Improving Vietnamese-English Medical Machine Translation

Nhu Vo[1], **Dat Quoc Nguyen**[2], **Dung D. Le**[1], **Massimo Piccardi**[3], **Wray Buntine**[1]
[1]College of Engineering and Computer Science, VinUniversity, Vietnam;
[2]VinAI Research, Vietnam; [3]University of Technology Sydney, Australia
[1]{nhu.vd, dung.ld, wray.b}@vinuni.edu.vn; [2]v.datnq9@vinai.io; [3]massimo.piccardi@uts.edu.au

Abstract

Machine translation for Vietnamese-English in the medical domain is still an under-explored research area. In this paper, we introduce MedEV—a high-quality Vietnamese-English parallel dataset constructed specifically for the medical domain, comprising approximately 360K sentence pairs. We conduct extensive experiments comparing Google Translate, ChatGPT (gpt-3.5-turbo), state-of-the-art Vietnamese-English neural machine translation models and pre-trained bilingual/multilingual sequence-to-sequence models on our new MedEV dataset. Experimental results show that the best performance is achieved by fine-tuning `vinai-translate` (Nguyen et al., 2022b) for each translation direction. We publicly release our dataset to promote further research.

Keywords: Vietnamese, English, Medical Machine Translation

1. Introduction

Almost all medical universities and hospitals in Vietnam use Vietnamese in their teaching and practices. Additionally, the majority of specialized educational materials created for students, doctors, and nurses are in English. Even though some undergraduate and many higher-degree medical programs now incorporate English, learners are still required to use Vietnamese in their daily professional interactions. Thus, the demand for high-quality Vietnamese-English medical machine translation (MT) has increased significantly.

For training an MT model, a suitable parallel dataset is needed (El-Kishky et al., 2020; Schwenk et al., 2021; Nguyen et al., 2022c). Previous Vietnamese-English data comes from publicly available resources (Tiedemann, 2012; Cettolo et al., 2015), and then a particular methodology for creating parallel sentences is followed (Doan et al., 2021; Nguyen et al., 2022a). There are two prominent high-quality and large-scale Vietnamese-English parallel datasets that have been made publicly available to date: PhoMT (Doan et al., 2021) and MTet (Ngo et al., 2022). However, PhoMT does not contain pairs from the medical domain, while MTet contains 13,410 medical sentence pairs. In addition, the COVID-19 - HEALTH Wikipedia dataset contains 4,273 Vietnamese-English sentence pairs in COVID-19 news.[1] These numbers of medical sentence pairs are small for high-quality medical translation training. This is a compelling motivation for the development of a dedicated high-quality Vietnamese-English parallel dataset to bridge the gap in the available resources for machine translation in the medical domain.

In this paper, as our first contribution, we introduce the MedEV dataset, a high-quality Vietnamese-English parallel corpus containing 358.7K sentence pairs in the medical domain. As our second contribution, we conduct a comprehensive empirical investigation using the MedEV dataset to improve the performance of neural machine translation (NMT) models within the medical health domain. In particular, we compare the performance of medical text translation among various translation tools and models, including Google Translate, ChatGPT (gpt-3.5-turbo), state-of-the-art Vietnamese-English NMT models, and pre-trained bilingual/multilingual sequence-to-sequence models. To the best of our knowledge, this marks the first empirical study focusing on Vietnamese-English medical machine translation.

We make the MedEV dataset publicly available for research and educational purposes.[2] We hope that MedEV, along with our empirical study, will serve as a foundational resource for future research and applications in the field of Vietnamese-English medical machine translation.

2. Our MedEV Dataset

Developing our MedEV dataset involves three main stages. First, we collect parallel document pairs in the medical domain and then preprocess the collected data. Second, we perform the alignment of parallel sentences within pairs of parallel documents. Last, we perform post-processing steps, which include removing duplicate sentences and manually verifying the quality of the validation and test splits.

[1] `https://www.elrc-share.eu`

[2] Our MedEV dataset is publicly available at: `https://huggingface.co/datasets/nhuvo/MedEV`.

