Enriching Biomedical Knowledge for Low-resource Language Through Large-Scale Translation

Long Phan[1]*, **Tai Dang**[1,3]*, **Hieu Tran**[1]*, **Trieu H. Trinh**[1,4]*,
Vy Phan[3], **Lam D. Chau**[2] and **Minh-Thang Luong**[1]
[1]VietAI Research
[2]Case Western Reserve University
[3]University of Massachusetts-Amherst
[4]New York University

Abstract

Biomedical data and benchmarks are highly valuable yet very limited in low-resource languages other than English, such as Vietnamese. In this paper, we use a state-of-the-art translation model in English-Vietnamese to translate and produce both pretrained and supervised data in the biomedical domains. Thanks to such large-scale translation, we introduce ViPubmedT5, a pretrained Encoder-Decoder Transformer model trained on 20 million translated abstracts from the high-quality public PubMed corpus. ViPubMedT5 demonstrates state-of-the-art results on two different biomedical benchmarks in summarization and acronym disambiguation. Further, we release ViMedNLI - a new NLP task in Vietnamese translated from MedNLI using the recently public En-vi translation model and carefully refined by human experts, with evaluations of existing methods against ViPubmedT5.

1 Introduction

In recent years, pretrained language models (LMs) have played an important and novel role in developing many Natural Language Processing (NLP) systems. Utilizing large pretrained models like BERT (Devlin et al., 2018), XLNET (Yang et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020) BART (Lewis et al., 2019), and T5 (Raffel et al., 2019) has become an effective trend in natural language processing. All these large models follow the Transformer architecture proposed by (Vaswani et al., 2017) with the attention mechanism. The architecture has been proven to be very suitable for finetuning downstream tasks leveraging transfer learning with their large pretrained checkpoints. Before the emergence of large Transformer LMs, traditional wording embedding gave each word a fixed global representation. Large pretrained models can derive word vector representation from a trained large corpus. This will give the pretrained model a better knowledge of the generalized representation of a trained language/domain and significantly improve performance on downstream finetune tasks. The success of pretrained models on a generative domain (BERT, RoBERTa, BART, T5, etc.) has created a path in creating more specific-domain language models such as CodeBERT (Feng et al., 2020) and CoTexT (Phan et al., 2021b) for coding languages, TaBERT (Yin et al., 2020) for tabular data, BioBERT (Lee et al., 2019) and PubmedBERT (Tinn et al., 2021) for biomedical languages.

Biomedical literature is getting more popular and widely accessible to the scientific community through large databases such as Pubmed[1], PMC[2], and MIMIC-IV (Johnson et al., 2021). This also leads to many studies, corpora, or projects released to further advance the Biomedical Natural Language Processing field (Lee et al., 2019; Tinn et al., 2021; Phan et al., 2021a; Yuan et al., 2022). These biomedical domain models leverage transfer learning from pretrained models (Devlin et al., 2018; Clark et al., 2020; Raffel et al., 2019; Lewis et al., 2019) to achieve state-of-the-art results on multiple Biomedical NLP tasks like Named Entity Recognition (NER), Relation Extraction (RE), or document classification.

However, few studies have been on leveraging large pretrained models for biomedical NLP in low-resource languages. The main reason is the lack of large biomedical pretraining data and benchmark datasets. Furthermore, collecting biomedical data in low-resource languages can be very expensive due to scientific limitations and inaccessibility.

We attempt to overcome the issue of lacking biomedical text data in low-resource languages by using state-of-the-art translation works. We start

[*] The first four authors contributed equally to this work

[1] https://pubmed.ncbi.nlm.nih.gov

[2] https://www.ncbi.nlm.nih.gov/pmc

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3131–3142
May 2-6, 2023 ©2023 Association for Computational Linguistics

