Enhanced Translation of Biomedical Texts via Domain Specific Embeddings

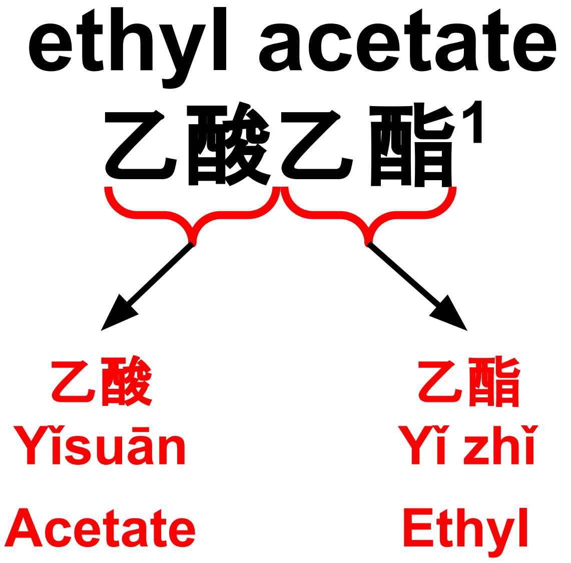
Presenter: Aaron Lin

DATASCI 266

December 10, 2024

Berkeley
UNIVERSITY OF CALIFORNIA

Domain Vocabulary is Difficult to Translate

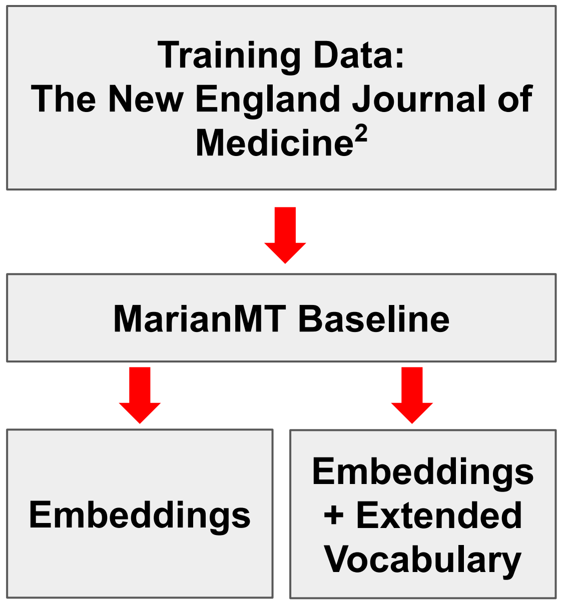


ethyl acetate

乙酸乙酯[1]

乙酸
Yǐsuān
Acetate

乙酯
Yǐ zhǐ
Ethyl



Training Data:
The New England Journal of Medicine[2]

