Enhanced Translation of Biomedical Texts via Domain Specific Embeddings

Aaron Lin
UC Berkeley School of Information
aaronlin@ischool.berkeley.edu

Abstract

Domain adaptation is a current issue in Neural Machine Translation (NMT) due to distinct vocabularies unique to different technical disciplines. In this paper, we present an approach to integrate domain knowledge from biomedical contexts using word embeddings. Experimental results show that incorporating an embedding with the existing tokenizer vocabulary improves performance as reflected by a nearly 5 percentage point increase in BLEU while the usage of the embedding in combination with extending the tokenizer vocabulary vocabulary using named entities caused drastic degradation in translation quality due to the low frequency of specific domain vocabulary across sentences.

1 Introduction

Neural Machine Translation (NMT) has the potential to unlock information across technical disciplines. In recent years, domain adaptation has become a hot topic for NMT research for its practical use in making information accessible across linguistic boundaries. One domain where linguistic barriers remain prevalent is the biomedical domain, where the English is the primary language of most texts which can present challenges for both researchers and medical professionals (Bawden et al., 2019). However, the predominantly monolingual nature of academic papers written in English presents a substantial challenge for non-native speakers, ranging from discouraging researchers from publishing their findings to, in extreme cases, contributing to medical malpractice (Rezaeian, 2015).

Accurate translation of domain nomenclature remains a prominent issue in NMT due to the low frequency of the occurrence of highly specific technical terms, as well as significant linguistic differences. One example from Xu et al. (2020) is the word"ethyl acetate" which translates to "乙酸乙酯" where the second set of characters "乙酯" corresponds to "ethyl" and "乙酸" corresponds to acetate.

The aim of this work is to improve translation of domain-specific vocabulary via the incorporation of pretrained biomedical embeddings with existing transformer models for NMT. By incorporating pretrained embeddings alongside transfer learning on a specific biomedical dataset, we hope to improve the translation of biomedical texts.

2 Background

Transformers have revolutionized the field of neural machine translation (NMT), establishing themselves as state-of-the-art due to their ability to capture long-range dependencies while requiring less time to train than their predecessors due to being more parallelizable (Vaswani et al., 2017). Unlike RNNs or LSTMs which process words sequentially, Transformers use a self-attention mechanisms that allow the model to generate direct connections between all words in a sentence, regardless of their position. Because of this, transformers are able to process complex sentence structures more efficiently and with greater accuracy.

Named entities refer to objects that can be identified by a proper name, and previous works use named entity recognition (NER) to preprocess data with named entity classes and boundary tags to increase the quality of translation (Li et al., 2018). Other studies have incorporated domain knowledge into NMT using lexicons(Wang et al., 2022) and ontologies(Remy, François and De Jaeger, P. and Demuynck, Kris, 2022). Different from these approaches, in this paper we use a biomedical em-

bedding to leverage vector similarities between similar terminologies. While previous studies have utilized embeddings to enhance the performance of NMT systems in low-resource settings (Qi et al., 2018), this study goes further by incorporating a Named Entity Recognition (NER) model to expand the vocabulary of our NMT model's tokenizer with domain-specific terms.

In this paper, we incorporate domain knowledge alongside using an existing biomedical embedding with an additional experimental variable of extending the model vocabulary with named entity recognition to improve vocabulary translation.

3 Methods

3.1 Models

The open-source MarianMT[1] model for English to Chinese translation was selected for this project as an already proven model developed by Junczys-Dowmunt et al. (2018) on the Microsoft Translator Team. The model itself is an encoder-decoder Transformer architecture with 6 layers for each component respectively and static positional embeddings.

In this study, the standard MarianMT model with transfer learning applied to the train dataset will be used as a baseline for comparison. There will be two additional experimental models; **Model I** which comprises the standard model with augmented embedding layer and **Model II** which utilizes the standard model and augmented embedding layer in combination with a tokenizer vocabulary extended by named entities. The experimental setups can be seen in Figure 1

3.2 Domain Knowledge Integration

To integrate biomedical domain knowledge into the model, an embedding layer was constructed using a pretrained embedding[2] from Zhang et al. (2019), which was created using data from PubMed and the Medical Information Mart for Intensive Care (Mimic-III) database. For both experimental models, the