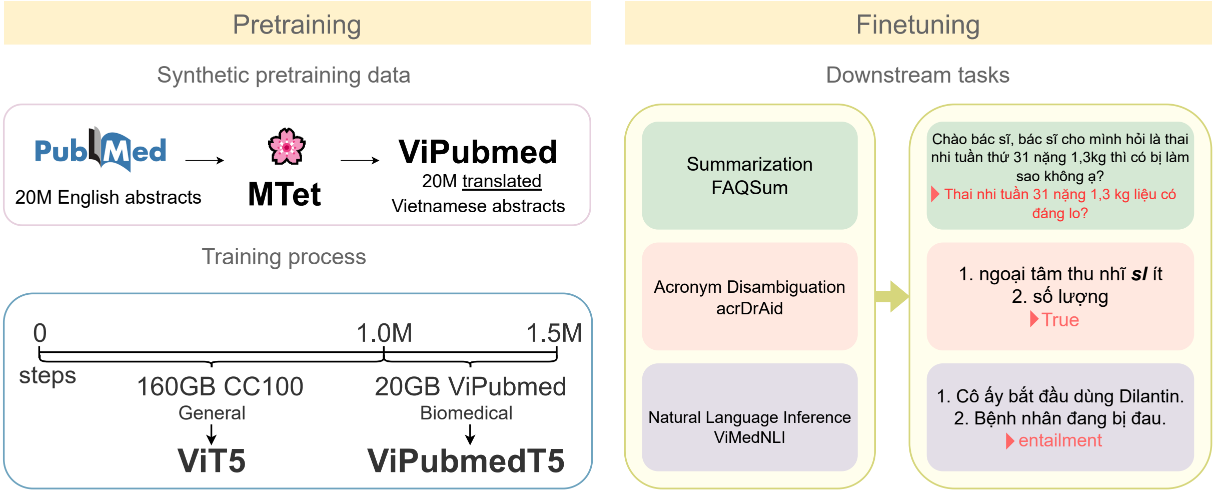
Pretraining
Finetuning
Synthetic pretraining data
PubMed
20M English abstracts
MTet
ViPubmed
20M translated Vietnamese abstracts
Training process
0
1.0M
1.5M
steps
160GB CC100
General
ViT5
20GB ViPubmed
Biomedical
ViPubmedT5
Downstream tasks
Summarization
FAQSum
Acronym Disambiguation
acrDrAid
Natural Language Inference
ViMedNLI
Chào bác sĩ, bác sĩ cho mình hỏi là thai nhi tuần thứ 31 nặng 1,3kg thì có bị làm sao không ạ?
Thai nhi tuần 31 nặng 1,3 kg liệu có đáng lo?
1. ngoại tâm thu nhĩ ***sl*** ít
2. số lượng
True
1. Cô ấy bắt đầu dùng Dilantin.
2. Bệnh nhân đang bị đau.
entailment

Figure 1: Overview of the pretraining and finetuning of ViPubmedT5

with the Vietnamese language and keep everything reproducible for other low-resource languages in future work.

We introduce ViPubmedT5, a pretrained encoder-decoder transformer model trained on synthetic Vietnamese biomedical text translated with state-of-the-art English-Vietnamese translation work. Meanwhile, we also introduced ViMedNLI, a medical natural language inference task (NLI), translated from the English MedNLI (Romanov and Shivade, 2018) with human refining.

We thoroughly benchmark the performance of our ViPubmedT5 model when pretrained with synthetic translated biomedical data with ViMedNLI and other public Vietnamese Biomedical NLP tasks (Minh et al., 2022). The results show that our model outperforms both general domain (Nguyen and Nguyen, 2020; Phan et al., 2022) and health-specific domain Vietnamese (Minh et al., 2022) pretrained models on biomedical tasks.

In this work, we offer the following contributions:

- A state-of-the-art English-Vietnamese Translation model (with self-training) on medical and general domains.
- A first Encoder-Decoder Transformer model ViPubmedT5 pretrained on large-scale synthetic translated biomedical data.
- A Vietnamese medical natural language inference dataset (ViMedNLI) that translated from MedNLI (Romanov and Shivade, 2018) and refined with biomedical expertise human.
- We publicize our model checkpoints, datasets, and source code for future studies on other low-resource languages.

2 Related Works

The development of parallel text corpora for translation and use for training MT systems has been a rapidly growing field of research. In recent years, low-resource languages have gained more attention from the industry, and academia (Chen et al., 2019b; Shen et al., 2021; Gu et al., 2018; Nasir and Mchechesi, 2022). Previous works include gathering more training data or training large multilingual models (Thu et al., 2016; Fan et al., 2021). Low-Language MT enhances billions of people's daily life in numerous fields. Nonetheless, there are specific domains crucial yet limited such as biomedical and healthcare, in which MT systems have not been able to contribute adequately.