MarianMT Baseline

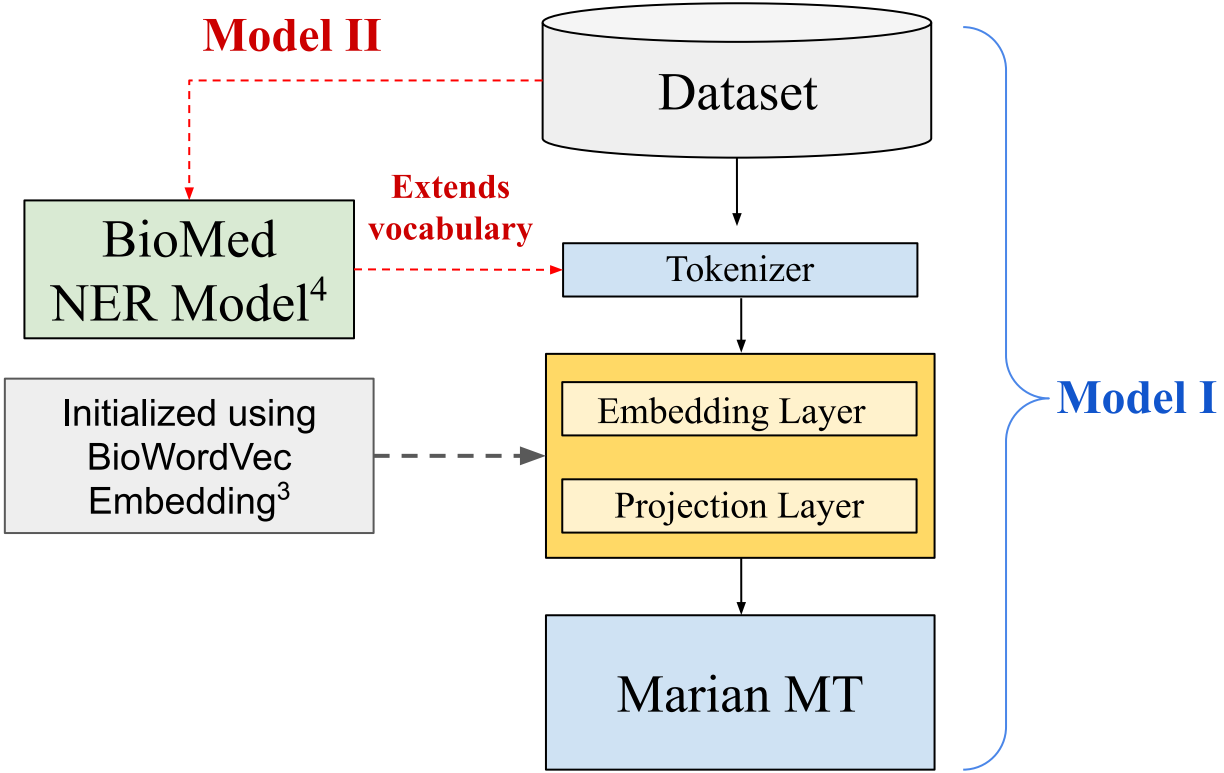
Embeddings

Embeddings + Extended Vocabulary

[1] Example from https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00457-0

[2] Data from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01621-8

Embeddings Are Used to Integrate Domain Vocabulary



[3] Updated version of BioWordVec available here alongside BioSentVec embedding https://github.com/ncbi-nlp/BioSentVec

[4] The existing biomedical NER model is available at https://huggingface.co/venkatd/BioMed_NER

Embedding Layer Improved Precision of Translations

Model	BLEU(%)	Brevity Penalty	BERTscore	TER
Baseline	**32.9**	.803	0.827 ± 0.074	46.6
Model I	**38.3**	.924	0.840 ± 0.099	46.4

Of all 150 patients enrolled, 105 (70 %) had received at least three previous TKIs .

Ground Truth	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3TKI .	在 纳 入 的150 例 患 者中, 105 例 (70%) 接受过至少 3 次TKI **治疗**.

Meaning: “treatment”

NMT Models Still Omitted Important Domain Vocabulary

Model	BLEU(%)	Brevity Penalty	BERTscore	TER
Baseline	**32.9**	.803	0.827 ± 0.074	46.6
Model I	**38.3**	.924	0.840 ± 0.099	46.4

Of all 150 patients enrolled, 105 (70 %) had received at least three previous TKIs .

Ground Truth	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3TKI.	在 纳 入 的150 例 患 者中, 105 例 (70%) 接受过至少 3 次TKI 治疗.

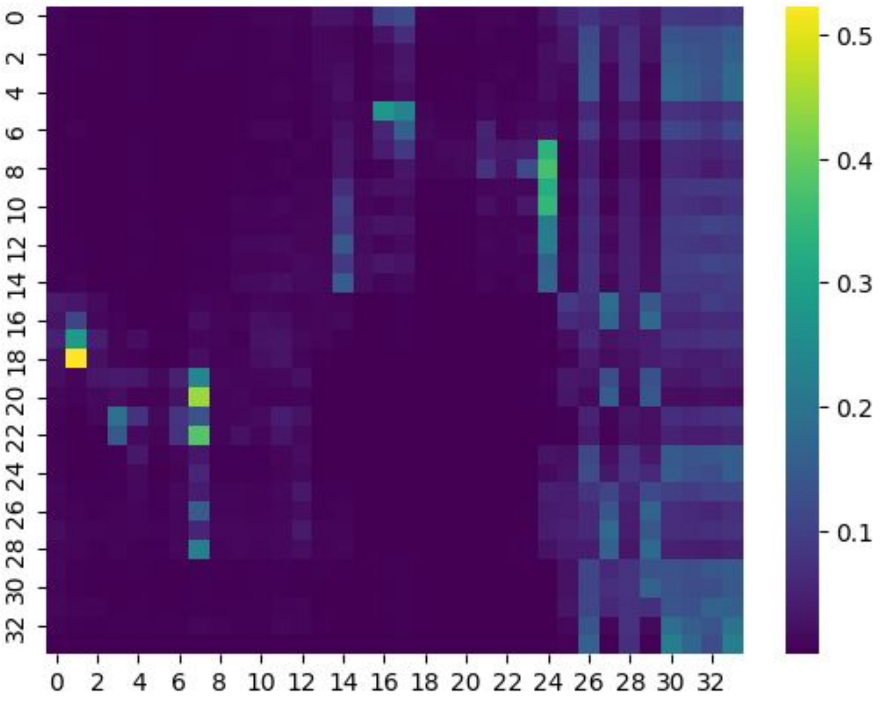
Meaning: “This experiment” and “All of them”

