文書単位翻訳モデルを用いた英日小説翻訳の改善

棚橋 優希 [1] 井佐原 均 [2]

豊橋技術科学大学

[1]tanahashi@lang.cs.tut.ac.jp [2]isahara@tut.jp

1 はじめに

近年, ニューラルネットワークを機械翻訳へ応用したニューラル機械翻訳 (以下 NMT) が高い成果を上げている. NMT は従来の統計的機械翻訳のように単語 1 語ごとを逐語的に翻訳するのではなく, 1 文全体を読み込んだ後に翻訳文を生成する. これにより, マニュアルや論文, ニュース記事といった形式の定まった文書のみならず, 小説のような自由度のある文書についても可読性が高い翻訳を行えるようになった. しかし, 既存研究の多くは先述したニュース記事などを翻訳の対象としており, 小説を対象とした機械翻訳の研究はほとんどない.

本研究では, NMT を用いて英日の小説翻訳を行うにあたり, ニュース記事などの翻訳ではあまり問題にならないが, 小説翻訳では大きく影響する問題を洗い出し, その対処法について検討する. 実験では, 小説翻訳における問題の一つである照応先曖昧性問題について, 文書単位翻訳モデルを用いることでこの問題に対処できることを確認する.

2 背景

2.1 小説翻訳

現在, 翻訳モデルの学習に利用できる小説の対訳コーパスは総数が少なく, また著者や翻訳者が統一されていない. そのため, 文章の表現に大きな差異があり, 例えば全く同じ英文に対する日本語訳が作品によって異なっている例がある. これにより, BLEU などの既存手法を用いた定量的な性能の評価が困難になっている.

ゆえに, 本研究では小説翻訳に係る具体的な障害を小問題として設定し, これを解決することで翻訳の改善に繋げる. NMT における既知の問題は複数存在するが, 特に英日の小説翻訳を行う際に影響するものとして特に固有名詞一貫性問題と照応先曖昧性問題を設定する.

2.2 固有名詞一貫性問題

NMT では, 計算時間を抑えるため出力する単語を制限し, 学習コーパス内に存在しない単語や出現頻度が低い単語 (低頻度語) はすべて同一の代替トークン (UNK) に置き換えるという処理を行う. このような置き換えを行なった場合, 低頻度語を翻訳する際に出力される単語について曖昧性が含まれる. 実際に Google 翻訳[*1]を用いて低頻度語を含む文を翻訳した結果を図 1 に示す.

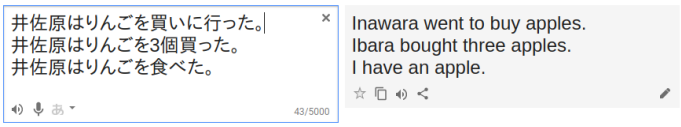

図 1 低頻度語を含む文の翻訳

小説翻訳において, 人名や地名といった低頻度になりがちな固有名詞の翻訳に曖昧性が含まれると, 全体を通して読んだ際にストーリーが破綻してしまう恐れがある. この問題を解決するために, 固有名詞は UNK でなく専用のユニークトークンで置換する, 注意機構を採用している場合注意の重みが最も高い原文側の単語で置換するなどの方法が考えられるが, ベストプラクティスは判明していない.

2.3 照応先曖昧性問題

英日の小説対訳コーパスには, 表 1 に示すような対訳が存在する. ex1 では, 英文側の主語が “You” であるのに対し, 和文では主語である “君は” が抜けている. また, ex2 では英文側の主語が “He” であるのに対し, 和文では固有名詞である “ピーター” が主語になっている. このような照応関係の対訳を含むコーパスで学習を行う場合, 出力に曖昧性が含まれる.

例えば ex1 の和文を英文に翻訳する際, 翻訳モデルは省略された主語が何であるかをこの 1 文から判断す

[*1] translate.google.com

表 1　照応先曖昧性問題の例

	En	Ja
ex1	You are nearly half an hour late.	30 分近く遅れたな
ex2	He did this because there is a saying in the Neverland that.	ピーターがそうしたのは，ネバーランドにはこんなことわざがあったからだ 。

ることはできない．また ex2 において英文から和文へ翻訳する際，学習コーパス内に “He” に対応する名詞の訳が異なる対訳が含まれていると，翻訳モデルは正しく主語を選択することができない．このように，照応関係がコーパス内で一貫していない対訳を含むコーパスで学習を行うことにより，翻訳に曖昧性が生まれる問題を照応先曖昧性問題と定義する．

3　文書単位翻訳モデル

NMT の代表的なモデルとして，エンコーダデコーダモデルからなる seq2seq[6] や，それにアテンション機構を加えた RNNSearch[1] などのモデルが有名である．これらのモデルでは，学習用の対訳を使ってあるソース言語文が与えられた際に正しいターゲット言語文へ翻訳される確率を最大化するように学習を行う．しかし，これらの学習は 1 文単位で行われるため，文同士の繋がりなどの情報は無視される．

近年では，翻訳対象の 1 文だけでなくより大域を考慮した翻訳を実現するための文書単位翻訳 (Document-level Translation) 研究が盛んに行われている．文書単位翻訳モデルでは，従来の翻訳モデルと異なり，現在翻訳対象となっている文だけでなく過去の文の情報 (コンテキスト) も参照して翻訳を行う．これにより，同じ文書内で用いられる単語の一貫性の担保や，多義性を孕む単語の正しい翻訳を行うことを可能にする．

