VNJPTranslate: A comprehensive pipeline for Vietnamese-Japanese translation

Hoang Hai Phan[1], Nguyen Duc Minh Vu[2], and Phuong Nam Dang[3]

[1]Hai.PH225715@sis.hust.edu.vn
[2]D3etMe4n@gmail.com
[3]Nam.DP225892@sis.hust.edu.vn

March 31, 2025

Abstract

Neural Machine Translation (NMT) driven by Transformer architectures has advanced significantly, yet faces challenges with low-resource language pairs like Vietnamese-Japanese (Vi-Ja). Issues include sparse parallel data and handling linguistic/cultural nuances. Recent progress in Large Language Models (LLMs) with strong reasoning, often refined via Reinforcement Learning (RL), enables high-quality synthetic data generation. We introduce VNJPTranslate, a pipeline designed to systematically address the Vi-Ja translation task. It features a targeted data augmentation strategy using advanced LLMs with Chain-of-Thought prompting for challenging segments identified via corpus analysis. Subsequently, we employ efficient fine-tuning techniques (Unsloth with QLoRA) on a capable, low-parameter autoregressive model (specifically, a fine-tuned version of the 1.8B parameter Sailor model, which is based on the Qwen architecture) to create a practical and high-performing translation system. This integrated approach aims to improve Vi-Ja translation quality significantly over existing baselines.

Keywords: Neural Machine Translation, Low-Resource Languages, Vietnamese-Japanese, Transformer Models, Large Language Models, Synthetic Data Generation, Chain-of-Thought Prompting, Data Augmentation, Efficient Fine-tuning, QLoRA, Unsloth, Sailor, Qwen.

1 Introduction

Neural Machine Translation (NMT) represents the state-of-the-art in automated translation, achieving remarkable performance, particularly for high-resource language pairs [Vaswani et al., 2017]. The introduction of the Transformer architecture [Vaswani et al., 2017] was pivotal, enabling models to capture long-range dependencies and train efficiently on massive datasets. However, this success does not uniformly extend to low-resource language pairs, where the scarcity of parallel corpora remains a critical bottleneck [Gu et al., 2018, Zhang et al., 2022a].

The Vietnamese-Japanese (Vi-Ja) pair is a pertinent example of this low-resource challenge [Ngo et al., 2022]. Beyond data scarcity, significant typological differences between Vietnamese (analytic, SVO, tonal) and Japanese (agglutinative, SOV, pitch-accent) and distinct cultural contexts introduce complexities that standard NMT models struggle to handle effectively with limited training data [Wang et al., 2024, Ngo et al., 2019]. Consequently, a noticeable quality gap persists compared to high-resource translation tasks.

Addressing this gap requires specialized approaches. Recent advances in Large Language Models (LLMs) offer promising avenues, particularly their capacity for high-fidelity text generation and sophisticated reasoning, often enhanced through techniques like Reinforcement Learning from Human Feedback (RLHF) [Ouyang et al., 2022, et al., 2025]. These capabilities can be harnessed to generate synthetic parallel data, augmenting scarce authentic resources [Shao et al., 2023a].

This paper introduces VNJPTranslate, a comprehensive pipeline specifically designed to improve Vi-Ja NMT. Our core contributions are twofold: (1) A targeted data preparation strategy that identifies difficult-to-translate segments (e.g., containing rare words) and leverages a powerful LLM with Chain-of-Thought (CoT) prompting and few-shot examples to generate high-quality synthetic translations for these segments. (2) The application of highly efficient fine-tuning techniques (Unsloth with 4-bit QLoRA) [Unsloth AI, 2024c, Dettmers et al., 2023] to adapt a capable autoregressive LLM (specifically, the fine-tuned 'thangvip/vilaw-sailor-instruct-v3' model [thangvip, 2024, Le et al., 2024], based on the Sailor/Qwen architecture [Dou et al., 2024, Qwen Team, 2024]) for the Vi-Ja translation task, resulting in a performant yet deployable model.

