて追加の学習データを作成することにより, 単文のみでの複文翻訳を行った. 本章では本実験において残った問題を今後の課題として以下にまとめる.

- 本研究では, 単文の類似文を追加して学習を行ったが, 私達が日頃生活している環境であると複文も存在する. そこで, 類似文検索によって複文の類似文をテスト文1文ごとに数文出力させる. そして, 追加することによって, 日頃に近い環境を再現でき, 精度向上が見込める. よって, 複文の類似文を追加する実験を検討する.
- 本実験では類似文検索において,TF を用いて同じ文字同士が含まれる割合で行った. しかし,TF-IDF のように更に精度の高い類似文検索を行うほうがより近い類似文を出力できる. よって他の方法を用いて類似文検索を行うことを検討したい.

第9章　おわりに

本研究では, 学習データが単文のみの場合における複文翻訳が翻訳可能かを調査するため単文の学習データを用いて複文翻訳の実験を行った. 実験結果より, 提案手法によって学習データを追加した複文翻訳のほうが単文のみの学習データよりも良い翻訳精度が得られることがわかった. 以上より学習データが単文のみの場合においても提案手法の方法を用いることにより, 翻訳精度の向上が図れることが分かる.

今後は, 学習データに複文学習データ 9 万文の類似文を追加させることにより, 翻訳精度を向上させることを目標としていきたい

謝辞

最後に, 三年間に渡り, 本研究のご指導をいただきました鳥取大学工学部知能情報工学科自然言語処理研究室の村田真樹教授, 村上仁一准教授に深く感謝すると共に, 厚く御礼申し上げます. そして, 日常の議論を通じて多くの知識や示唆を頂いた同研究室の皆様に深謝いたします.

参考文献

[1] 中澤敏明：“ 機械翻訳の新しいパラダイム：ニューラル機械翻訳の原理 ”, 情報管理, Vol. 60, No. 5,pp. 209-306, 2017.

[2] 呉浩東：“ 機械翻訳の原理と研究動向 ”, マテシス・ウニウェルサリス, Vol. 20, No. 5, pp. 27-41, 2019.

[3] Seiichiro Kondo, Kengo Hotate, Tosho Hirasawa, Masahiro Kaneko, Mamoru Komachi : “Sentence Concatenation Approach to Data Augmentation for Neural Machine Translation”, camera-ready for NAACL Student Research Workshop, pp 7, 2021.

[4] 文の構造：https://upwrite.jp/grammar/text_structure, 2023/2/6 時点.

[5] OpenNMT : https://pypi.org/project/OpenNMT-py/1.2.0/

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

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

[8] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Au-tomatic Evaluation of Translation Quality for Distant Language Pairs. Conference on Em-pirical Methods on Natural Language Processing (EMNLP), 2010.

[9] 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.