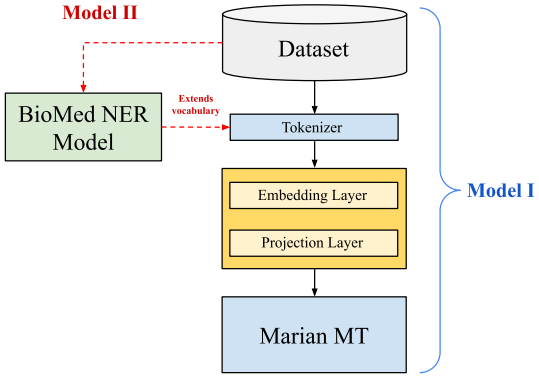

Figure 1: The framework of the proposed approaches. **Model I** indicated on the right in blue details the construction of domain knowledge integration via an embedding layer initialized using BioWordVec embeddings followed by a linear layer used to project the embeddings into the model space. **Model II** indicated in red on the left shows the use of the BioMed NER Model to extend the model vocabulary using the named entity outputs

cleaned vocabulary from the MarianMT tokenizer was passed through BioSentVec prior to training to align the tokens with the pretrained medical embeddings and map them in the vocabulary. In **Model II**, the NER-enhanced model, named entities extracted from the dataset during pre-processing were added to the model vocabulary and included during the embedding layer initialization. The intention of this inclusion is to allow the model to incorporate domain information for more unique tokens by extending the tokenizer vocabulary.

3.3 Datasets

A cleaned parallel corpus constructed of sentence pairs from articles The New England Journal of Medicine (NEJM) biomedical texts in the Chinese-English language pair was obtained with a train and test sets comprised of 62127 and 2102 sentence pairs respectively.

A DeBERTa-based named entity recognition (NER) model pre-trained on sentences in biomedical contexts was used to extract named entities from the training set[3](Venkatd, 2023). These named entities were then deduplicated . The distribution of the top 10 entity categories after de-duplication is presented in Appendix A.1.

[1] https://huggingface.co/Helsinki-NLP/opus-mt-en-zh

[2] https://github.com/ncbi-nlp/BioSentVec The updated BioWordVec is available here alongside the BioSentVec model in word2vec bin format

[3] https://huggingface.co/venkatd/BioMed_NER

Model	BLEU(%)	Brevity Penalty	BERTscore	TER
Baseline	32.9	0.803	0.827 ± 0.074	46.6
Model I	38.3	0.924	0.840 ± 0.099	46.4
Model II	4.42	0.354	0.688 ± 0.113	79.0

Table 1: The BLEU, BERTscore, and TER for the three models. Note that BERTscore is represented by the mean and standard deviation across the dataset.

3.4 Evaluation

To evaluate the performance of the baseline and two experimental models, three metrics were selected: BLEU score, BERTScore, and Translation Edit Rate (TER). BLEU measures precision by assessing the overlap between the generated and reference translations (Papineni et al., 2002). BERTScore evaluates semantic similarity using contextual embeddings (Zhang* et al., 2020), and TER quantifies the edit distance needed to match the reference translation, with lower scores indicating fewer necessary corrections (Snover et al., 2006). This combination of metrics captures complementary aspects of translation quality: precision, semantic understanding, and accuracy.

4 Results and Discussion

4.1 Results

Based on the results in Table 1, the best-performing model was **Model I** (MarianMT extended with BioSentVec embeddings), which achieved a BLEU score improvement of approximately 5.4 percentage points compared to the baseline model. In terms of BERTscore, the baseline and **Model I** demonstrated slightly improved mean performance, with **Model I** increasing by .013. However, the standard deviation of the BERTscore increased by 0.025 in **Model I**, indicating that translation quality may be less consistent than in the baseline. The TER decreased slightly by around 0.2, suggesting that the total number of edits required to correct the translations to the ground truth remained largely unchanged. BLEU scores reflect precision in translation, BERTScore captures semantic similarity, and TER roughly evaluates the accuracy of the translation. From these interpretations, the results indicate that **Model I** had translations that were more precise and had similar levels semantic similarity and accuracy compared to the baseline.

By far the worst performance was observed in **Model II**, which had a BLEU score nearly an order of magnitude smaller than both other models at just 4.42 percentage points. The BERTscore also decreased by .15 and the TER increased by over 32 compared to the Baseline and Model I. These metrics indicate that **Model II** translations were drastically worse in terms of precision and accuracy while having greatly distorted semantic similarity in relation to the ground truth translations.

4.2 Discussion

The sharply decreased BLEU score for **Model II** can be attributed to the difference in brevity penalty (BP) shown in Table 1 between **Model II** (0.354) and the other models, Baseline (0.803) and **Model I** (0.924). A BP value of 0.354 for **Model II** indicates that its translations are significantly shorter than the reference translations, which correspondingly leads to a decreased BLEU score. As an example, for the english input sentence "the safety and antileukemic activity of asciminib in patients with Philadelphia chromosome @-@ positive leukemia are unknown" the generated translation was "在患者中的和尚未确定" which would represent the meaning as "in patients and have not yet been determined". In this case, the significant biomedical vocabulary such as "antileukemic", "asciminib", and "Philadelphia chromosome" were completely omitted from the target translations.

