2W-03

英日機械翻訳のための対訳コーパスフィルタリングの検討

本田 志遠† 正木 亮太郎† 梶原 智之‡

† 愛媛大学工学部工学科 ‡ 愛媛大学大学院理工学研究科

1 はじめに

深層学習に基づく自然言語処理では、非常に大規模なコーパスを用いてモデルを訓練することが一般的である。これによって近年の自然言語処理は大きく性能を改善しているが、一方でモデル構築のためのコストは増大している。

例えば機械翻訳では、Common Crawl や WikiMatrix のような、数百万から数千万文対の対訳コーパスを用いて翻訳器を訓練することが多い。しかし、このようなウェブから自動収集されたデータには多くのノイズが含まれる。ノイズを含む訓練データは、統計的機械翻訳などの従来モデルよりも現在のニューラル機械翻訳モデルに対してより有害である場合が多い[1]と言われており、自動収集された対訳コーパスから有害な文対を取り除く対訳コーパスフィルタリング[2]が近年盛んに研究されている。

本研究では、英語から日本語への機械翻訳を対象とする対訳コーパスフィルタリングの手法について検討する。最大規模の英日対訳コーパスである JParaCrawl[3]を用いた実験の結果、提案手法によって訓練データを半減させた場合に、全ての対訳コーパスを訓練に用いるよりも高い翻訳品質を短い訓練時間で達成できた。

2 提案手法

本研究では、英日機械翻訳における対訳コーパスフィルタリングのために、以下の 3 種類のノイズを扱う。各ノイズの例を表 1 に示す。

A) 短すぎる文・長すぎる文
B) 非英語文・非日本語文
C) 意味的に対応しない文対

対訳コーパスフィルタリングの手法として、タイプ B のノイズを除外するために 2 つの手法を考え、合計 4 種類の手法を検討する。その上で、無作為に同規模の文対を除外するベースラインと比べて有効な手法を組み合わせ、最終的な提案手法とする。本研究で検討する 4 種類の対訳コーパスフィルタリング手法を以下に示す。

表 1：対訳コーパスに含まれるノイズの例

ノイズ	対訳文
A	【英】RA: Guy J 【日】RA: Guy J ニュース
B-1	【英】Ding Ye On, Lee Bong In 【日】丁用根、李鳳仁
B-2	【英】H-696 - - Wigs online store 【日】H-696 - - 通販 wigs2you.com
C	【英】You will always need to have the back up 7 computer I was using XP. 【日】私は Photoshop 7 の 2 つのレイヤーを持っています。

A) 文字数が少なすぎる文／多すぎる文を除外

カンマやピリオドなどの記号、スペースや改行などの空白文字を除外した上で、文字数を数える。そして、閾値を超えて短い文や長い文を含む文対を訓練用データから除外する。

B-1) 対象言語だと判定されなかった文を除外

言語判定ツール langdetect[1]を用いて、対訳コーパスに含まれる各文の言語を判定する。そして、ソース文が英語と判定されなかった場合またはターゲット文が日本語と判定されなかった場合に、その文対を訓練用データから除外する。

B-2) 非対象言語の文字の割合が高い文を除外

カンマやピリオドなどの記号、スペースや改行などの空白文字を除外した上で、文字種の割合を数える。そして、ソース文における英字の割合またはターゲット文における平仮名・片仮名・漢字の割合が閾値を下回る場合に、その文対を訓練用データから除外する。

C) 文ベクトルの類似度が低い／高い文対を除外

Multilingual Universal Sentence Encoder (mUSE) [4]を用いて対訳文における文間の意味的類似度を推定する。mUSE は英語や日本語を含む 16 言語に対応した多言語文符号化器であり、各言語の文を共通のベクトル空間で表現できる。

本研究では、ソース文とターゲット文をそれぞれ mUSE によってベクトル化し、それらの余弦類似度によって文間の意味的類似度を推定する。そして、閾値を超えて類似度の高い文対や低い文対を訓練用データから除外する。

Towards Parallel Corpus Filtering for English-to-Japanese Machine Translation

† Shion Honda (g520387b@mails.cc.ehime-u.ac.jp)
Ryotaro Masaki (g520391k@mails.cc.ehime-u.ac.jp)
‡ Tomoyuki Kajiwara (kajiwara@cs.ehime-u.ac.jp)
Ehime University

