UNIVERSITY OF SCIENCE
VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

FINAL PROJECT

Vietnamese-To-English Machine Translation for Ancient Texts with T5

CS418 – Introduction to Natural Language Processing

Team Member

22125056 – Le Phat Minh

22125072 – Vo Thinh Phat

22125085 – O Hon Sam

22127384 – Duong Quang Thang

Theoretical Teacher : Dinh Dien

Laboratory Teacher : Luong An Vinh

Nguyen Hong Buu Long

Ho Chi Minh City, 02/01/2025

MEMBERS

Student ID	Full name	Contribution percentage
22125056	Le Phat Minh	30%
22125072	Vo Thinh Phat	30%
22125085	O Hon Sam	30%
22127384	Duong Quang Thang*	10%

*: With the Prof. Dinh Dien's permission, Duong Quang Thang was allowed to participate in the project and attend the CS418 course as an auditor (not officially registered and not receiving grades for this course).

Contents

1 Project Overview

In this project, we built a Vietnamese-to-English machine translation system for Ancient Texts by fine-tuning the Text-to-Text Transfer Transformer (T5) model [1].

To perform the project, we follow main steps:

– **Data Collection:** We gathered data by crawling from websites and digitized ancient texts from books.

– **Data Augmentation:** To improve translation quality, we employed data augmentation techniques; this included using large language models for generating paraphrases and translators expand the dataset with diverse examples.

– **Model Fine-Tuning:** We fine-tuned on multiple T5 variants using the same hyperparameters. Among them, the fine-tuned `VietAI/envit5-translation` model [2] achieved the highest BLEU, making it the chosen model for the web system implementation.

– **Web User Interface:** We implemented a simple web interface using the Flask library [16], allowing users to input text and receive its translation.

2 Data Collection and Automatic Alignment

2.1 Data Collection

The corpus that we have built up during this project comprises notable Vietnamese literary, with the majority written in Nom characters. To ensure comprehensiveness and authenticity, we sourced these works from reputable archives, libraries, and specialized compilations from famous translators all over the world. Each text has been carefully verified to ensure the high quality of the bilingual training corpus for the machine translation task.

Despite the limited free resource of ancient texts on the Internet, our group has collected more than 15,000 pair of ancient Vietnamese - English sentences from both bilingual (such as The Tale of Kieu) and monolingual resources (The Vietnamese resource comes from thivien.net and the English resource we get from many open libraries and websites on the Internet). The detailed information about the Vietnamese literary and the contribution of each literary to the corpus is listed in the Table 1 below:

Title	Number of sentences
Dumb Luck - Số đỏ	2663
Chinese poem before the 10th century	297
A song of sorrow inside royal harem - Cung oán ngâm khúc	355
Exhortation to Military Officers - Hịch tướng sĩ	111
Prison diary - Nhật ký trong tù	478
Proclamation of Victory - Bình Ngô Đại Cáo	123
The marvelous neuter at Blue Creek - Bích Câu kỳ ngộ	649
The constant mouse - Trinh thử	522
The quarrel of the six beasts - Lục súc tranh công	448
Poems of Hàn Mặc Tử	233
Poems of Nguyễn Bỉnh Khiêm	88
Poems of Nguyễn Khuyến	107
Poems of Tố Hữu	155
Vietnamese anthology	1784
Poems of Hồ Xuân Hương	31
Lament of the soldier wife (two versions) - Chinh phụ ngâm	471
The tale of Kieu (two versions) - Truyện Kiều	6508
Poems of Lục Vân Tiên	216
Total	**15239**

Table 1: List of literary works and the corresponding number of sentences used for fine-tuning