The remainder of this paper is structured as follows: Section 2 discusses related work in NMT architectures, low-resource translation techniques, and the use of LLMs for data augmentation. Section 3 details the VNJPTranslate pipeline, covering dataset preparation and model fine-tuning. Section 3.3 presents our evaluation setup and results. Finally, Section 4 concludes the paper. (A dedicated Conclusion section would be added in a full paper).

2 Related Works

2.1 Neural Machine Translation Architectures

The field of NMT evolved from early sequence-to-sequence (seq2seq) models based on Recurrent Neural Networks (RNNs) [Sutskever et al., 2014, Cho et al., 2014]. While pioneering, these models faced limitations with long sequences. The Transformer architecture [Vaswani et al., 2017], relying entirely on self-attention mechanisms, overcame these limitations, enabling parallelization and superior modeling of dependencies, quickly becoming the dominant paradigm. Frameworks like T5 [Raffel et al., 2020] and its multilingual variants (e.g., mT5 [Xue et al., 2021]) further solidified the effectiveness of the Transformer encoder-decoder structure for translation by leveraging massive pre-training.

More recently, large-scale decoder-only autoregressive models (e.g., GPT series [Brown et al., 2020], Qwen series [Qwen Team, 2024]) have gained prominence. While standard translation benchmarks often still favor optimized encoder-decoder models [Li et al., 2023, Bheemaraj et al., 2024], decoder-only models offer advantages in generative fluency and flexibility via prompting [Slator, 2024]. Techniques like few-shot learning [Brown et al., 2020] and Chain-of-Thought (CoT) prompting [Wei et al., 2022] allow these models to perform complex tasks, including translation or related data generation, with minimal task-specific training [Lal et al., 2024]. Our work leverages a decoder-only architecture for its adaptability and fine-tuning efficiency.

2.2 Low-Resource Neural Machine Translation

Addressing data scarcity in NMT is a significant research area [Zhang et al., 2022a]. Common techniques include transfer learning (leveraging models trained on high-resource pairs) [Gu et al., 2018], multilingual NMT (training a single model for multiple language pairs) [Aharoni et al., 2019, Xue et al., 2021], and back-translation (generating synthetic source sentences from monolingual target data). Data augmentation through paraphrasing or other manipulations of existing parallel data is also employed [Hou et al., 2022]. Our work focuses on synthetic data generation directly from the source language using advanced LLMs, complementing existing parallel data rather than relying solely on target monolingual data (as in back-translation) or complex transfer learning setups. Ngo et al. [2022] specifically investigated synthetic data generation for Vi-NMT, providing context for our targeted approach. Ngo et al. [2019] highlighted the challenges posed by rare words in low-resource Vi-NMT, motivating our targeted refinement strategy.

2.3 LLMs for Data Augmentation and Optimization

Modern LLMs, particularly those exhibiting strong reasoning capabilities enhanced via RL [Ouyang et al., 2022, et al., 2025], are increasingly used for data augmentation. Their ability to generate fluent and contextually relevant text makes them suitable for creating synthetic parallel corpora [Shao et al., 2023a]. CoT prompting has been shown to improve the quality of generation for complex tasks [Wei et al., 2022, Shao et al., 2023b], which we adapt for generating high-fidelity translations.

Furthermore, related optimization techniques can enhance NMT pipelines. Knowledge Distillation (KD) allows transferring knowledge from large teacher models (potentially used for data generation) to smaller student models [Hinton et al., 2015, Kim and Rush, 2016, Sun et al., 2021]. Active Learning (AL) focuses training efforts on the most informative data samples [Zhang et al., 2022b, Olsson, 2009, Beyer et al., 2022]. While not explicitly implementing KD or AL in the current pipeline, our approach aligns with the principle of leveraging powerful models (for targeted data generation) and efficiently training a smaller model. The use of efficient fine-tuning methods like QLoRA [Dettmers et al., 2023] via libraries like Unsloth [Unsloth AI, 2024c] further contributes to optimizing the training process for practical deployment.