Previous works using MT systems for biomedical tasks includes (Neves et al., 2016; Névéol et al., 2018). Additionally, several biomedical parallel (Deléger et al., 2009) have been utilized just for terminology translation only. Pioneer attempts to train MT systems using a corpus of MEDLINE titles (Wu et al., 2011), and the use of publication titles and abstracts for both ES-EN and FR-EN language pairs (Jimeno-Yepes et al., 2012). However, none of these works targets low-resource languages. A recent effort to train Vietnamese ML systems for biomedical and healthcare is Minh et al. (2022). These, however, do not utilize the capability of MT systems, instead relying on manual crawling. Therefore, this motivation has led us to employ MT systems to contribute high-quality Vietnamese datasets that emerged from the English language. To the best of our knowledge, this is the first work utilizing state-of-the-art machine translation to translate both self-supervised and supervised learning biomedical data for pretrained models in a low-resource language setting.

2.1 Pubmed and English Biomedical NLP Studies

The Pubmed[3] provides access to the MEDLINE database[4] which contains titles, abstracts, and metadata from medical literature since the 1970s. The dataset consists of more than 34 million biomedical abstracts from the literature collected from sources such as life science publications, medical journals, and published online e-books. This dataset is maintained and updated yearly to include more up-to-date biomedical documents.

Pubmed Abstract has been the main dataset for almost any state-of-the-art biomedical domain-specific pretrained models (Lee et al., 2019; Yuan et al., 2022; Tinn et al., 2021; Yasunaga et al., 2022; Alrowili and Shanker, 2021; Phan et al., 2021a). In addition, many well-known Biomedical NLP/NLU benchmark datasets are created based on the unlabeled Pubmed corpus (Doğan et al., 2014; Nye et al., 2018; Herrero-Zazo et al., 2013; Jin et al., 2019). Recently, to help accelerate research in biomedical NLP, Gu et al. (2020) releases BLURB (**B**iomedical **L**anguage **U**nderstanding & **R**easoning **B**enchmark), which consists of multiple pretrained biomedical NLP models and benchmark tasks. It is important to note that all of the top 10 models on the BLURB Leaderboard[5] are pretrained on the Pubmed Abstract dataset.

2.2 English-Vietnamese Translation

Due to its limitation of high-quality parallel data available, English-Vietnamese translation is classified as a low-resource translation language (Liu et al., 2020). One of the first notable parallel datasets and En-Vi neural machine translation is ISWLT'15 (Luong and Manning, 2015) with 133K sentence pairs. A few years later, PhoMT (Doan et al., 2021) and VLSP2020 (Ha et al., 2020) released larger parallel datasets, extracted from publicly available resources for the English-Vietnamese translation.

Recently, VietAI[6] curated the largest 4.2M high-quality training pairs from various domains and achieved state-of-the-art on English-Vietnamese translation (Ngo et al., 2022). The work also focuses on En-Vi translation performance across multiple domains, including biomedical. As a result, the project's NMT outperforms existing En-Vi translation models (Doan et al., 2021; Fan et al., 2020) by more than 2% in the BLEU score.

3 Improvements on Biomedical English-Vietnamese Translation through Self-training

To generate a large-scale synthetic translated Vietnamese biomedical corpus, we first look into improving the existing English-Vietnamese translation system in the biomedical translation domain. Previous work from Ngo et al. (2022) has shown that En-Vi biomedical bitexts are very rare, even for large-scale bitext mining. Therefore, we look into self-training to leverage the available monolingual English biomedical data.

Self-training approach has been experimented with in He et al. (2019) and utilized to improve translation on low-resource MT systems (Chen et al., 2019a). The advantage of this method is that the source side of the monolingual corpus can be domain-specific data for translation. However, the shortcoming is that the generated targets can be low-quality and affect the machine translation performance. Therefore, we start with the English-Vietnamese machine translation model from Ngo et al. (2022), denoted bT_A, which achieves state-of-the-art results on both En-Vi biomedical and general translation domains.