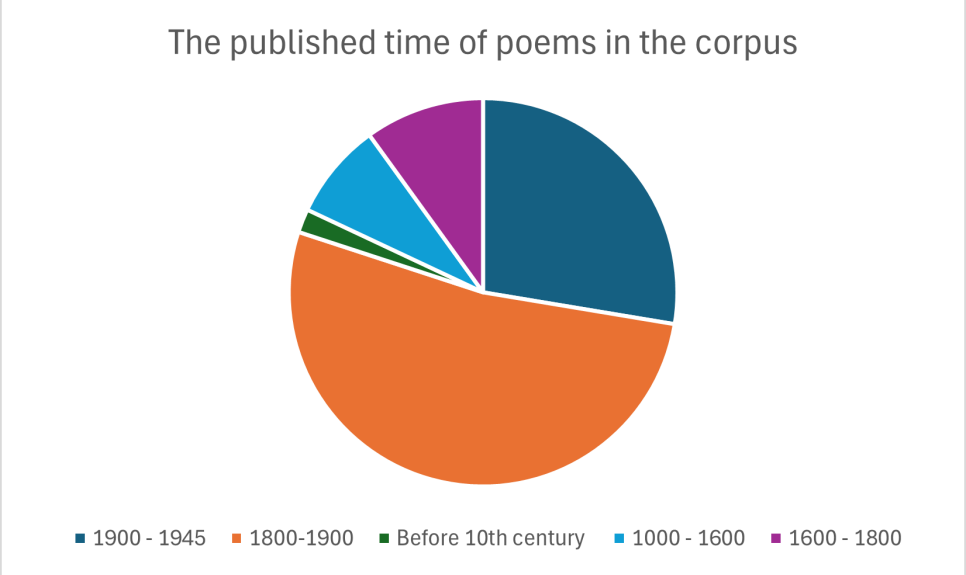

Figure 1: The published time of poems in the corpus

As we have shown in the Figure 1, more than half of the poems are published between 19th and 20th century. However, we still try to maintain the variation of our data in the corpus by collecting poems from various dynasties.

2.2 Automatic Alignment

After getting the resource and preprocesing the data, the final step of building a corpus for machine translation is aligning the Vietnamese sentence (source) with the target sentence in English. Except for the data coming from the bilingual/multilingual resources, aligning the independent monolingual resources is a big problem since they often have different number of sentences.

Since we don't really have a dictionary to map a Vietnamese input to an English output and we also don't have more than one translation in a majority of poems, we can't apply traditional methods for automatic alignment such as Levenshtein [18].

In this project, we utilize Language-agnostic BERT Sentence Embedding (LaBSE [17]) for automatic alignment. The approach is simple: we encode the sentences in Vietnamese and English using the sentence-embedding model LaBSE, then, with each Vietnamese sentence, we compute the cosine similarity of that Vietnamese embeddings to the English embeddings within a "window". By taking the difference in the number of sentences between Vietnamese and English poems, the window is specified by selecting all sentences whose distances from the source sentence are lower than the difference. The result of that operation will be a vector of similarity scores of size $1 \times N$ (N is the size of the window). Continuing the operations above will result in a 2D matrix of size $M \times N$, M is the number of sentences in Vietnamese poems.

Through several experiments, we find out that the threshold 0.55 will ensure that the alignment is correct. Thus, we apply this threshold to filter all scores satisfied the constraint. However, it's possible that there exists more than one score that's greater than 0.55. In this case, we decide to choose the sentence which is closest to the source one.

In many cases, the correct pairs of Vietnamese-English sentence don't have the similarity score greater than the threshold. In this case, our group defines an "acceptable" threshold, which is 0.45, and we choose the pair of sentences having the highest similarity score: if the score of that pair is greater than our acceptable threshold, then the pair is also included in the final alignment result.

With the alignment algorithm above, we can align automatically about 35 to 60% of the sentence (the result has been verified carefully with human inspection). The result is not really high because the algorithm depends greatly on the sentence-embedding model, which is not so good especially when it meets old Vietnamese words

3 Data Augmentation Methodology

Given the limited availability of parallel data for Ancient Vietnamese to English translation, it is crucial to employ synthetic data generation or data augmentation techniques to enhance the dataset.

Three primary techniques can be utilized:

- Back-translation [3].
- Perturbation [4].
- Data synthesis using pre-trained models [5].

For this study, we adopt data synthesis using pre-trained models since they offers sufficient performance across various evaluation metrics, including traditional BLEU [12], BLEURT [14], and SacreBLEU [13].

Building on the methodology outlined in Data Augmentation for English-Vietnamese Neural Machine Translation: An Empirical Study [5], we propose the following pipeline for synthetic data generation.

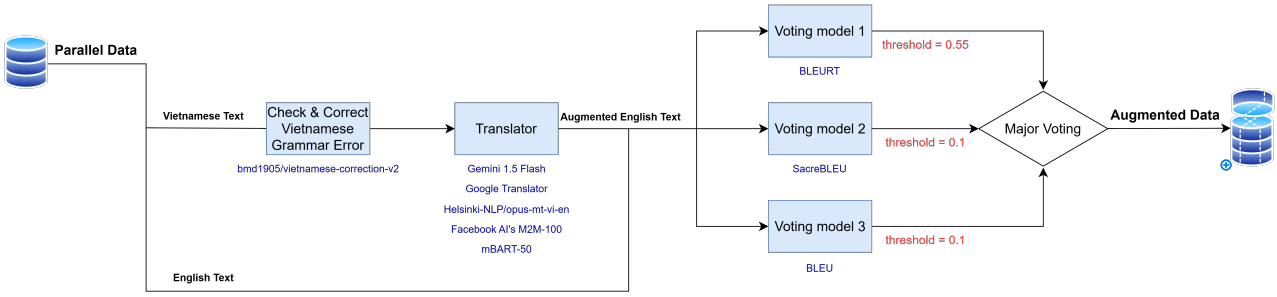

Figure 2: Data Augmentation Pipeline.