3 The VNJPTranslate Pipeline

Our methodology focuses on creating a high-quality dataset tailored for Vi-Ja translation and efficiently fine-tuning a suitable NMT model.

3.1 Dataset Preparation

We employ a multistage process to construct and refine our parallel corpus, aiming to maximize quality while managing resources efficiently.

3.1.1 Initial Corpus Generation

A baseline parallel corpus is first established. We utilize available Vietnamese text resources (potentially large scale, drawing inspiration from dataset construction methodologies like those described in HIRANO et al. [2023] for Japanese datasets). This source text is translated into Japanese using a computationally inexpensive smaller LLM optimized for throughput. This step produces an initial, potentially large, but possibly noisy Vi-Ja dataset.

3.1.2 Targeted Refinement Strategy

To enhance translation quality, especially for sentences susceptible to errors in the initial translation, we applied a targeted refinement using a more powerful LLM:

1. **Identification of Challenging Sentences:** We analyze the target (Japanese) side of the initial corpus using a Bag-of-Words (BoW) approach to identify sentences containing low-frequency Japanese words. This heuristic targets sentences likely dealing with less common concepts or specific terminology, which often pose challenges for NMT models trained on limited data [Ngo et al., 2019]. We flag sentences containing words below a frequency threshold set to capture approximately 15% of the corpus, focusing refinement efforts.

2. **Contextual Few-Shot Example Retrieval:** For each flagged Vietnamese source sentence (S_{vi}), we retrieve contextually relevant examples

to guide the advanced LLM's translation process. Using the BM25 relevance scoring algorithm [Robertson and Zaragoza, 2009], we identify the top-3 most similar source sentences (S'_{vi}) from a high-quality subset of the initial corpus (or a dedicated clean set). Their corresponding initial translations (T'_{ja}) are used as few-shot demonstrations within the prompt, a technique known to improve in-context learning performance [Brown et al., 2020, Zhang et al., 2024, LangChain Team, 2024].

3. **CoT-Prompted Synthetic Translation Generation:** We employ a powerful reasoning LLM, DeepSeek-V3 [Neontri, 2025, DeepSeek AI, 2025], noted for its multilingual proficiency and reasoning capacity [GeeksforGeeks, 2025], to re-translate the identified challenging source sentences (S_{vi}). Crucially, we utilize Chain-of-Thought (CoT) prompting [Wei et al., 2022] to explicitly guide the model through intermediate reasoning steps before producing the final translation, enhancing translation fidelity [Shao et al., 2023a, K2view, 2024]. For each flagged S_{vi}, we generate two diverse but high-quality Japanese translations (T^1_{ja}, T^2_{ja}) by slightly varying generation hyperparameters (e.g., temperature set to 0.7 and 0.85, respectively).

The final training corpus combines the initial baseline translations (for non-flagged sentences) with the higher-quality, CoT-generated synthetic pairs (S_{vi}, T^1_{ja}) and (S_{vi}, T^2_{ja}) for the flagged, challenging sentences. This creates an augmented dataset enriched specifically where the baseline model likely struggles.

3.2 VNJPTranslate Model Fine-tuning

3.2.1 Base Model Selection

Our chosen model for fine-tuning is 'thangvip/vilaw-sailor-instruct-v3' [thangvip, 2024], a model developed within the context of Vietnamese legal NLP [Le et al., 2024]. This model is based on the Sailor project's 1.8B parameter models [Dou et al., 2024], which utilize the Qwen architecture [Qwen Team, 2024]. It offers strong performance, particularly in Vietnamese contexts, and its architecture supports efficient fine-tuning.

