Exploring Parameter-Efficient Fine-Tuning and Backtranslation for the WMT 25 General Translation Task

Felipe Ribeiro Fujita de Mello[1], **Hideyuki Takada**[1]
[1] Ritsumeikan University, Japan
Correspondence: is0596kh@is.ritsumei.ac.jp

Abstract

In this paper, we explore the effectiveness of combining fine-tuning and backtranslation on a small Japanese corpus for neural machine translation. Starting from a baseline English→Japanese model (COMET = 0.460), we first apply backtranslation (BT) using synthetic data generated from monolingual Japanese corpora, yielding a modest increase (COMET = 0.468). Next, we fine-tune (FT) the model on a genuine small parallel dataset drawn from diverse Japanese news and literary corpora, achieving a substantial jump to COMET = 0.589 when using Mistral 7B. Finally, we integrate both backtranslation and fine-tuning—first augmenting the small dataset with BT generated examples, then adapting via FT—which further boosts performance to COMET = 0.597. These results demonstrate that, even with limited training data, the synergistic use of backtranslation and targeted fine-tuning on Japanese corpora can significantly enhance translation quality, outperforming each technique in isolation. This approach offers a lightweight yet powerful strategy for improving low-resource language pairs.

1 Introduction

Neural MT for Japanese benefits from recent large language models (LLMs) and recipe-driven data augmentation, but publicly documented, *small-corpus* workflows are scarce. This paper focuses on a minimalist, engineering-first pipeline that couples (i) supervised fine-tuning (FT) on a small Japanese corpus with (ii) backtranslation (BT) to expand coverage. Our objectives are:

- To give a **clear blueprint** that other researchers can adopt even with limited computing resources.
- To perform **transparent evaluation**, using well-established metrics such as **COMET** and **BLEU/chrF**.

2 Related Work

Research on improving neural machine translation (NMT) for Japanese has increasingly relied on two complementary techniques: backtranslation and fine-tuning. Early large-scale systems demonstrated that backtranslation is particularly effective for low-resource settings, as it leverages abundant monolingual corpora to generate synthetic parallel data. This method augments scarce bilingual datasets and helps reduce domain mismatch, which is a persistent challenge in English–Japanese translation.

Kiyono et al. (2020) investigated English–Japanese news translation at WMT 2020, showing that the combination of synthetic data through backtranslation and subsequent fine-tuning significantly improved performance over a baseline. Extending this line of work, Le et al. (2021) explored fine-tuning with domain-specific corpora and demonstrated that backtranslation enhanced adaptation to the news domain in the WMT 2021 shared task. Their study highlighted the importance of tailoring fine-tuning schedules when working with Japanese corpora.

Further refinements were presented by Morishita et al. (2022) in WMT 2022, who introduced a system that incorporated both extensive backtranslation and selective fine-tuning. Their approach confirmed that even moderate-scale synthetic corpora, when carefully integrated, yield measurable improvements in translation accuracy for Japanese. Similarly, Kudo et al. (2023) reported results from WMT 2023 where backtranslation and iterative fine-tuning were applied to robustly adapt transformer-based systems, demonstrating strong gains for English–Japanese translation.

In parallel, multilingual NMT research has also highlighted the value of backtranslation. Xu et al. (2021) proposed an auxiliary language framework, leveraging backtranslation across multiple lan-

Proceedings of the Tenth Conference on Machine Translation, pages 765–768

guage pairs, including Japanese. Their results suggest that cross-lingual signals derived from back-translation not only improve individual language directions but also enhance multilingual consistency.

These studies illustrate the central role of back-translation in augmenting limited Japanese corpora and show that fine-tuning, when combined with synthetic data, can consistently raise translation quality. They provide the empirical foundation for our own work for low-resource Japanese NMT.

3 System Architecture

Our proposed method combines fine-tuning on a small parallel Japanese–English dataset with back-translation to augment the available training data. As illustrated in Figure 1, monolingual Japanese sentences are first translated into English using the a pretrained model to create synthetic parallel pairs. These synthetic pairs are then used with the original data to fine-tune a pretrained model. The resulting system benefits from both the linguistic diversity of backtranslation and the domain adaptation of fine-tuning, leading to improved translation quality as measured by COMET, BLEU, and chrF++.