3.1 Check and Correct Grammar Mistakes

We use `bmd1905/vietnamese-correction-v2` model [6] to check and correct grammatical errors for Vietnamese text. We consider this grammatical correction like a machine translation task, where the wrong grammar text and the right grammar text are treated as the source language and the target language, respectively.

3.2 Translator Models

Specifically, we integrate the following translator models into our pipeline:

- Gemini 1.5 Flash [7].
- Google Translator [9].
- Helsinki-NLP/opus-mt-en-vi [8].
- Facebook AI's M2M-100 [10]
- mBART-50 [11].

Gemini 1.5 Flash is particularly noteworthy because of its accessibility (15 requests per day and 1,500 requests per minute under the free API key version) [7]. This makes it a practical choice for resource-constrained scenarios, combining stability and balanced performance.

3.3 Quality Control Through Voting Models

To ensure the quality of the generated synthetic data, we apply the Majority Voting technique to decide whether to keep the newly generated augmented English. In detail, majority voting counts the votes of the base learners and predicts the final labels as the label with the

majority of votes. In comparison to unweighted averaging, majority voting is less biased towards the outcome of a particular base learner as the effect is mitigated by majority vote count [15]. Thus, if two of the three classification models (voting models) agree on the translation output of Translator.

By conducting several experiments, we observe that the translation quality is ensured by the lower bound of 0.55 for BLEURT, 0.1 for SacreBLEU, and 0.1 for traditional BLEU.

3.4 Summary of Total Augmented Data

The data augmentation process yielded a total of 2,181 augmented lines.

4 Model Fine-Tuning

After collecting and cleaning the dataset, we proceeded to the fine-tuning phase. This step is also the most critical in obtaining a final model capable of effectively understanding ancient Vietnamese texts and translating them into English.

We conducted fine-tuning on various T5 model variants using the same **hyperparameter configuration** shown in Table 2.

Hyperparameter	**Value**
Number of Epochs	100
Learning Rate	1×10^{-4}
Batch Size	128
Weight Decay	0.01

Table 2: Hyperparameter Configuration for Fine-Tuning

For **training infrastructure**, we use NVIDIA A100 80GB and use FP16 for faster computation.

5 Experiment

5.1 Experimental Setup

To evaluate the performance of multiple models, we divided our dataset into training and testing sets. The test set comprised 1,524 sentence pairs, accounting for 10% of the data, while the remaining 90% was used for training.

We compare the models across three stages:

- **Pretrained**: Performance of the base model without any fine-tuning.
- **Fine-tuned**: Performance after training on collected dataset.
- **Fine-tuned (Aug.)**: Performance after training on collected dataset and its augmented version.

Initially, we faced a dilemma in selecting the optimal model for deployment between `VietAI/envit5-translation` and `NlpHUST/t5-vi-en-base` due to conflicting metrics:

- `VietAI/envit5-translation` achieves higher BLEURT scores.
- `NlpHUST/t5-vi-en-base` has better SacreBLEU scores.

To make an informed decision, we evaluate both models at three aforementioned stages.

5.2 Evaluation Metrics

To assess the quality of the translations, we calculated BLEU scores using two distinct metrics:

- SacreBLEU [13]: SacreBLEU offers significant improvements over the traditional BLEU metric by providing a standardized, reproducible, and user-friendly evaluation tool. The results are shown in Table 3.

- BLEURT [14]: BLEURT offers a more accurate, nuanced, and reliable metric compared to traditional methods like BLEU. The results are shown in Table 4.

5.3 Results

Model Name	**SacreBLEU Scores**		
	Pretrained	**Fine-tuned**	**Fine-tuned (Aug.)**
VietAI/envit5-translation	1.863	5.306	6.557
NlpHUST/t5-vi-en-base	2.041	2.987	3.287

Table 3: Comparison of SacreBLEU Scores (on a scale of 100) Between Pretrained, Fine-tuned, and Fine-tuned on Augmented Data Models

Model Name	**BLEURT Scores**		
	Pretrained	**Fine-tuned**	**Fine-tuned (Aug.)**
VietAI/envit5-translation	27.658	32.178	33.812
NlpHUST/t5-vi-en-base	26.608	27.642	27.736

Table 4: Comparison of BLEURT Scores (on a scale of 100) Between Pretrained, Fine-tuned, and Fine-tuned on Augmented Data Models

5.4 Conclusion

We selected the fine-tuned `VietAI/envit5-translation` model for the backend because it achieved the highest scores in both SacreBLEU (6.557) and BLEURT (33.812) evaluations after applying data augmentation technique and fine-tuning process.