[1] https://github.com/Mimino666/langdetect

表 2：BLEU による英日機械翻訳の品質評価

	訓練用データの文対数	翻訳品質（BLEU）
フィルタリングなし	1,000 万	17.9
ベースライン（無作為抽出）	500 万	17.7
フィルタリング手法 A	766 万 → 500 万	16.5
フィルタリング手法 B-1	969 万 → 500 万	16.8
フィルタリング手法 B-2	697 万 → 500 万	18.6
フィルタリング手法 C	681 万 → 500 万	18.1
提案手法（B-2 + C）	519 万 → 500 万	18.9

3 実験設定

訓練用データには、最大規模の英日対訳コーパスである JParaCrawl[3]を用いた。本研究では、JParaCrawl に含まれる 1,000 万文対の訓練用データを半分の 500 万文対に削減する実験を行う。検証用および評価用には、WMT20[2]におけるニュース翻訳タスクの対訳データ（検証用 1,998 文対、評価用 1,000 文対）を用いた。

翻訳器には、Joey NMT[3]によって構築した Transformer モデル[5]を用いた。このモデルは、512 次元の埋込層および隠れ層を持つ 6 層の自己注意ネットワークであり、8 個の注意ヘッドを持つ。バッチサイズを 4,096 トークンとし、最適化手法には adam を使用した。前処理には SentencePiece[4]の 1-gram 言語モデル（語彙サイズは 32,000）によるサブワード分割を行った。翻訳品質の評価には SacreBLEU[5]を用いた。

各手法の閾値は、訓練用データと検証用データの比較によって設定した。手法 A では、訓練用データにおいて 40 文字未満の英語文・220 文字以上の英語文・100 文字以上の日本語文の各割合が検証用データよりも高いため、これらに該当する文対を訓練用データから除外した。手法 B-2 では、英字の割合が 90%未満の英語文および平仮名・片仮名・漢字の割合が 85%未満の日本語文を含む文対を訓練用データから除外した。手法 C では、0.4 未満の類似度または 0.75 以上の類似度を持つ文対を訓練用データから除外した。

4 実験結果

実験結果を表 2 に示す。各手法で該当するノイズを除外した訓練用データから、500 万文対を無作為抽出した。例えば、手法 A では、短文や長文を除外した訓練用データが 766 万文対あり、この中から無作為に 500 万文対を選択して翻訳器を訓練した際の翻訳品質が BLEU=16.45 である。

無作為に訓練用データを選択するベースラインと比較して、手法 A および手法 B-1 は翻訳品質が低下したが、手法 B-2 および手法 C は有効であった。これらの 2 手法を組み合わせた提案手法は、訓練用データ全体を用いるよりも BLEU において 1 ポイント高い翻訳品質を達成した。

提案手法は、翻訳品質の改善に加えて、訓練時間の短縮にも貢献した。TITAN RTX の GPU1 枚を用いる訓練において、検証用データの BLEU が収束するまでに 1,000 万文対の訓練用データ全体を用いる場合には 21 万ステップ（約 175 時間）の訓練が必要だが、提案手法によって 500 万文対に削減した場合には 12.4 万ステップ（約 103 時間）と、約 40%の訓練時間を短縮できた。

5 おわりに

本研究では、英日機械翻訳のための対訳コーパスフィルタリングの手法について検討した。1,000 万文対の訓練用データを半減させる評価実験において、非英語文や非日本語文の除外および意味的に対応しない文対の除外が有効であることを確認した。今後の課題として、他のドメインにおける検証や、他の言語判定ツールや文符号化器を用いた検証を進める予定である。

参考文献

[1] Huda Khayrallah, Philipp Koehn. On the Impact of Various Types of Noise on Neural Machine Translation. In Proc. of WNGT, pp.74-83, 2018.
[2] Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, Francisco Guzmán. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In Proc. of WMT, pp.726-742, 2020.
[3] Makoto Morishita, Jun Suzuki, Masaaki Nagata. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In Proc. of LREC, pp.3603-3609, 2020.
[4] Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, Ray Kurzweil. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. In Proc. of RepL4NLP, pp.250-259, 2019.
[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention Is All You Need. In Proc. of NIPS, pp.5998-6008, 2017.

[2] https://www.statmt.org/wmt20/translation-task.html
[3] https://github.com/joeynmt/joeynmt
[4] https://github.com/google/sentencepiece
[5] https://github.com/mjpost/sacrebleu