英語版医療機器不具合用語集を対象とした深層学習による機械翻訳の精度評価

谷川原 綾子[*1], 横井 英人[*2], 上杉 正人[*3]
[*1] 北海道科学大学 保健医療学部, [*2] 香川大学医学部附属病院 医療情報部
[*3] 北海道情報大学 医療情報学部

Machine translation of English medical device adverse event terminology using deep learning

Ayako Yagahara[*1], Hideto Yokoi[*2], Masahito Uesugi[*3]
[*1] Faculty of Health Sciences, Hokkaido University of Science
[*2] Dept. of Medical Informatics, Kagawa University Hospital
[*3] Hokkaido Information University

抄録：医療機器不具合用語集の日本版(JFMDA 用語集)と国際版(IMDRF 用語集)のマッピング効率化に向けたシステム構築を進めている。本研究では，深層学習を用いて IMDRF 用語集の自動翻訳を実施し，その精度評価を行った。翻訳用モデルとして，sequence-to-sequence ベースの学習済み公開モデルである mBART と 100 言語の翻訳が可能なモデルである transformer ベースの m2m-100,Open AI が公開している GPT-3，Google が公開している googletrans を取得した。加えて，医療機器関連対訳コーパスからオリジナルの翻訳モデル，mBART と m2m-100 をファインチューニングしたモデルも生成した。IMDRF 用語集の対訳文からテストデータを抽出し，各モデルにおける翻訳精度を評価したところ，googletrans の BLEU スコアが 27.3 と最も高く，目視評価でも 78%と最良の翻訳品質と判定された。GPT-3 では、目視評価にて googletrans に次ぐ 76%であった。mBART50 はファインチューニングにより BLEU はわずかに向上したが，目視評価にて品質は低下と判断された。m2m-100 は，ファインチューニングしたモデルにて BLEU が低下し，品質も低下した。自作モデルは BLEU が最低となり，目視評価でも最低の品質と判定された。

キーワード Transformer, sequence-to-sequence **機械翻訳，医療機器不具合用語集**

1. はじめに

本邦では、医療機器産業連合会(JFMDA)より医療機器不具合報告における標準用語の使用を推進するため、不具合用語集(JFMDA 用語集)を公開している[1]。一方、国際的には、国際医療機器規制当局フォーラム(IMDRF)からも、医療機器の不具合、有害事象等の標準コードを収載した用語集(IMDRF 用語集)が英語で公開されており[2]、米国や欧州における有害事象報告において採用されている。国内の報告書提出においては、JFMDA 用語集の使用が優先されるが、IMDRF 用語集との相互運用も国際整合の観点から重要とされ、JFMDA にて用手的なマッピング作業を進めている。両用語集は定期的に更新されるため、その都度、マッピングに多大な人的リソースと時間が必要となる。我々は、マッピングの効率化に向けて、コンピュータによる支援システムの構築を進めており、本研究では、そのシステムの一部である深層学習を用いた英日自動翻訳アルゴリズムの検討を行った。

2. 方法

1) 対訳データの生成

自作モデル生成用のコーパスとして，医薬品医療機器総合機構が提供している医療機器基準データベースの一般的名称とその定義文から対訳データ(JMDN データ)を生成した。加えて，IMDRF 用語集の対訳データを取得し，JMDN データと合わせたオリジナル対訳データを作成した。この対訳データは SentencePiece によるサブワード化が行われた。訓練用，検証用，テスト用データの生成では，IMDRF 対訳データをそれぞれランダムに 70:15:15 の割合で分割した。JMDN データは全て訓練用データとした。

2) 翻訳モデルの取得

本研究では，学習済みモデル，1)の訓練データでファインチューニングされた学習済みモデル，

自作モデルを使用した。学習済みモデルは，googletrans[3], Facebook AI research が公開した mBART [4], m2m-100 [5]、OpenAI が開発している GPT-3 を使用した。 m2m-100 については 418M パラメータモデル(m2m-100-418M)と 1.2B パラメータモデル(m2m-100-1.2B)を使用した。

3) 公開モデルのファインチューニングと自作モデルの生成

2)で取得した Facebook AI research から公開されている 3 つの翻訳モデルに対して，1)で生成した訓練用と検証用データを用いてファインチューニングを実施した。自作モデルの生成においても上記と同じデータを使用した。学習には，Facebook が提供している fairseq[6]を使用した。

4) 自動翻訳の精度評価

全モデルのテストデータを各モデルに入力し，翻訳生成文を得た。この生成文を Bilingual Evaluation Understudy (BLEU) にて評価した。

BLEU に加え，テストデータからランダムに 50 文抽出し，評価者 2 名による目視評価も実施した。意味的な整合性が取れていると判断された生成文の割合を算出した。

3. 結果・考察

1) 精度評価

各モデルの精度は Table.1 に示す．すべての指標において最良の精度を示したのは googletrans であった。BLEU では、googletrans に次いで高値であったモデルは，m2m-100-1.2B であった。目視評価では googletrans に次いで，GPT-3 の値が高かった。オリジナル対訳データのみにて作成されたモデルは，すべての指標において最低値を示した。この原因は，学習データの量が十分ではなかったためと考えている。

2) 翻訳結果の特徴

公開モデルにおいて，多義語の訳を間違える例が見られた。例えば，“Vessel perforation(訳：血管穿孔)”を googletrans が「船舶用穴あけ」と誤訳した。また、英単語の発音をカタカナ表記で訳されることも多かった。例として，“Aphonia (訳：失声症)”を「アフォニア」と出力していた。これらの事例に対して、ファインチューニングされたモデルは，正答もしくは正答に近い訳が出力された。

ファインチューニングされた全てのモデルは，ファインチューニング前と比較して，目視評価にて妥当と判断される生成文が大幅に減少した。その原因は、一部の単語が誤った単語に置き換わってしまったためであった。今後、この原因について調査を進めたいと考えている。

Table.1 機械翻訳の評価

	BLEU	目視評価
Transformer (オリジナルコーパス)	11.3	10%
mBART50	15.1	62%
mBART50 (finetuning あり)	21.2	34%
m2m-100-418M	17.0	62%
m2m-100-418M (finetuning あり)	14.8	18%
m2m-100-1.2B	21.4	62%
m2m-100-1.2B (finetuning あり)	13.5	18%
GPT-3	20.6	76%
googletrans	**27.3**	**78%**

参 考 文 献

[1] 日本医療機器産業連合会: 医療機器不具合用語集の活用について. Available: https://www.jfmda.gr.jp/activity/committee/fuguai/.

[2] The International Medical Device Regulators Forum: Adverse Event Terminology. Available: https://www.imdrf.org/working-groups/adverse-event-terminology.

[3] googletrans. Available: https://pypi.org/project/Googletrans/.

[4] Tan Y, Tran C, Li X, et al: Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, arXiv, 2020.

[5] Fan A, Bhosale S, Schwenk H, et al: Beyond English-Centric Multilingual Machine Translation, arXiv, 2020.

[6] Ott M, Edunov S, Baevski A, et al: fairseq: A Fast, Extensible Toolkit for Sequence Modeling, arXiv, 2019. Available: https://github.com/pytorch/fairseq.