We use bT_A to translate and generate a synthetic parallel biomedical dataset with 1M pairs of English-Vietnamese biomedical abstracts from the Pubmed Corpus. The new 1M En-Vi biomedical pairs are then concatenated with the current high-quality En-Vi translation dataset from MTet (Ngo et al., 2022) and PhoMT (Doan et al., 2021), increasing from 6.2M to 7.2M En-Vi sentence pairs total. To verify the effectiveness of our new self-training data, we re-finetune the bT_A model on this 7.2M bitexts corpus. We report the model performance on the medical test set from MTet and the general test set from PhoMT in Table 1 (the translation performances on other domains like News, Religion, and Law are reported in Appendix A for

[3] https://pubmed.ncbi.nlm.nih.gov

[4] https://www.nlm.nih.gov/bsd/pmresources.html

[5] https://microsoft.github.io/BLURB/leaderboard.html

[6] https://vietai.org

Table 1: BLEU Scores Results for En-Vi Translation on MTet Medical and PhoMT General Test Sets

Model	**Finetune Datasets**	**MTet Medical Test Set**	**PhoMT General Test Set**
M2M100	CCMatrix + CCAligned	30.18	35.83
Google Translate	-	38.60	39.86
SOTA	MTet+PhoMT	38.69	45.47
Ours	MTet+PhoMT +1M Self-training Pubmed Abstracts	**45.61**	**46.01**

Notes: The best BLEU scores are in bold. The state-of-the-art (SOTA) model and MTet dataset are from Ngo et al. (2022); PhoMT dataset and Google Translate's result are from Doan et al. (2021). M2M100 model is from Fan et al. (2020).

further reference).

The results show that our model outperforms existing Machine Translation systems in English-Vietnamese translation by applying self-training. We obtain a significant gain of 6.61 BLEU Score (38.69->45.61) on the MTet Medical test set. Our model with self-training also achieves state-of-the-art results on the PhoMT general domain test set by 0.53 BLEU Score (45.47->46.01). This shows that our approach not only improves the English-Vietnamese translation performance in the biomedical context but also generalizes to general translation. We further discuss our self-training model performance on other translation domains in Appendix A.

4 ViPubmed

After developing a new state-of-the-art machine translation system for English-Vietnamese translation in the biomedical domain in Section 3, we apply the system, denoted bT_B, on downstream translation to generate the first large-scale synthetic translated biomedical corpus for Vietnamese.

To ensure that our translated ViPubmed dataset contains up-to-date biomedical research (for example, Covid-19 diseases and Covid-19 vaccines), we use the newest Pubmed22[7] which contains approximately 34 million English biomedical abstracts published. The raw data is compressed in XML format. We then parse these structured XMLs to obtain the abstract text with Pubmed Parser[8] (Achakulvisut et al., 2020).

The machine translation model bT_B is an Encoder-Decoder Transformer based model with 512 token-length for input and output. Therefore, we filter out English Pubmed abstracts with more than 512 tokens. For fair size comparison with the unlabeled dataset of other health-related Vietnamese pretrained models (discussed in Section 7.2), we take a subset of 20M biomedical abstracts (20GB of text) for translation and leave a larger subset for future releases. We then translate the 20M English biomedical abstracts with the bT_B model using 4 TPUv2-8 and 4 TPUv3-8.

5 ViMedNLI

Along with an unlabeled dataset for pretraining, we also introduce a benchmark dataset generated by translation and refined with human experts. We start with a natural language inference (NLI) task as it is less prone to errors in biomedical entity translation compared to named-entity recognition (NER) or relation extraction (RE) tasks. The process of creating the ViMedNLI is shown in Figure 2.

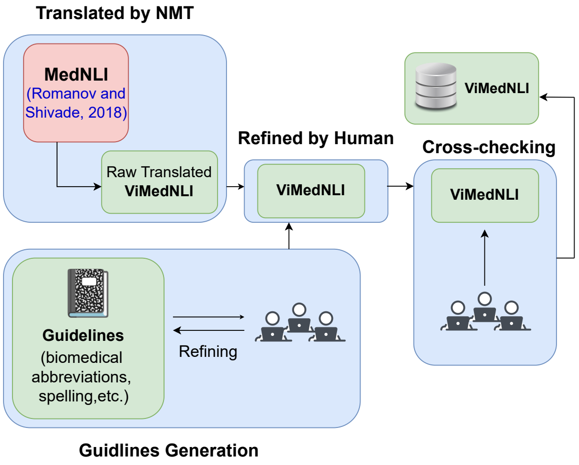

Figure 2: The Process of ViMedNLI Corpus Creation

[7] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline
[8] https://github.com/titipata/pubmed_parser