3.2.2 Efficient Fine-tuning with Unsloth and QLoRA

To maximize training efficiency, we employ the Unsloth library [Unsloth AI, 2024c, Shakil, 2024], which significantly accelerates training speed and reduces memory usage compared to standard methods [Unsloth AI, 2024b, 2025b]. Specifically, we utilize 4-bit QLoRA (Quantized Low-Rank Adaptation) [Dettmers et al., 2023] implemented via Unsloth [Unsloth AI, 2024a, 2025a]. QLoRA enables fine-tuning large models with dramatically less memory by quantizing the base model weights and training low-rank adapters, while largely preserving model performance. This combination allows effective fine-tuning of the 1.8B parameter Sailor-based model on accessible hardware (Colab's Tesla T4 and Kaggle's P100).

3.2.3 Training Procedure

The chosen 'thangvip/vilaw-sailor-instruct-v3' model is fine-tuned using a standard supervised learning objective on the augmented parallel corpus created in Section 3.1. Training hyperparameters (learning rate schedule, batch size, number of epochs) are tuned based on performance on a held-out validation set, primarily optimizing for the BLEU score [Papineni et al., 2002].

3.3 Evaluation Setup

Performance is evaluated on the Vietnamese-Japanese test set from the Tatoeba portion of the OPUS corpus collection, commonly used for evaluating Helsinki-NLP models [Helsinki-NLP, 2020]. We use the BLEU score [Papineni et al., 2002] as the primary metric. We compare our fine-tuned VNJPTranslate model against the established Helsinki-NLP/opus-mt-ja-vi baseline [Helsinki-NLP, 2020].

Table 1: Comparison of BLEU Scores on Vi-Ja Tatoeba Test Set

Model	BLEU Score
Helsinki-NLP/opus-mt-ja-vi	20.30
thangvip/vilaw-sailor-instruct-v3 (fine-tuned)	28.30

Table 1 presents the comparison, showing a significant improvement achieved by our fine-tuned model.

4 Conclusion

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation in the wild: Findings and challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 391–402, 2019. URL `https://aclanthology.org/W19-5335/`.

Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Understanding the success of knowledge distillation – a data augmentation perspective. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=kk_izlQ0F4`.

Nagaraju Naik Bheemaraj, Srinidhi Vadlamani, Sravani Goli, and Radhika Mamidi. Machine translation with large language models: Decoder only vs. encoder-decoder, 2024. URL `https://arxiv.org/abs/2409.13747`.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Ad-*

vances in neural information processing systems, volume 33, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL https://doi.org/10.3115/v1/D14-1179.

DeepSeek AI. deepseek-ai/deepseek-v3-0324, March 2025. URL https://huggingface.co/deepseek-ai/DeepSeek-V3-0324.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 1–21, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6031f04c9532b906742acd4827447d4c-Paper-Conference.pdf.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. Sailor: Open language models for south-east asia, 2024. URL https://arxiv.org/abs/2404.03608.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

GeeksforGeeks. Deepseek r1 vs v3: A head-to-head comparison of two ai models. GeeksforGeeks Article, January 2025. URL https://www.geeksforgeeks.org/deepseek-r1-vs-v3/.

Jiatao Gu, Yong Wang, Kyunghyun Tran, and Victor OK Li. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3361–3366, 2018. URL https://aclanthology.org/D18-1376/.

Helsinki-NLP. opus-mt-ja-vi, 2020. URL https://huggingface.co/Helsinki-NLP/opus-mt-ja-vi.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Masanori HIRANO, Masahiro SUZUKI, and Hiroki SAKAJI. llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology, 2023. URL https://arxiv.org/abs/2305.12720.

Feifan Hou, Haochen Zhang, Wenyi Ding, Jinlan Tang, and Joey Hui. Data augmentation with diversified rephrasing for low-resource neural machine translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 37–43, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-srw.5.

K2view. Chain-of-thought reasoning supercharges enterprise llms. K2view Blog, October 2024. URL https://www.k2view.com/blog/chain-of-thought-reasoning-supercharges-enterprise-llms/.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327, 2016. URL https://aclanthology.org/D16-1139/.