6 Web User Interface

Finally, to make our product more accessible, we developed a web user interface using Flask [16]. Users can input Vietnamese ancient text, and our system will process it with the model and return the translated English text, which is displayed on the web interface.

Now, let's look at some examples of translations produced by our system.

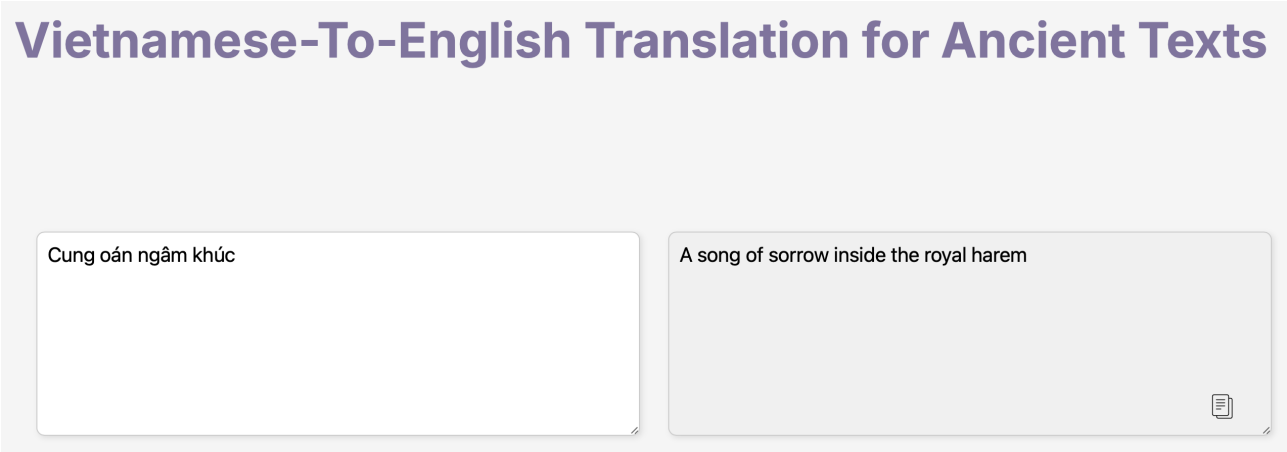

Figure 3: Example 1–A song of sorrow inside the royal harem.

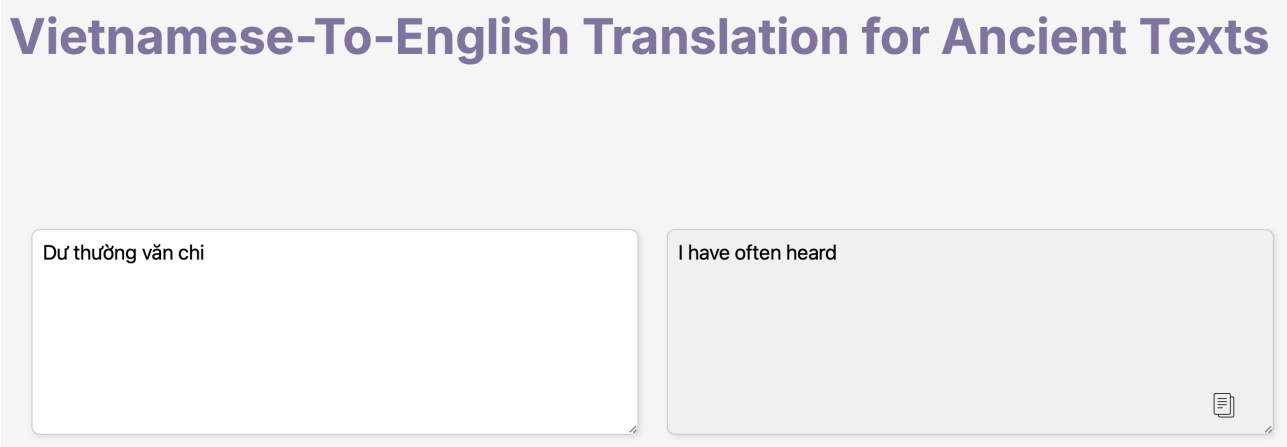

Figure 4: Example 2–I have often heard.