Genre	Total		Training				Validation				Test			
	#doc	#pair	#doc	#pair	#en/s	#vi/s	#doc	#pair	#en/s	#vi/s	#doc	#pair	#en/s	#vi/s
Article Abstracts	22,580	196,276	21,397	186,528	23.87	31.67	583	4,883	24.38	32.32	600	4,865	24.51	32.24
MSD Manuals	2,796	123,302	2,652	117,101	25.96	36.39	87	3,079	26.98	37.98	57	3,122	26.67	37.30
Thesis Summaries	783	23,084	731	21,940	30.27	36.9	35	571	28.83	37.33	17	573	25.40	31.92
Article Translations	1,059	16,134	1,000	15,328	25.92	34.56	31	406	25.57	33.83	28	400	26.30	35.41
All	27,218	358,796	25,780	340,897	25.09	33.76	736	8,939	25.61	34.65	702	8,960	25.4	34.12

Table 1: Statistics of our MedEV dataset. **#doc**: The number of parallel document pairs. **#pair**: The number of parallel sentence pairs. **#en/s**: The average number of word tokens per English sentence. **#vi/s**: The average number of syllable tokens per Vietnamese sentence.

2.1. Data collection and pre-processing

We collect 27,218 parallel document pairs from publicly accessible resources across four genres, including: (i) 22,580 bilingual Vietnamese-English abstracts derived from scientific articles published in medical, clinical, and pharmaceutical journals based in Vietnam; (ii) 2,796 English documents and their corresponding Vietnamese-translated versions from the MSD Manuals website;[3] (iii) 783 bilingual Vietnamese-English summaries extracted from doctoral dissertations from official websites of medical universities in Vietnam; and (iv) 1,059 English scientific articles and their Vietnamese translations, completed by Vietnamese medical doctors.

Here, these document pairs are available in either HTML web pages or in PDF/DOC/DOCX files. To process HTML web pages, we crawl and extract parallel text pairs using the DownThemAll[4] tool and the "BeautifulSoup" library.[5] For PDF/DOC/DOCX files, we download and convert them into the plain text format.[6] Afterward, we manually eliminate unnecessary elements such as headers, footers, footnotes, and page numbers from articles, and then extract the bilingual abstract/summary pairs.

To extract sentences for the next stage of parallel sentence alignment, we automatically segment each text document into sentences, using the Stanford CoreNLP toolkit for English (Manning et al., 2014) and the VnCoreNLP toolkit for Vietnamese (Nguyen et al., 2018; Vu et al., 2018).

2.2. Sentence pair alignment

Following the PhoMT alignment approach (Doan et al., 2021), we align parallel sentences within a parallel document pair, as follows: (**1**) Translate each English source sentence into Vietnamese by using the pre-trained model `vinai-translate` (Nguyen et al., 2022b).[7] (**2**) Align English-Vietnamese sentence pairs via an "intermediate" alignment between the Vietnamese-translated versions of the English source sentences and the Vietnamese target sentences. This is done by using alignment toolkits Hunalign (Varga et al., 2005) and Bleualign (Sennrich and Volk, 2011). (**3**) Select sentence pairs that were aligned by both of these alignment toolkits.

Hunalign and Bleualign include 99% and 95% of Vietnamese/English sentences from our raw dataset into their output, respectively, resulting in an alignment coverage rate of 93+% of Vietnamese/English sentences to be included in the alignment output of about 390K sentence pairs.

2.3. Data post-processing

Out of the 390K English-Vietnamese sentence pairs generated in the previous stage, we exclude 14K sentence pairs with SacreBLEU scores (Post, 2018) falling outside the range of $[5, 95)$. Subsequently, we also remove 16K duplicate sentence pairs, both within and across all document pairs, resulting in a dataset of 358,885 unique sentence pairs. This dataset is randomly split at the document level, following a sentence pair ratio of 0.95 / 0.025 / 0.025, thus yielding a total of 340,897 sentence pairs for training, 8,982 for validation, and 9,006 for test.

To assess the dataset's quality, we conduct a manual examination within our validation and test sets. This evaluation task is carried out by two third-year medical undergraduates,[8] who are responsible for determining if each sentence pair is misaligned (i.e. completely different sentence meaning or partly preserving the sentence meaning). Each examiner independently assesses a total of 8,982 + 9,006 = 17,988 sentence pairs within an average of 90 hours. Then, we perform a cross-checking process and find that 43 validation sentence pairs (0.48%) and 46 test sentence pairs (0.51%) exhibits misalignment. Given the tiny percentage of

[3] https://www.msdmanuals.com/professional

[4] https://www.downthemall.org/

[5] https://pypi.org/project/beautifulsoup4/

[6] We use the "pdftotext" Python library to extract content from PDF files, typically formatted in two columns.