3.1 Implementation Details

The system builds on top of `AutoTokenizer` and `AutoModelForCausalLM`, enabling flexible experimentation with Mistral 7B[1] (Jiang et al., 2023). Parameter-efficient fine-tuning is employed to reduce computational demands, while training routines follow established best practices with gradient accumulation, mixed precision (`torch.float16`), and GPU offloading.

3.2 Dataset

For our experiments, we relied on the Japanese–English *WikiCorpus* released by Kyoto University[2]. This corpus consists of parallel sentences extracted from Wikipedia, providing high-quality and naturally occurring examples of Japanese usage. Given the limited scope of our study, we sampled a total of approximately 1,500 sentence pairs for training and validation.

3.3 Tokenization

For Japanese text, we adopt `fugashi`[3], a MeCab wrapper optimized for Python, which provides robust morphological analysis and segmentation. This ensures that the tokenizer can handle Japanese corpora effectively, producing consistent subword units that align with both training and backtranslation data.

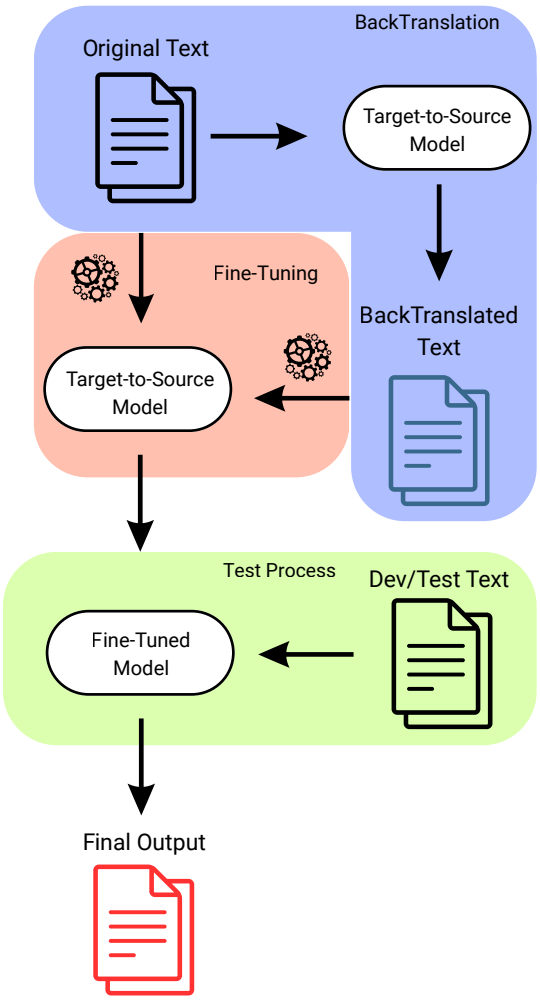

Figure 1: Overview of the proposed method

3.4 Backtranslation

Backtranslation (BT) is implemented by first using a pretrained model (Japanese → English) on the available parallel data as shown in Figure 2. Using this model, synthetic English sentences are generated from monolingual Japanese corpora. These synthetic pairs are then added to the original parallel dataset, effectively enlarging the training corpus. This augmentation proved crucial in mitigating data scarcity, providing additional coverage for domain-specific and colloquial expressions.

[1] https://huggingface.co/mistralai/Mistral-7B-v0.3

[2] https://alaginrc.nict.go.jp/WikiCorpus/

[3] https://github.com/polm/fugashi

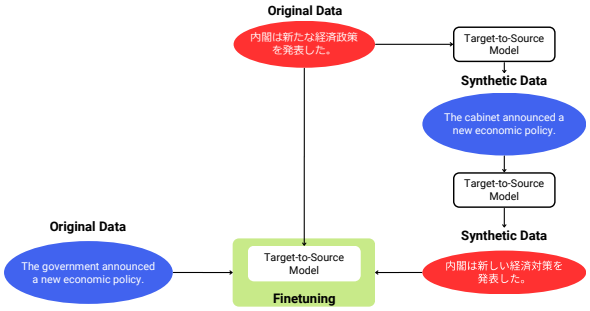

Figure 2: Overview of the backtranslation steps to generated synthetic data to serve as input along with the original data

3.5 Fine-Tuning Procedure

Fine-tuning (FT) is performed on a small, high-quality parallel dataset of Japanese corpora. The fine-tuning focuses on adapting pre-trained Mistral 7B to the translation domain. To make this process more efficient, we employ parameter-efficient fine-tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA) (Hu et al., 2021). This approach enables effective adaptation to Japanese with limited resources, making fine-tuning feasible even under hardware constraints.