Table 2: Some Examples of Abbreviations and Spelling Refining

#	MedNLI	Translated by NMT	Refined by human
Abbreviations Refining			
1	Electrocardiograms revealed no QRS changes	Điện tâm đồ cho thấy không có thay đổi về QRS .	Điện tâm đồ cho thấy không có thay đổi về phức độ QRS
2	Patient has no PMH	Bệnh nhân không có PMH	bệnh nhân không có tiền sử bệnh
3	Patient is post op	Bệnh nhân đã hồi phục *(Patient is recovered)*	Bệnh nhân hậu phẫu thuật
Spelling Refining			
4	The infant was born at herm	Đứa bé được sinh ra ở Herm *(The baby was born at Herm)*	Đứa bé được sinh đủ tháng (The baby was born at term)
5	The patient had an sotesophytes	Bệnh nhân có sinh cảm *(The patient has flu)*	Bệnh nhân bị viêm xương khớp *(The patient had osteophytes)*
6	Patient has delerium	Bệnh nhân có hội chứng delerium *(Patient has delerium syndrome)*	Bệnh nhân bị mê sảng *(Patient has delirium)*

#1: Abbreviation in English can be used in both English and Vietnamese.
#2: Abbreviation can only be used in English. In Vietnamese, abbreviation is different.
#3: Abbreviation in English is wrong when translated to Vietnamese.
#4: The word *"term"* is misspelled as *"herm"*.
#5: The word *"osteophytes"* is misspelled as *"sotesophytes"*.
#6: The word *"delirium"* is misspelled as *"delerium"*.

5.1 MedNLI

MedNLI (Romanov and Shivade, 2018) is an NLI dataset annotated by doctors and grounded in the patients' medical history. Given a premise sentence and a hypothesis sentence, the relation of the two sentences (entailment, contradiction, neutral) is labeled by two board-certified radiologists. The source of premise sentences in MedNLI is from MIMIC-III (Johnson et al., 2016), a large open-source clinical database. The dataset has been widely studied and benchmarked by the Biomedical NLP research community[9] (Peng et al., 2019; Phan et al., 2021a; El Boukkouri et al., 2020; Alrowili and Shanker, 2021; Kanakarajan et al., 2019).

5.2 Dataset Challenges

We follow the same procedures discussed in Section 4 to translate the same training, development, and test sets released in Romanov and Shivade (2018). The time and resources to translate the dataset are negligible as there are a total of 14522 samples.

However, upon translating the dataset with NMT, we find out that the English clinical note domain has a distinct sublanguage with unique challenges (abbreviations, inconsistent punctuation, misspellings, etc.). This observation has also been addressed in Friedman et al. (2002) and Meystre et al. (2008). Such differences in clinical language representation challenge the translation output and our quest to release a high-quality medical dataset.

[9] https://paperswithcode.com/dataset/mednli

Table 3: Statistics of finetuned datasets

Corpus	Train	Dev	Test	Task	Domain
acrDrAid, pairs	4000	523	1130	Acronym disambiguation	Medical
FAQSum, documents	10621	1326	1330	Abstractive summarization	Healthcare
ViMedNLI, pairs	11232	1395	1422	Inference	Clinical

5.3 Human Refining

The unique challenges of clinical data under translation settings (discussed in Section 5.2) require us to work with humans who not only have expertise in biomedical knowledge but are also sufficient in both English and Vietnamese languages to refine the dataset. Therefore, we collaborate with pre-medical Vietnamese students who studied at well-known U.S. Universities to refine the ViMedNLI datasets.

The refining process starts with a comprehensive guidelines document with thorough annotation instructions and examples. Then, as clinical notes contain a significant amount of technical abbreviations that the machine translation system can not translate initially (Section 5.2), we work with the medical annotators to create abbreviations and their expansion forms. To make sure the expansion form of these abbreviations generalizes well in real-world settings, we verify the use case of these words through multiple Vietnamese medical websites, blogs, and online dictionaries. Hence, we decided to keep the original English abbreviations, replace them with a Vietnamese expansion form, or replace them with a Vietnamese abbreviation. Some examples of this process are shown in Table 2.