[7] https://github.com/VinAIResearch/VinAI_Translate

[8] Examiners have a proficient English level at IELTS 7.0+ and GPA 3.5+/4.0.

misalignment at the sentence level in both the validation and test sets, we assert that our training set maintains a high standard of quality. Finally, we remove those misaligned pairs, resulting in a final count of 8,939 high-quality sentence pairs for validation and 8,960 for test. Table 1 shows the statistics of our MedEV dataset.

3. Experiment Setup

3.1. Experimental models

Our experimental setup focuses on using the MedEV dataset to explore: (i) the dataset's quality as demonstrated by its usage in improving neural machine translation (NMT) models' performance in the medical health domain; and (ii) a comparison of medical text translation performance among a well-known translation engine - Google Translate, a large language model - ChatGPT (gpt-3.5-turbo), pre-trained multilingual translation models `SeamlessM4T` (Communication et al., 2023) and `M2M100` (Fan et al., 2021), state-of-the-art Vietnamese-English NMT models `vinai-translate` (Nguyen et al., 2022b) and `envit5-translation` (Ngo et al., 2022), and pre-trained sequence-to-sequence models mBART (Liu et al., 2020) and `envit5-base` (Ngo et al., 2022).

mBART is pre-trained on a dataset of 25 languages, that contains 300GB of English texts and 137 GB of Vietnamese texts. Subsequently, `vinai-translate` is fine-tuned using mBART on a dataset of 9M sentence pairs, including 3M high-quality pairs in PhoMT (Doan et al., 2021) and an additional 6 million pairs from the noisier datasets CCAligned (El-Kishky et al., 2020) and WikiMatrix (Schwenk et al., 2021). On the other hand, `envit5-base` is a bilingual variant of the T5 model (Raffel et al., 2020), pre-trained on a dataset consisting of 80GB of English texts and 80GB of Vietnamese texts. Furthermore, `envit5-translation` is fine-tuned using `envit5-base` on a dataset of 6.2M high-quality sentence pairs from both PhoMT and the MTet dataset (Ngo et al., 2022).

3.2. Implementation details

On our MedEV dataset, we fine-tune the models `vinai-translate`, `envit5-translation`, mBART, and `envit5-base` for 5 epochs with AdamW (Loshchilov and Hutter, 2019), using HuggingFace "transformers" library (Wolf et al., 2020). We use an initial learning rate of 5e-5 and a maximum sequence length of 256. We employ mixed precision training (fp16), using 4 NVIDIA A100 GPUs, a batch size of 4 for each GPU, with 8 steps of gradient accumulation and 1250 warm-up steps.

	Model	Validation set		Test set	
		En2Vi	**Vi2En**	**En2Vi**	**Vi2En**
w/o FT	Google Translate	47.37	38.50	47.86	39.26
	ChatGPT 0-shot	34.38	29.79	34.45	30.39
	ChatGPT 1-shot	35.28	31.27	35.23	31.70
	ChatGPT 8-shot	36.09	31.87	36.02	32.57
	ChatGPT 16-shot	36.32	32.14	35.69	32.90
	ChatGPT 32-shot	34.92	32.08	36.37	32.94
	SeamlessM4T medium	31.04	21.57	31.25	21.65
	M2M100 418M	28.30	22.46	28.26	22.56
	vinai-translate	44.24	33.28	44.60	33.44
	envit5-translation	42.86	31.33	43.23	32.00
FT	vinai-translate	**52.21**	**42.66**	**52.14**	**42.38**
	envit5-translation	51.14	41.47	51.27	41.17
	mBART	51.23	41.67	51.18	41.51
	envit5-base	50.10	40.66	49.94	40.36

Table 2: BLEU scores. "FT" denotes fine-tuning.

We use beam search with a beam size of 5 for decoding. The performance is computed using metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005). Here, we calculate the case-sensitive BLEU score using SacreBLEU (Post, 2018). Each model is evaluated after every 1000 training steps, and the model checkpoint that yields the highest BLEU score on the validation set is selected for evaluation on the test set.

For ChatGPT (gpt-3.5-turbo), we conduct zero-shot/few-shot "in-context" learning. In the n-shot setting, we randomly select n samples from the training set for the prompt content for each validation/test sample. Note that for $n = 32$, since the "gpt-3.5-turbo" model limits requests to 4096 tokens, we restrict randomly sampled training sentences with a length of fewer than 64 tokens. Please refer to the prompt construction template in the Appendix A. In a preliminary experiment, we find that a temperature value of 0.2 yields the best performance score. Therefore, we report all our ChatGPT results using a fixed temperature of 0.2.