3.6 Evaluation Metrics

Evaluation is conducted using both automatic and human-oriented metrics. Automatic scores include:

- **COMET** (Rei et al., 2020): a neural-based quality estimator, used as the primary evaluation metric.
- **BLEU** (Papineni et al., 2002) and **chrF** (Popović, 2015): reference-based metrics to provide comparability with prior work.

These metrics are computed on the validation set at each epoch and the final models are selected based on the best COMET score.

3.7 System Configuration

Our system is implemented using Hugging Face's `transformers` [4] library. The key hyperparameters and settings are summarized in Table 1.

Component	Configuration
Model	Mistral 7B (decoder-only)
Architecture	Transformer decoder, 32 layers, hidden size 4096
Training epochs	5–8
Batch size	128 (with gradient accumulation)
Minibatch size	4 per device (before accumulation)
Learning rate	2×10^{-5} (cosine schedule)
Max learning rate	3×10^{-5}
Warmup steps	500
Optimizer	AdamW
Weight decay	0.01
Dropout	0.1
Gradient clipping	1.0
Precision	Mixed (`float16`)
Decoding	Beam size 3, max new tokens 256, no sampling; length penalty 1.0
Logging	Save best checkpoint on COMET
Number of updates	10,000

Table 1: System configuration for fine-tuning with backtranslation.

4 Experiments

4.1 Setup

We fine-tune on 1.5k seed pairs and their BT-augmented counterparts (same domain). We segment documents on blank lines, translate at paragraph level, and enforce paragraph-count parity. We then merge to document level, verify, and score. Finally, we compared the results on several baselines to verify the system output compared to a state-of-art model.

4.2 Results

The results in Table 2 show several consistent trends. First, applying backtranslation (BT) to the baseline Mistral 7B model provided only a marginal gain in COMET (0.468 vs. 0.460) while simultaneously lowering BLEU, suggesting that synthetic data alone cannot compensate for the absence of high-quality parallel supervision.

In contrast, fine-tuning (FT) on the small but high-quality Japanese parallel dataset yielded a substantial improvement, raising COMET to 0.589 and demonstrating the strong impact of targeted adaptation. When FT was combined with BT, the model achieved the highest COMET score of 0.597, confirming that the synergy between synthetic augmentation and fine-tuning is beneficial.

However, BLEU slightly decreased compared to FT alone, indicating that n-gram overlap metrics do not always align with adequacy-oriented metrics like COMET. This divergence highlights the importance of using multiple evaluation measures: while BLEU and chrF++ capture surface similarity, COMET better reflects semantic adequacy and fluency.

Overall, the results suggest that FT is the main driver of quality improvement in low-resource

[4] https://github.com/huggingface/transformers

Japanese translation, while BT plays a supporting role by diversifying the training signal.

Model	BLEU	chrF	COMET
Mistral 7B Base	0.63	–	0.460
Mistral 7B Base + BT	0.18	–	0.468
Mistral 7B FT	1.97	–	0.589
Mistral 7B FT + BT	1.41	15.87	0.597

Table 2: Experimental results on Mistral 7B

5 Limitations

Our approach faces three main limitations. First, training on a small corpus makes the system highly sensitive to overfitting, requiring early stopping and regularization. Second, the effectiveness of backtranslation depends on the reverse model, as low-quality outputs can add noise; simple filtering methods such as length-ratio checks and language identification are necessary to maintain data quality. Finally, since the experiments rely on Wiki-derived text, there is a risk of domain shift when applying the model to other contexts, which may require domain adaptation.

6 Conclusion

In this work, we investigated the combined use of fine-tuning (FT) and backtranslation (BT) to improve English–Japanese neural machine translation under small-data conditions. The results show that parameter-efficient fine-tuning combined with carefully filtered backtranslation can provide a practical and effective blueprint for improving Japanese translation, even with limited computational resources. Future work will explore domain adaptation and scaling synthetic data generation to further enhance robustness.

References

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155, Online. Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 128–136, Singapore. Association for Computational Linguistics.

Giang Le, Shinka Mori, and Lane Schwartz. 2021. Illinois Japanese -> English News Translation for WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 144–153, Online. Association for Computational Linguistics.

Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase, and Jun Suzuki. 2022. NT5 at WMT 2022 general translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 318–325, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3029–3041, Punta Cana, Dominican Republic. Association for Computational Linguistics.