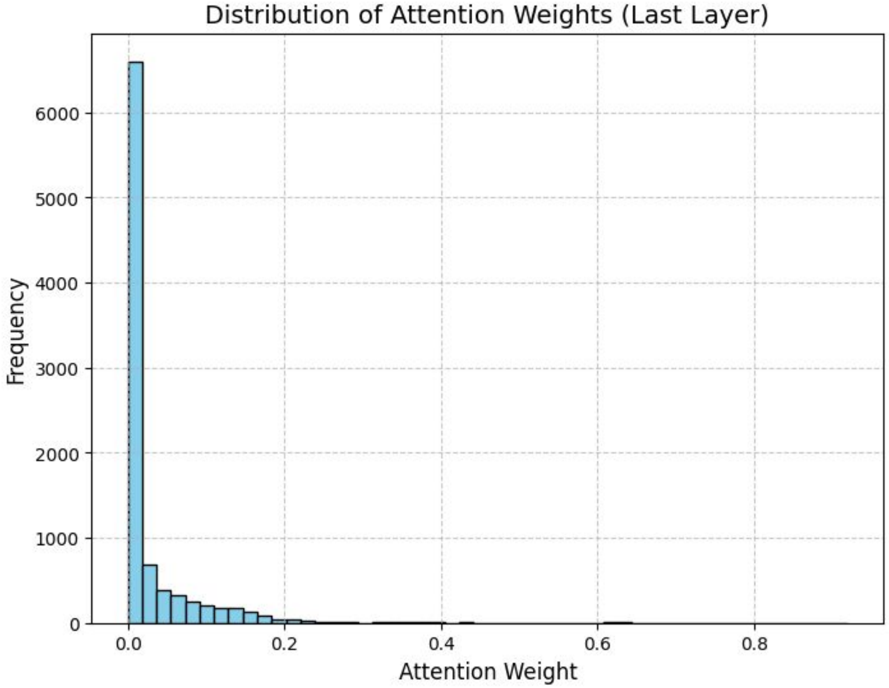
Extended Vocabulary Led to Fragmented Translations

Model	BLEU(%)	Brevity Penalty	BERTscore	TER
Baseline	32.9	.803	0.827 ± 0.074	46.6
Model I	38.3	.924	0.840 ± 0.099	46.4
Model II	4.42	.354	0.688 ± 0.113	79.0

Ground Truth	asciminib 用于 费城 染色体 阳性 白血病 患者 的 安全性 和 抗 白血病 活性 尚未 明确.	The safety and antileukemic activity of asciminib in patients with Philadelphia chromosome @-@ positive leukemia are unknown
Model II	在患者中的和尚未确定	In patients and have not yet been determined

Sparse Attention Weights Lead to Poor Translations





Conclusions and Future Work

Conclusions

- Incorporating biomedical embeddings boosted translation quality with a BLEU score improved by 5.4%
- Using NER to extend model vocabulary lead to drastically decreased model performance

Future Work

- Incorporation of knowledge graphs and dynamic knowledge selection

[4] Interesting paper about dynamic knowledge graph guided NMT https://pmc.ncbi.nlm.nih.gov/articles/PMC10240026/

References (Powerpoint Only)

[1] 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.

[2] Boxiang Liu and Liang Huang. 2021. Paramed: a parallel corpus for english–chinese translation in the biomedical domain. BMC Medical Informatics and Decision Making

[3] 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

[4] Venkatd. 2023. Biomedner. . Accessed: 2024-12-08

[2] Updated version of BioWordVec available here alongside BioSentVec embedding https://github.com/ncbi-nlp/BioSentVec