Yash Kumar Lal, Jivnesh Sandhan, Ravi Tejomurtula, Pawan Goyal, and Laxmidhar Behera. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages, 2024. URL https://arxiv.org/abs/2409.04512.

LangChain Team. Dynamic few shot example selection. LangSmith Documentation, 2024. URL https://docs.smith.langchain.com/guides/develop/datasets/dynamic-few-shot. Accessed: 2025-03-30.

Thang V. Q. Le, Dinh-Hong Vu, Van-Huy Pham, Anh-Cuong Le, and Nguyen P. Nguyen. A framework for vietnamese question-answering in law domain. In *2024 IEEE 9th International Conference on Data Science in Cyberspace (DSC)*, pages 726–731, 2024. doi: 10.1109/DSC63484.2024.00108.

Haode Li, Jingjing Xu, Qi Liu, Wenqiang Lei, Chunheng Wang, and Tat-Seng Chua. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder, 2023. URL https://arxiv.org/abs/2304.03730.

Neontri. Breaking down deepseek: Key features and risks. Neontri Blog, March 2025. URL https://neontri.com/insights/breaking-down-deepseek-key-features-and-risks.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. Overcoming the rare word problem for low-resource language pairs in neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 207–214, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5228. URL https://aclanthology.org/D19-5228/.

Thi-Vinh Ngo, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha, and Le-Minh Nguyen. An efficient method for generating synthetic data for low-resource machine translation – an empirical study of chinese, japanese to vietnamese neural machine translation. *Applied Artificial Intelligence*, 36(1):

2101755, 2022. doi: 10.1080/08839514.2022.2101755. URL https://doi.org/10.1080/08839514.2022.2101755.

Fredrik Olsson. Active learning literature survey. 2009. URL https://www.semanticscholar.org/paper/Active-Learning-Literature-Survey-Olsson/e910293a70053429c787a7cb82242e24e73e6d41.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen Team. Qwen2.5: A Party of Foundation Models. Blog Post, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Muhammad Shakil. Unsloth guide: Optimize and speed up llm fine-tuning. DataCamp Tutorial, October 2024. URL https://www.datacamp.com/tutorial/unsloth-guide-optimize-and-speed-up-llm-fine-tuning.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models, 2023a. URL https://openreview.net/forum?id=CI8OFzFA6s.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models, 2023b. URL https://arxiv.org/abs/2302.00618.

Slator. A primer on decoder-only vs encoder-decoder models for ai translation. Slator Article, October 2024. URL https://slator.com/a-primer-on-decoder-only-vs-encoder-decoder-models-for-ai-translation/.

Fusheng Sun, Jianhao Liu, Fei Huang, Zhiyuan Luo, and Yang Liu. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2320–2330. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.181. URL https://aclanthology.org/2021.acl-long.181/.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

thangvip. thangvip/vilaw-sailor-instruct-v3. https://huggingface.co/thangvip/vilaw-sailor-instruct-v3, 2024.

Unsloth AI. Unsloth documentation - 4-bit quantization. Unsloth Documentation / Hugging Face Model Cards, 2024a. URL https://huggingface.co/collections/unsloth/quantized-models-6663296a85a6240a41b4631b.

Unsloth AI. Unsloth documentation - key features. Unsloth Documentation, 2024b. URL https://unsloth.ai/docs.

Unsloth AI. unslothai/unsloth: Finetune llama 3.3, deepseek-r1, gemma 3 & reasoning llms 2x faster with 70% less memory! https://github.com/unslothai/unsloth, 2024c.

Unsloth AI. Fine-tuning guide - unsloth documentation. Unsloth Documentation, March 2025a. URL https://unsloth.ai/fine-tuning.