4. Experimental Results

Tables 2 and 3 present the BLEU, TER and METEOR scores obtained by all experimental models for both translation directions: English-to-Vietnamese (En2Vi) and Vietnamese-to-English (Vi2En). In the "without fine-tuning" (w/o FT) setting, the automatic translation engine Google Translate consistently outperforms both `vinai-translate` and `envit5-translation`, achieving the best scores. In contrast, ChatGPT tends to produce lower scores in most cases while `SeamlessM4T` and `M2M100 418M` exhibit the poorest performance, significantly behind the superior results of Google Translate. This is likely due to Google Translate being trained on some paral-

	Model	English-to-Vietnamese						Vietnamese-to-English					
		< 10	[10, 20)	[20, 30)	[30, 40)	[40, 50)	≥ 50	< 10	[10, 20)	[20, 30)	[30, 40)	[40, 50)	≥ 50
		5.16%	24.96%	28.40%	19.59%	10.40%	11.48%	13.99%	39.85%	26.94%	11.27%	3.93%	4.02%
w/o FT	Google Translate	43.17	45.16	47.33	48.20	48.57	48.16	34.08	37.86	38.97	39.47	41.01	42.41
	ChatGPT 0-shot	30.46	32.15	33.39	34.67	35.55	35.61	26.15	27.53	29.73	31.63	34.07	35.54
	ChatGPT 1-shot	31.67	33.22	34.16	35.49	36.17	35.87	27.32	28.78	31.20	33.05	34.99	36.75
	ChatGPT 8-shot	34.08	34.06	35.19	36.23	36.69	36.28	27.97	29.72	32.15	33.38	35.85	37.75
	ChatGPT 16-shot	29.95	32.91	34.93	36.02	36.60	36.34	28.10	29.97	32.42	33.89	35.93	38.34
	ChatGPT 32-shot	34.94	34.82	35.42	36.67	37.18	36.18	28.39	30.04	32.49	33.87	36.27	38.11
	SeamlessM4T medium	25.78	29.30	30.58	32.20	32.60	29.81	16.06	19.82	22.35	22.93	24.54	19.74
	M2M100 418M	24.07	27.08	28.07	29.04	29.66	27.14	19.40	20.55	22.56	23.98	24.59	24.17
	vinai-translate	31.53	43.07	44.51	44.77	43.70	43.92	28.81	30.99	33.03	34.36	36.01	38.01
	envit5-translation	38.72	41.77	42.75	43.73	44.08	42.59	27.07	28.31	31.76	33.89	35.53	37.12
FT	vinai-translate	48.64	**50.58**	**50.93**	**51.59**	**51.63**	**52.92**	**38.07**	**39.97**	**41.24**	**41.80**	**44.59**	47.12
	envit5-translation	**49.97**	50.50	50.30	50.81	51.27	51.99	35.32	38.07	40.11	41.32	44.44	**47.28**
	mBART	48.85	49.83	50.18	50.43	51.00	51.61	37.88	38.91	40.44	40.22	43.89	46.01
	envit5-base	49.11	49.13	48.95	48.88	49.12	49.98	35.43	37.62	39.05	38.89	42.02	44.14

Table 4: BLEU scores on the test set w.r.t. sentence lengths of reference sentences (i.e. the number of words including punctuations). The number below each length bucket indicates the percentage of sentences in that bucket.