Model II's lower BLEU score and brevity penalty suggest that the model struggles to generate complete translations, likely due to the low-frequency vocabulary introduced by adding the named entities that were extracted from the training set. Since the majority of the named entity words appear infrequently in the training data (99% of the words ap-

pear less than 100 times), the model may not adequately capture their meaning or generate them in the output. This results in shorter translations that lack important entities, leading to a lower BLEU score and incomplete translation compared to the reference. In contrast, **Model I** and the Baseline model, which do not rely on these low-frequency entities, produce more balanced translations with higher brevity penalties and more reliable BLEU scores.

In the comparison between **Model I** and the baseline, the results are consistent with expectations that the incorporation of the domain knowledge. **Model I**'s improved performance stems from its ability to integrate biomedical domain knowledge effectively, which enhanced semantic similarity and precision in translations while maintaining a reasonable brevity penalty and edit rate. These improvements highlight the benefit of incorporating domain-specific embeddings into machine translation pipelines.

Target	**Baseline**	**Model I**
在**本试验**纳入**的全部**150 例患者中, 105 例 (70%) 接受过至少 3 种 TKI **治疗**.	在纳入的 150 例患者中, 105 例 (70%) 接受过至少 3 次 TKI .	在纳入的 150 例患者中, 105 例 (70%) 接受过至少 3 次 TKI **治疗**.

Table 2: An example of the generated translations with the target on the left, **Baseline** in the middle, and **Model I** onf the right.

The example of the translations in Table 2 demonstrates the improvement of **Model I** over the Baseline. Although the outputs are nearly the same, the Baseline lacks the characters "**治疗**" which means "treatment" whereas these characters were correctly added to the end of the **Model I** generated outputs. However, it is also notable that both the Baseline and **Model I** excluded "**本试验**" which means "this experiment" as well as "**的全部**" which means "all of them". This suggests that while the embedding layer enhances domain-specific accuracy, challenges remain in ensuring full content coverage for technical terms. These omissions could point to limitations in the training data or embedding initialization.

5 Conclusion

In this paper we performed domain adaptation using a pre-trained biomedical embedding from BioWordVec alongside transfer learning using the MarinMT English to Chinese model and found that the incorporation of the embedding increased performance with an improved BLEU score by nearly 5 percentage points. Experiments to extend the model vocabulary using named entities identified from the training corpus in combination with integrating the biomedical embedding yielded severely decreased transformation quality, likely due to the infrequency of the named entities added to the model vocabulary. Future work could look at expanding these findings over a larger dataset. Additional strategies, such as adjusting sequence length penalties or leveraging external knowledge sources during decoding, could further improve translation completeness and fidelity.

Limitations

This paper mainly addresses translation in the English to Chinese direction. The experimental models will not work in translating from Chinese to English unless an existing biomedical embedding is provided. Training the model with the extended vocabulary is GPU intensive and will drastically increase runtime.

Ethics Statement

This project is focused on the development of machine translation systems for biomedical texts, with a focus on improving the translation of domain-specific vocabulary. The ethical considerations of this work include ensuring responsible use of artificial intelligence, mitigating risks of misuse, and promoting accessibility in scientific communication. The datasets employed in this research are sourced from publicly available and ethically curated repositories. No sensitive or private data is used, all data was originally sourced from a reputable journal.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie

Névéol, Felipe Soares Mariana Neves, Karin Verspoor Amy Siu, and Maika Vicente Navarro. 2019. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Remy, François and De Jaeger, P. and Demuynck, Kris. 2022. Taming large lexicons : translating clinical text using medical ontologies and sentence templates. In *EmP : 1st RADar conference on Engineer meets Physician, Proceedings*, page 5.

Mohsen Rezaeian. 2015. Disadvantages of publishing biomedical research articles in english for non-native speakers of english. *Epidemiol Health*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Venkatd. 2023. Biomed$_n$*er*. . Accessed: 2024-12-08.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Tingjun Xu, Junhong Zhou Weiming Chen, Jingfang Dai, Yingyong Li, and Yingli Zhao. 2020. Neural machine translation of chemical nomenclature between english and chinese. *Journal of Cheminformatics*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yijia Zhang, Zhihao Yang Qingyu Chen, Hongfei Lin, and Zhiyong Lu. 2019. Bioword-vec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*.

A Appendix

A.1 Named Entity Data

Distribution of named entities identified in training set by group.

Entity Group	Count
Detailed description	14676
Diagnostic procedure	9421
Lab value	5498
Sign symptom	4932
Medication	4268
Date	3496
Biological structure	3050
Therapeutic procedure	2497
Disease disorder	2058
History	1895

Table 3: The counts of named entities added to the tokenizer vocabulary by group. Top 10 categories are displayed in descending order.