Aside from the English medical abbreviations, there are several grammatical and spelling mistakes the machine translation system does not understand, translating either into Vietnamese meanings or even failing to translate. Human refining is therefore required. The phrase *"The infant was born at herm"*, for example, was translated as "Đứa bé được sinh ra ở Herm". The word *"herm"*, which should be spelled as *"term"*, is misspelled and has no medical meaning. The accurate translation should be "Đứa bé được sinh đủ tháng" (*"The infant was born at term"*). Table 2 shows more examples of spelling refining cases.

Additionally, the machine translation system occasionally produces incorrect Vietnamese meanings when translating words with proper English spelling and grammar. Considering the sentence *"The patient had post-term delivery"* as an example. Despite having the meaning "Bệnh nhân sinh muộn", it was mistranslated as "Bệnh nhân sinh non" (*"The patient had pre-term delivery"*). Another example is "Narrowing of the vessels", which means "Thu hẹp các mạch" rather than "Thu hẹp các" (no meaning).

6 ViPubmedT5

With an unlabeled synthetic translated ViPubmed Corpus (Section 4) and a benchmark ViMedNLI dataset (Section 5), we pretrain and finetune a Transformer-based language model (Vaswani et al., 2017) to verify the effectiveness of our approach in enriching Vietnamese biomedical domain with translation data. We explain our model and the pretraining settings we applied in this section.

6.1 Model Architecture

We adopt the Transformer encoder-decoder model proposed by Vaswani et al. (2017), the ViT5 (Phan et al., 2022) checkpoints, and T5 framework [10] implemented by Raffel et al. (2019). ViT5 is the first monolingual Vietnamese Transformer model; the model achieves state-of-the-art results on multiple Vietnamese general tasks, including generation and classification. The ViT5 publication releases 2 model sizes - base and large. We train ViPubmedT5 using the base setting (220 million parameters) and leave larger models for future work.

6.2 Pretraining

We pretrain our ViPubmedT5 on 20GB of translated biomedical data ViPubmed (Section 4). We leverage the Vietnamese checkpoints in the original ViT5 work (Phan et al., 2022) and continuously pretrain the model on the synthetic biomedical-specific data for another 500k steps. Previous works (Lee et al., 2019; Tinn et al., 2021) have shown that this approach will allow pretrained language models to learn a better representation of

[10] https://github.com/google-research/text-to-text-transfer-transformer

Table 4: Tests results on Vietnamese health and biomedical tasks

Domain	Datasets	Metrics	**PhoBERT**	**ViT5**	**ViHealthBERT**	**ViPubmedT5**
			(+news)	(+cc100)	(+health text mining)	(+translated ViPubmed)
Healthcare	FAQSum	RougeL	41.16	**61.3**	43.85	60.6
Medical	acrDrAid	Mac-F1	82.51	88	86.7	**89.04**
Clinical	ViMedNLI	Acc	77.29	77.85	79.04	**81.65**

Notes: The best scores are in bold, and the second best scores are underlined. PhoBERT & ViHealthBERT scores on FAQSum and acrDrAid are from Minh et al. (2022)

biomedical language context while maintaining the core Vietnamese language representation.

We train ViPubmedT5 using the same spans-masking learning objective as Raffel et al. (2019). During self-supervised training, spans of biomedical text sequences are randomly masked (with sentinel tokens). The target sequence is formed as the concatenation of the same sentinel tokens and the real masked spans/tokens.

7 Experiments

7.1 Benchmark dataset

We finetune and benchmark our pretrained ViPubmedT5 model on two public Vietnamese biomedical-domain datasets acrDrAid & FAQSum, (Minh et al., 2022) and our released ViMedNLI (Section 5). Detailed statistics of the three datasets are shown in Table 3.

- **acrDrAid** (Minh et al., 2022) is a Vietnamese Acronym Disambiguation (AD) dataset that contains radiology reports from Vinmec hospital[11], Vietnam. The task is correctly identifying the expansion of an acronym in a given radiology report context. The dataset is annotated by three expert radiologists. The acrDrAid has 135 acronyms and 424 expansion texts in total.