	Model	English-to-Vietnamese				Vietnamese-to-English			
		Article Abstracts	MSD Manuals	Thesis Summaries	Article Translations	Article Abstracts	MSD Manuals	Thesis Summaries	Article Translations
w/o FT	Google Translate	40.06	56.86	49.20	52.79	32.17	48.38	**44.82**	40.92
	ChatGPT 0-shot	30.48	39.05	34.18	39.59	25.79	36.14	31.94	35.14
	ChatGPT 1-shot	31.42	39.66	35.02	39.78	26.69	38.14	33.25	35.75
	ChatGPT 8-shot	32.40	40.23	36.08	40.13	27.26	39.46	34.08	36.23
	ChatGPT 16-shot	31.91	39.96	36.28	40.38	27.50	40.08	33.85	36.13
	ChatGPT 32-shot	32.97	40.30	36.77	40.51	27.37	40.23	34.41	36.52
	SeamlessM4T medium	25.56	38.02	28.01	40.32	17.94	26.09	21.63	28.92
	M2M100 418M	23.13	34.36	24.35	37.69	19.36	26.20	23.77	28.30
	vinai-translate	37.99	53.46	37.03	48.74	28.07	39.79	35.82	39.34
	envit5-translation	37.44	50.85	41.04	46.89	24.51	42.86	33.65	38.13
FT	vinai-translate	**45.69**	**60.77**	**50.74**	**50.92**	**33.25**	**54.54**	42.22	**41.86**
	envit5-translation	44.73	60.29	50.02	50.09	32.32	54.26	40.54	37.92
	mBART	45.54	59.18	50.21	45.54	33.13	52.83	41.86	36.54
	envit5-base	43.58	58.13	48.16	44.00	32.08	51.09	39.11	36.07

Table 5: BLEU scores on the test set for each genre.

	Model	En2Vi		Vi2En	
		TER↓	METEOR↑	TER↓	METEOR↑
w/o FT	Google Translate	46.30	0.704	56.52	0.665
	ChatGPT 0-shot	59.35	0.625	66.68	0.608
	ChatGPT 1-shot	58.47	0.629	64.88	0.614
	ChatGPT 8-shot	57.80	0.634	63.74	0.621
	ChatGPT 16-shot	58.57	0.629	63.46	0.622
	ChatGPT 32-shot	57.48	0.638	63.32	0.623
	SeamlessM4T medium	61.69	0.576	76.13	0.498
	M2M100 418M	64.79	0.537	75.16	0.518
	vinai-translate	48.69	0.685	61.93	0.626
	envit5-translation	49.98	0.673	67.63	0.627
FT	vinai-translate	**42.22**	**0.740**	**52.24**	**0.685**
	envit5-translation	42.23	0.733	53.50	0.678
	mBART	42.99	0.732	53.03	0.678
	envit5-base	43.43	0.720	54.07	0.666

Table 3: TER and METEOR scores on the test set.

lel resource in the medical domain. As for ChatGPT, it generally attains better scores when more training pairs are used in the few-shot setups. When it comes to the "fine-tuning" setting, all fine-tuned models outperform Google Translate on both validation and test sets in both translation directions. Here, `vinai-translate` achieves the best scores, surpassing Google Translate by a substantial margin. Specifically, it outperforms Google Translate by 4+ BLEU points in English-to-Vietnamese translation and by 3+ BLEU points in Vietnamese-to-English translation.

Tables 4 and 5 show BLEU scores on the test set for English-to-Vietnamese and Vietnamese-to-English translation directions regarding each sentence length bucket and resource genre, respectively. We find from Table 4 that in medical texts, as the sentence length increases, the probability of encountering common words that match between the machine-translated text and the reference text also increases, resulting in higher BLEU scores. For shorter sentences, the translation system may offer synonymous words or medical terms that do not align perfectly with the reference text. As shown in Table 5, the highest BLEU scores are reported for MSD Manuals, which are composed of documents written by doctors on common diseases classified under the ICD-10 code system. The following are

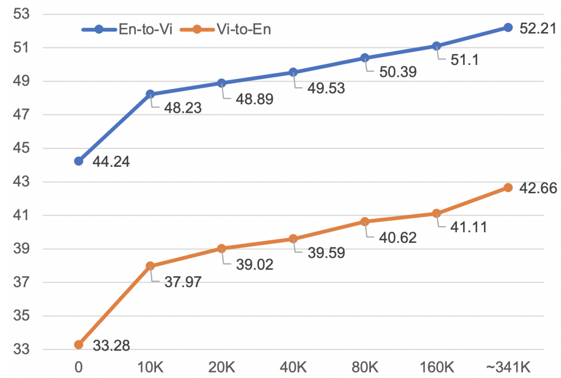

Figure 1: BLEU scores on the validation set when fine-tuning `vinai-translate` with different training sizes for both translation directions.

the scores reported for Thesis Summaries and Article Translations. On the contrary, the remaining resource genre, Article Abstracts (including article titles, abstracts, and keywords), contains more medical terminology than the other categories, resulting in lower BLEU scores.