文書単位翻訳モデルには階層 RNN 構造を用いたモデル [8]，階層アテンション構造を用いたモデル [3]，メモリネットワークを用いたモデル [2]，トランスフォーマモデル [7] の文書レベル拡張 [9] など，コンテキストの取得方法が異なる複数のモデルが存在する．

4　目的

照応先曖昧性問題を緩和するための方法として，3 節で述べた文書単位翻訳モデルを用いることを考える．過去の文情報を参照することにより，前文に照応のてがかりが含まれる場合について出力における曖昧性を緩和することができると考えられる．

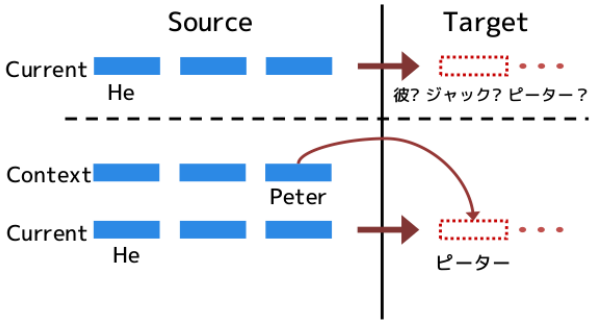

図 2　文書単位翻訳モデルによる照応先曖昧性解消の例

本研究では，実際に Zhang らの文書単位翻訳モデルを用いて実際に小説翻訳を行うことにより，照応先曖昧性問題がどの程度緩和されているかを確かめる．

5　実験

5.1　実験設定

対訳コーパスとして，Project Gutenberg[*2] と青空文庫[*3] より集めた 113 作品を使用する．そのうち開発・評価用の 5 作品 (1,254 文) を除いた 85,408 文の対訳を学習に用いる．

また，今回の実験で用いる Zhang らのモデルでは，文書単位翻訳用モデルを作成する前に小説とは別の大規模な対訳コーパスにて事前学習を行う必要がある．そのため，本研究ではこの事前学習用大規模コーパスとして，ASPEC[4] を使って事前学習を行う．

これらの小説コーパスと大規模コーパスのいずれにも，前処理として SentencePiece[*4] を使ったバイト対符号化 (BPE) によるサブワード正規化 [5] を適用する．ASPEC と小説コーパスでは用いられる語彙に大きな異なりが存在するため，サブワードの学習には翻訳で用いるコーパスとは無関係なデータを使用し，どちらのコーパスにおいても未知トークンの割合を極力下げ

[*2] http://www.gutenberg.org/
[*3] https://www.aozora.gr.jp/
[*4] https://github.com/google/sentencepiece

るように調整する．実際には日本 Wikipedia のダンプデータより 500 万文, 英 Wikipedia のダンプデータより 500 万文を混合した計 1,000 万文のデータを用いる．さらに, サブワードモデルのボキャブラリサイズは 32K に設定する．また, コンテキストの情報を使わない場合との比較を行うため, 小説コーパスのみを用いて学習した翻訳モデル, それに加え大規模コーパスでの事前学習を行なった翻訳モデルでの実験も行なう．

5.2 モデルパラメータ

実験で用いるトランスフォーマモデルでは, 隠れ層の次元数を 512 とし, マルチヘッドアテンションのヘッド数は 8 に設定する．またエンコーダ, デコーダのスタック数は共に 6 にする．ミニバッチ数は 6250 単語/1 ステップとし, 事前学習・本学習共に全部で 20 万ステップの学習を行う．本学習では, 事前学習モデルのうち最も ASPEC 開発セットの値が良かったステップ数のモデルを基に小説コーパスでの学習を行う．今回コンテキストとして扱う情報は, 翻訳対象の文から見て過去 2 文のソース文とする．

5.3 性能評価

実験では, 開発・評価用の 5 作品から開発用の 1 作品を除いた残りの 4 作品について英日翻訳を行う．出力された結果について, Moses の multi-bleu.perl を用いて 4-gram までの BLEU 値を計算する．さらに, 照応先曖昧性問題を緩和できているかを確かめるため, 英文側に He, She といった三人称が現れる文を列挙し, それに対応する日本語訳が誤っていない単語であるかを検証する．評価結果は本発表にて報告する．

6 おわりに

本研究では, NMT を用いた小説翻訳を行うにあたり影響しうる問題として固有名詞一貫性問題と照応先曖昧性問題を定義した．また, そのうち照応先曖昧性問題について, 文書単位翻訳モデルを用いることで問題の影響を緩和できるか検証することを目的とし, 実験の諸設定を行なった．本発表では, 実際の翻訳例を示しつつ文書単位翻訳モデルが問題の改善に寄与できているかを確かめた．

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] Sameen Maruf and Gholamreza Haffari. Document Context Neural Machine Translation with Memory Networks. In *ACL*, pages 1275–1284, 2018.

[3] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *EMNLP*, 2018.

[4] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In *LREC*, pages 2204–2208, 2016.

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

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. *CoRR*, abs/1409.3215, 2014.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, pages 5998–6008. 2017.

[8] Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting Cross-Sentence Context for Neural Machine Translation. In *EMNLP*, pages 2826–2831, 2017.

[9] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the Transformer Translation Model with Document-Level Context. In *EMNLP*, pages 533–542, 2018.