- **FAQ Summarization** (Minh et al., 2022) is a Vietnamese summarization dataset collected from FAQ sections of multiple *healthcare* trustworthy sites. For each FAQ section, the question text is the input sequence, and the title is a target summary.

- **ViMedNLI** is our released dataset discussed in Section 5.

7.2 Baseline

To verify the effectiveness of our proposed methods, we compare our ViPubmedT5 model with other state-of-the-art Vietnamese pretrained models:

- **PhoBERT** (Nguyen and Nguyen, 2020) is the first public large-scale monolingual language model pretrained for the Vietnamese language. The model follows the original RoBERTa (Liu et al., 2019) architecture. PhoBERT is trained on a 20GB word-level Vietnamese news corpus.

- **ViT5** (Phan et al., 2022) is the most recent state-of-the-art Vietnamese pretrained model for both generation and classification tasks. The model is trained on a general domain CC100-vi corpus.

- **ViHealthBERT** (Minh et al., 2022) is the first domain-specific pretrained language model for Vietnamese healthcare. After initializing weights from PhoBERT, the model is trained on 25M health sentences mined from different sources.

8 Results

The main finetuned results are shown in Table 4. The main takeaway is that training on synthetic translated biomedical data allows ViPubmedT5 to learn a better biomedical context representation. As a result, ViPubmedT5 achieves state-of-the-art in Medical and Clinical contexts while performing slightly worse than ViT5 in healthcare topics.

On the healthcare domain (FAQSum), ViPubmedT5 approximates the current state-of-the-art result (60.6 and 61.3) while outperforming the other models by a large margin (43.85). The slight difference in performance to ViT5 signifies a difference in data distribution in PubMed abstracts (scientific

[11] https://vinmec.com/

writing) and FAQSum (dialogues between patients and doctors).

For both medical and clinical datasets, ViPubmedT5 outperforms other existing models. There are also notable improvements from the general domain ViT5 to ViPubmedT5 (88->89.04 in acrDrAid and 77.85->81.65 in ViMedNLI). This indicates that the translated ViPubmed corpus contains biomedical knowledge that low-resource Vietnamese pretrained models can leverage.

Meanwhile, our newly translated ViMedNLI can serve as a robust baseline dataset for Vietnamese BioNLP research. Both health and biomedical domain models (ViHealthBERT & ViPubmedT5) perform better than general domain models (PhoBERT & ViT5) on the ViMedNLI dataset. This shows that our translated and refined ViMedNLI dataset is high-quality and has robust biomedical contexts.

9 Scaling to Other Languages

Our novel way of utilizing a state-of-the-art NMT system to generate synthetic translated medical data for pretrained models is not limited to the Vietnamese language and is scalable to many other low-resource languages. Recent works focus on improving the quality of multiple low-resource NMT systems (NLLB Team et al., 2022; Fan et al., 2020; Bañón et al., 2020). These new state-of-the-art NMTs make the approach discussed in this paper more practical to produce synthetic translated biomedical data, enriching the Biomedical NLP research knowledge in multiple low-resource languages.

10 Conclusion

We utilize the state-of-the-art translation model MTet to scale up the very low-resourced yet highly valuable biomedical data in Vietnamese. Namely, ViPubMedT5, a T5-style Encoder-Decoder Transformer pretrained on a large-scale translated corpus of the biomedical domain that demonstrated state-of-the-art results on both inference and acronym disambiguation in the biomedical domain. We also introduced ViMedNLI, a machine-translated and human-expert refined benchmark in natural language inference to further grow the Vietnamese suite of benchmarks and data in biomedical data.

11 Limitations

Although our pretrained model trained on synthetic translated biomedical data produces state-of-the-art results on downstream tasks for the Vietnamese language, the approach hugely depends on the quality of the NMTs for other low-resource languages. Thanks to recent studies and contributions from the Vietnamese research community (Section 2.2), the English-Vietnamese translation system has proven strong enough for us to conduct the experiments discussed in this work. However, the NMT's actual performance needed before making the translated biomedical data useful for pretrained models is still a question that requires further studies.

12 Acknowledgements

We would like to thank the Google TPU Research Cloud (TRC) program and Soonson Kwon (Google ML Ecosystem programs Lead) for their supports.

References

Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. 2020. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979.