Figure 1 presents BLEU scores on the validation set for both translation directions when fine-tuning `vinai-translate` with different numbers of training sentence pairs. Here, using only 10K sentence pairs helps substantially improve the baseline scores by 4+ points: from 44.24 to 48.23 for English-to-Vietnamese and from 33.28 to 37.97 for Vietnamese-to-English. Additional 330K+ pairs produce 4+ more points, increasing from 48.23 to 52.21 and from 37.97 to 42.66. These scores clearly demonstrate the positive impacts of larger training sizes.

5. Conclusion

In this paper, we have presented a high-quality MedEV dataset of about 360K parallel sentence pairs from 27K documents in the medical domain. We conduct experiments on MedEV to compare strong baselines and demonstrate the effectiveness of the NMT model `vinai-translate` in Vietnamese-English medical machine translation. We hope that the public release of our dataset will be a major step in the direction of more extensive Vietnamese-English machine translation in the medical field. In future work, we will explore the translation quality when combining our MedEV with other general domains PhoMT and MTet.

6. Ethical Statement

Data are collected from publicly available websites, such as journals and universities, but also from `www.msd.com`. The content extracted from these sources cannot be used for public or commercial purposes. Therefore, the content also contains no private data about the patients.

7. References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–14.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *arXiv preprint*, arXiv:2308.11596.

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned:

A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5960–5969.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Chi Nguyen Ngo, Trieu H. Trinh, Long Phan, Hieu Trung Tran, Tai Dang, Hieu Chi Nguyen, Minh Le Nguyen, and Minh-Thang Luong. 2022. MTet: Multi-domain Translation for English and Vietnamese. *arXiv preprint*, arXiv:2210.05610.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 2582–2587.

Linh The Nguyen, Nguyen Luong Tran, Long Doan, Manh Luong, and Dat Quoc Nguyen. 2022a. A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*, pages 1726–1730.

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022b. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell*, pages 5543–5544.

Vinh Van Nguyen, Ha Nguyen, Huong Thanh Le, Thai Phuong Nguyen, Tan Van Bui, Luan Nghia Pham, Anh Tuan Phan, Cong Hoang-Minh Nguyen, Viet Hong Tran, and Anh Huu Tran. 2022c. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5494–5502.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, pages 175–182.

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.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, Németh László, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference*

Recent Advances in Natural Language Processing.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A. Prompt Design

Zero-shot Setting:

- For English to Vietnamese translation:

```
I want you to act as a translator to
    translate text from English to
    Vietnamese in the medical domain.
Now with the following English INPUT
    text:
INPUT= [English input sentence from
    validation/test set]
What is the corresponding Vietnamese-
    translated OUTPUT text?
```

- For Vietnamese to English translation:

```
I want you to act as a translator to
    translate text from Vietnamese to
    English in the medical domain.
Now with the following Vietnamese INPUT
    text:
INPUT= [Vietnamese input sentence from
    validation/test set]
What is the corresponding English-
    translated OUTPUT text?
```

Few-shot Setting:

- For English to Vietnamese translation:

```
I want you to act as a translator to
    translate text from English to
    Vietnamese in the medical domain.
    For instance, consider the following
     English INPUT text:
INPUT= [shot 1 source]
[shot 2 source]
[shot n source]

You would generate a corresponding
    Vietnamese OUTPUT text as follows:
OUTPUT= [shot 1 reference]
[shot 2 reference]
[shot n reference]

Now with the following English INPUT
    text:

INPUT= [English input sentence from the
    validation/test set]

What is the corresponding Vietnamese-
    translated OUTPUT text?
```

- For Vietnamese to English translation:

```
I want you to act as a translator to
    translate text from Vietnamese to
    English in the medical domain. For
    instance, consider the following
    Vietnamese INPUT text:}
INPUT= [shot 1 source]
[shot 2 source]
[shot n source]

You would generate a corresponding
    English OUTPUT text as follows:
OUTPUT= [shot 1 reference]
[shot 2 reference]
[shot n reference]

Now with the following Vietnamese INPUT
    text:

INPUT= [Vietnamese input sentence from
    validation/test set]

What is the corresponding English-
    translated OUTPUT text?
```

The output from the ChatGPT API may sometimes include model-generated sentences in addition to the translation results. We manually check the output and remove these sentences. For instance:

- The model repeats sentences from the prompt: *“The corresponding English-translated text is:”, “The corresponding Vietnamese-translated OUTPUT text is:”*
- The model adds new sentences in the response content: *“Possible English translation:”, “Possible OUTPUT:”, “Possible translation:”*