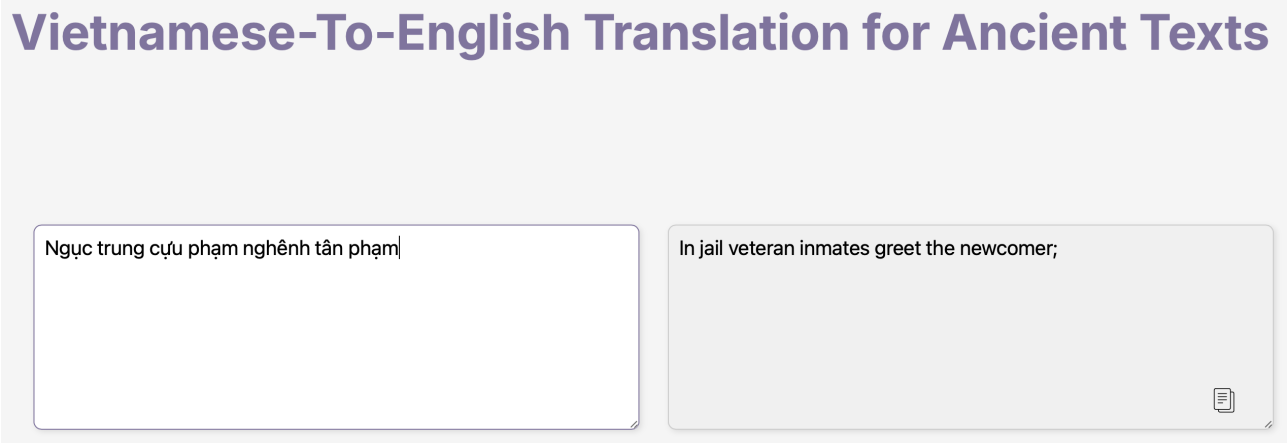

Figure 5: Example 3–In jail veteran inmates greet the newcomer.

References

[1] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv. https://arxiv.org/abs/1910.10683

[2] Ngo, C., Trinh, T. H., Phan, L., Tran, H., Dang, T., Nguyen, H., Nguyen, M., & Luong, M.-T. (2022). MTet: Multi-domain Translation for English and Vietnamese. arXiv. https://doi.org/10.48550/arxiv.2210.05610

[3] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

[4] Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O. Mann, T. A. (2021). Data Augmentation Can Improve Robustness. Advances in Neural Information Processing Systems (NeurIPS), 34, 29935–29948. https://proceedings.neurips.cc/paper/2021/file/fb4c48608ce8825b558ccf07169a3421-Paper.pdf.

[5] Pham, N. L. (2022). Data Augmentation for English-Vietnamese Neural Machine Translation: An Empirical Study. SSRN. https://ssrn.com/abstract=4216607 or http://dx.doi.org/10.2139/ssrn.4216607

[6] bmd1905. (2023). Vietnamese Correction v2. Hugging Face. https://huggingface.co/bmd1905/vietnamese-correction-v2.

[7] Google AI. Gemini API Documentation. Accessed January 2, 2025. https://ai.google.dev/gemini-api/docs/models/gemini

[8] Tiedemann, J., & Thottingal, S. (2020). OPUS-MT: Open Source Neural Machine Translation Models. *Helsinki-NLP*. Available at https://huggingface.co/Helsinki-NLP/opus-mt-en-vi.

[9] Nidhal Imame. (2020). Deep Translator: A Flexible Library for Multiple Translation Providers. GitHub Repository. https://github.com/nidhaloff/deep-translator.

[10] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020). Beyond English-Centric Multilingual Machine Translation. arXiv. https://arxiv.org/abs/2010.11125.

[11] Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2020). Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. arXiv. https://arxiv.org/abs/2008.00401.

[12] 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 on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[13] 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, Brussels, Belgium. Association for Computational Linguistics.

[14] Sellam, Thibault, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning Robust Metrics for Text Generation. In Proceedings of the ACL, 2020. https://arxiv.org/abs/2004.04696

[15] ScienceDirect. Majority Voting. Available at https://sciencedirect.com/topics/computer-science/majority-voting. Accessed January 2, 2025.

[16] Grinberg, M. (2018). Flask web development: developing web applications with python. O'Reilly Media, Inc.

[17] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, *"Language-agnostic BERT Sentence Embedding,"* arXiv preprint arXiv:2007.01852, 2022. [Online]. Available: https://arxiv.org/abs/2007.01852

[18] Y. Li and B. Liu, "A normalized Levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.