Unsloth AI. Finetune phi-4 with unsloth. Unsloth Blog/Announcement, January 2025b. URL https://unsloth.ai/blog/phi-4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Xinrun Wang, Joseph B Adebayo, Wejdan Alhakami, Diane El Baz, and Ahmed A Abd El-Latif. Innovations and challenges in neural machine translation: A review. *Electronics*, 13(13):2531, 2024. doi: 10.3390/electronics13132531. URL https://doi.org/10.3390/electronics13132531.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.41. URL https://doi.org/10.18653/v1/2021.naacl-main.41.

Ruiqing Zhang, Chuanqiuyue Jiang, Diptesh S Sachan, Graham Neubig, and Yang Liu. A survey on low-resource neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5636–5644. International Joint Conferences on Artificial Intelligence Organization, 2022a. doi: 10.24963/ijcai.2022/788. URL https://doi.org/10.24963/ijcai.2022/788.

Yongqi Zhang, Zhengyuan Zhu, Zhiling Chen, and Yuen-Hsien Yang. Effective in-context example selection through data compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 871–880, Miami, Florida, October 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.51.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates, 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.414/.

A Dataset Token Count

This appendix provides additional details on the dataset used in the VNJP-Translate pipeline. In our analysis, we examine the token count of the corpus.

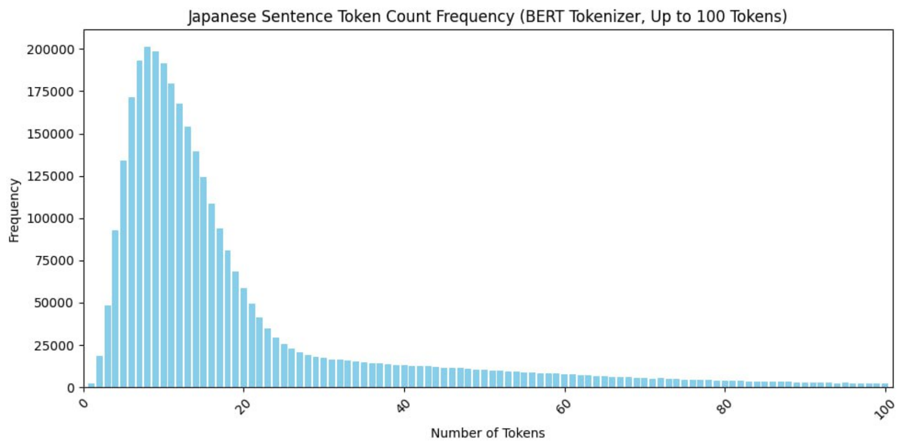

Figure 1: Japanese Token Count Frequency in Corpus

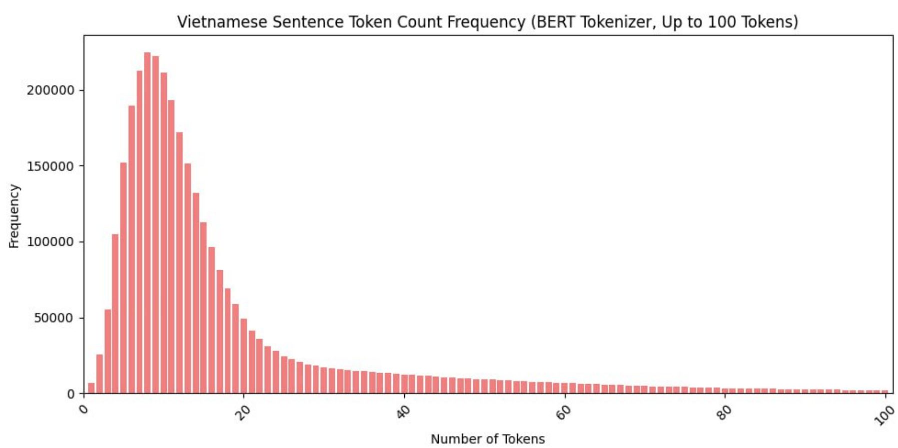

Figure 2: Vietnamese Token Count Frequency in Corpus