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Peng-Jen Chen, Jiajun Shen, Matt Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc'Aurelio Ranzato. 2019a. Facebook ai's WAT19 myanmar-english translation task submission. *CoRR*, abs/1910.06848.

Peng-Jen Chen, Jiajun Shen, Matthew Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc'Aurelio Ranzato. 2019b. Facebook AI's WAT19 Myanmar-English translation task submission. In *Proceedings of the 6th Workshop on Asian Translation*, pages 112–122, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of biomedical informatics*, 42 4:692–701.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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. *CoRR*, abs/2110.12199.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, 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.*, 22:107:1–107:48.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. Codebert: A pre-trained model for programming and natural languages. *CoRR*, abs/2002.08155.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of Biomedical Informatics*, 35(4):222–235. Sublanguage - Zellig Harris Memorial.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Thanh-Le Ha, Van-Khanh Tran, and Kim-Anh Nguyen. 2020. Goals, challenges and findings of the VLSP 2020 English-Vietnamese news translation shared task. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 99–105, Hanoi, Vietnam. Association for Computational Lingustics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *CoRR*, abs/1909.13788.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Antonio Jimeno-Yepes, Élise Prieur, and Aurélie Névéol. 2012. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14:146 – 146.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *CoRR*, abs/1909.06146.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. Mimic-iv.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Kamal raj Kanakarajan, Suriyadeepan Ramamoorthy, Vaidheeswaran Archana, Soham Chatterjee, and Malaikannan Sankarasubbu. 2019. Saama research at MEDIQA 2019: Pre-trained BioBERT with attention visualisation for medical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 510–516, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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. *CoRR*, abs/2001.08210.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–44.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. Vihealthbert: Pre-trained language models for vietnamese in health text mining. In *Proceedings of the Language Resources and Evaluation Conference*, pages 328–337, Marseille, France. European Language Resources Association.

Muhammad Umair Nasir and Innocent Mchechesi. 2022. Geographical distance is the new hyperparameter: A case study of finding the optimal pre-trained language for English-isiZulu machine translation. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 1–8, Seattle, USA. Association for Computational Linguistics.

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang

Luong. 2022. Mtet: Multi-domain translation for english and vietnamese.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *CoRR*, abs/2003.00744.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Benjamin E. Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. *CoRR*, abs/1806.04185.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021a. Scifive: a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598.

Long N. Phan, Hieu Tran, Daniel Le, Hieu Nguyen, James T. Anibal, Alec Peltekian, and Yanfang Ye. 2021b. Cotext: Multi-task learning with code-text transformer. *CoRR*, abs/2105.08645.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Jiajun Shen, Peng-Jen Chen, Matt Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. *ArXiv*, abs/1909.13151.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *CoRR*, abs/2112.07869.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Cuijun Wu, F. Xia, Louise Deléger, and Imre Solti. 2011. Statistical machine translation for biomedical text: are we there yet? *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2011:1290–9.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *CoRR*, abs/2005.08314.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

A Results for En-Vi Translation with Self-training

We explore a more drastic measure to expand the amount of data in the biomedical domain by self-training. We also explore how a large expansion in the number of biomedical bitexts affects the performance of our model on other domains such as Law, Religion, and News by using the MTet multi-domain test set (Ngo et al., 2022).

Table 5: BLEU Scores Results for En-Vi Translation on MTet Multi-domain Test Set

Model	**MTet Multi-domain Test Set**			
	Medical	News	Religion	Laws
SOTA	38.69	**51.47**	**41.44**	36.43
Ours	**45.61**	51.003	40.68	**39.51**

Notes: The best BLEU scores are in bold. The state-of-the-art (SOTA) model and MTet dataset are from Ngo et al. (2022); Our model trained with self-training approach on an extra 1M En-Vi synthetic biomedical abstracts is discussed in Section 3

The improvement is not evident across all domains when tested on a diverse domain test set (MTet). For example, while there are notable improvements in the Medical and Law domain, the model performs worse in the Religion and News domains. This can be attributed to the context representation of biomedical Pubmed Abstract data. Scientific abstracts tend to be more formal and academic for knowledgeable audiences with more domain expertise. Therefore, training on such data allows the Machine Translation system to perform better not only on the trained domain (Medical) but also on other formally presented domains, such as Law, while at the same time performing slightly worse on other domains (News and Religion).