2020年度（令和2年度）　卒業論文

複文翻訳のための単文による学習データの作成

電気情報系学科 卒業論文検印	
学科長	

指導教員

村田 真樹
村上 仁一

鳥取大学工学部　電気情報系学科
自然言語処理研究室
B17T2061X 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、今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.

今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、学習データとみなし, 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、私達が普段用いる文は複文が多い。そこで今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.

今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、学習データとみなし, 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側の翻訳結果のほうがよい結果が得られた.

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，本研究の手法について述べる．第 4 章では，実験環境について述べる．第 5 章では，実験結果を述べる．第 6 章では，本実験の考察を述べる．第 7 章では，本実験の今後の課題について述べる．第 8 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。

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　提案手法

本章では, 本研究で扱う提案手法の説明を行う.

3.1　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が不可能と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元々ある単文の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.

3.2　学習データの作成手順

以下に, 作成する単文+単文の学習データの手順の説明を行う.

1. まず初めに, 複文のテスト文があるデータからコマンドを用いてランダムに 10 文を抽出する.
2. 抽出した文と学習データの単文との間で類似文検索にかける.
3. 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+単文の学習データを作成する.
4. 作成したデータを元からある単文の学習データとともに NMT に学習させる.

3.3 類似文検索

単文+単文の学習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, 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, その上位数個を出力として出している.

このようにして, 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未だに膨大で学習しきれない場合, 更に類似文検索を行う. これを学習できるデータ量まで行う.

しかし, 元々の複文のテスト文が少ないと何度も類似文検索を行う必要がなく, 一回で済む場合もある.

第4章　実験環境

4.1　使用NMTエンジン

本実験には,Open-NMT[3] を使用した.

4.2　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として, 電子辞書などの例文より抽出した単文データ及び複文を用いる. データの内訳を表 4.2.1 に示す.

表 4.2.1: 学習データ

単文データ	160,000
単文+単文データ	25,000
ディベロップメント文	1,000
複文テストデータ	10

4.3 実験方法と目的

4.3.1 実験方法

NMT を用いて, 複文の翻訳を行う.

行う翻訳の学習データは 2 つで, 学習データ 1 を単文データ 16 万文のみ, 学習データ 2 を単文データ 16 万文+単文+単文データ 2 万 5 千文とし翻訳を行う. そして, それぞれの翻訳結果を人手による対比較評価で行う. 今回自動浄化の結果を用いていないのは, 自動評価の値では実験の評価手法としては, 参考にならない数値が出力されてしまうためである.

4.3.2 目的

本実験の目的としては,2 つの学習データでそれぞれ学習し, 翻訳させた時に提案手法の方がよい翻訳結果となっていることを調査するため.

第5章　実験結果

5.1　出力結果

5.1.1　人手評価結果

複文のテスト文を翻訳させた10文を用いて, 学習データ1(単文データのみ) と学習データ2(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5.1.1 に人手評価で対比較評価を行った結果を示す.

- 学習データ1○　　：学習データ1の方が良い
- 学習データ2○　：学習データ2の方が良い
- 差なし　　：翻訳精度に明確な差がない

表 5.1.1: 学習データ1と学習データ2の評価結果 (10文中)

学習データ1○	学習データ2○	差なし
0文	4文	6文

表5.1.1 の結果より, 学習データ1ち学習データ2の翻訳結果を比較すると, 学習データ2の翻訳結果がよく, 精度の向上が確認できる．また, 学習データ1と学習データ2において翻訳結果に差がない文は10文中6文あった.

5.1.2 翻訳結果

表 5.1.2 に NMT で出力された複文翻訳の結果例を示す.

表 5.1.2: 学習データ 2 ○の出力例

入力文	卒業 式 は 、 3 月 2 2 日 に 行わ れる 予定 で ある 。
参照文	The graduation ceremony is scheduled to be held on March 22 .
学習データ 1	The commencement will be held in the large hall on the 10th .
学習データ 2	The graduation ceremony will be held on March 22 .

表 5.1.2 より, 同一の複文日本文を入力文として, 学習データ 1 と学習データ 2 をそれぞれ学習させたものをそれぞれ翻訳にかけた. 学習データ 1 では, 意味の通らない不適切な文が出力されているのに対して, 学習データ 2 では, ある程度意味が汲み取れるような比較的正しい文が出力された.

第6章　考察

6.1　提案手法の問題点

本節では提案手法の問題点を考察する.

本研究では類似文検索を行った. その手法としては文中に同じ文字同士が含まれている割合を計算し, その値の高い順に出力するという形である. しかし, この手法は精度の高い類似文検索とは言えない. そのため, 翻訳結果の精度自体を改善しようとするためには, 類似文検索の方法について改善が必要であると考える.

6.2　翻訳結果について

本研究の翻訳結果は, 全て一文で出力された. 翻訳結果としては一文で出力されたほうが参照文に近い出力結果となり, 良い翻訳結果と言えるかもしれない. しかし, 単文+単文の学習データを学習させているため,2 文の出力結果が得られても良いはずである. だが, 結果として一文も出力されなかった. これは類似文検索においてより近い類似文を得られ, それらを組み合わせることができればいくつか出力として得られるのではないかと考える.

6.3　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 4 文得られ, どちらとも良くない (差がない) と判断した文は 6 文得られた. 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.

第7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, 複文のテスト文をランダムに 10 文抽出し, 類似文検索, 学習データの作成と翻訳の一連の流れを行ったが, 抽出する文を 50 文や 100 文に増やしたところ, 翻訳結果の精度が非常に低くなってしまった. そのため, 抽出文を増やした時の翻訳精度の改善がある.

- 本実験では単文+単文の学習データの作成において, 全ての類似文同士を組み合わせた. しかし実際は, 一つの単文につき出力された類似文同士を組み合わせだけでよく, そのためのプログラムの作成がある. そうすることにより, 全てを組み合わせた場合よりも遥かに学習データが削減されるため, 多くの複文の類似文検索を行えることになる.

- 本実験では類似文検索において, 同じ文字同士が含まれる割合で行った. しかし,word2vec 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. よって他のツールを用いて類似文検索を行うことを検討したい.

- 本実験において, 複文翻訳の出力結果例は一文のみの結果しか出力されなかった. しかし, 単文+単文の学習データを NMT に学習させているため,2 文の複文翻訳結果が出力されてもよく, そのような結果の出力が出力されるように工夫を行うことを検討したい.

第8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, 学習データが複文の場合において, 提案手法の方法を用いることによって複文翻訳の精度を向上させていきたい.

謝辞

最後に, 一年間に渡り, 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. また, 共同研究させていただいた茨城大学工学部情報工学科の佐々木稔氏を始め, 参考にさせていただいた論文の著者の方々に対して, 深く感謝申し上げます.

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 ”,2017.

[2] 呉浩東： ”機械翻訳の原理と研究動向 ”, 2019.

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