TransionMT's Submission to the Indic MT Shared Task in WMT 2025

ZeBiao Zhou and Hui Li and XiangXun Zhu and KangZhen Liu
Transsion, Shenzhen, Guangdong, China

Abstract

This study tackles the WMT 2025 low-resource Indic language translation task (EnglishAssamese, EnglishManipuri) by proposing a cross-iterative back-translation and data augmentation strategy using dual pre-trained models.Leveraging IndicTrans2_1B and NLLB_3.3B, the approach alternates fine-tuning and back-translation to iteratively generate high-quality pseudo-parallel corpora. Monolingual data relevance is enhanced via semantic similarity filtering with all-mpnet-base-v2, while training data is cleaned and normalized to improve quality. During inference, outputs from both fine-tuned models are combined to further boost translation performance in low-resource scenarios.

1 Introduction

India boasts a rich linguistic ecosystem, yet numerous languages suffer from limited digital resources. These low-resource languages face significant challenges in the construction and application of machine translation systems. Particularly, Assamese and Manipuri in the northeastern region not only lack parallel corpora and high-quality monolingual corpora but also exhibit large differences in linguistic structure and writing systems, posing additional difficulties for the training of Neural Machine Translation (NMT) models. Under low-resource conditions, traditional neural machine translation methods cannot fully leverage the advantages of large-scale data, resulting in limited model generalization ability and translation quality. Therefore, exploring how to efficiently utilize limited bilingual and monolingual data and effectively transfer cross-lingual knowledge has become a core issue in improving the translation performance of low-resource languages.

To tackle this challenge, this study participates in the WMT 2025 Low-Resource Indic Language Translation track, focusing on two translation directions:EnglishAssamese and EnglishManipuri, and proposes a cross-iterative back-translation and data augmentation method based on dual pre-trained models.The study selects open-source IndicTrans2_1B and NLLB_3.3B as the core translation models, combines multiple rounds of iterative back-translation to generate high-quality pseudo-parallel corpora, uses semantic similarity filtering technology to enhance the alignment between monolingual data and the task, and reduces the interference of noisy data on training through strict data cleaning and standardization operations. During the inference phase, the outputs of the two models are compared and selected, with the optimal result serving as the final translation output. This study aims to verify the effectiveness of cross-model collaborative back-translation mechanisms, data similarity augmentation, and multi-source result fusion in low-resource translation tasks, providing reusable technical routes and empirical experience for future multilingual low-resource machine translation research.

2 Dataset

All parallel data used in this study are derived from the official bilingual data provided by the WMT 2025 Low-Resource Indic Language Translation track, covering two directions: EnglishAssamese (en-as) and EnglishManipuri (en-mni). The data scale is shown in Table 1. Among them, the en-as direction contains 54,000 training sentence pairs, the en-mni direction contains 23,000 training sentence pairs, and both directions provide validation sets and test sets respectively.

A sampling analysis of the official test set reveals a concentrated domain distribution: healthcare accounts for 65.29%, entertainment and sports for 23.56%, and culture for 11.15%. To maximize domain consistency with the test set during the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1029–1033
November 8–9, 2025.

Language Pair	Train	Val	Test
en-as	54,000	2,000	2,000
en-mni	23,000	1,000	1,000

Table 1: Scale of WMT 2025 Official Bilingual Dataset

data augmentation phase, the study collects English monolingual data from the NLLB open corpus, BPCC open-source dataset, and specific website crawls. The open-source semantic similarity model all-mpnet-base-v2 is used to calculate the semantic similarity between the collected data and the test set samples. Sampling and filtering are performed in high-similarity data according to the above domain proportions, ultimately obtaining approximately 100,000 highly relevant English monolingual sentences for back-translation to generate pseudo-parallel corpora.

During the data cleaning phase, strict processing is uniformly applied to bilingual and monolingual data: removing sentences containing URLs, HTML tags, and non-linguistic characters; eliminating samples that failed to be translated or deviated from the source language in back-translation; standardizing symbols, checking and correcting English capitalization rules for the first letter; and removing duplicate sentences and abnormally short sentences. These operations significantly reduce the proportion of noisy data and ensure that the data domain distribution is highly consistent with the official test set, providing high-quality data support for subsequent cross-iterative back-translation and model optimization.

3 System Methodology

3.1 Pre-trained Models

This study is based on two open-source multilingual neural machine translation pre-trained models: IndicTrans2_1B (Kunchukuttan et al., 2023) and NLLB_3.3B (Fan et al., 2022).

• **IndicTrans2_1B**: A Transformer-based machine translation model optimized for 22 official languages of India and various related languages. It performs excellently in many-to-many, many-to-one, and one-to-many translation tasks, especially suitable for handling Indic languages with complex morphology and scarce training data (Kunchukuttan et al., 2023).

• **NLLB_3.3B (No Language Left Behind)**: A large-scale multilingual translation model proposed by Meta AI, covering more than 200 languages and possessing strong generalization ability in cross-lingual transfer (Fan et al., 2022).

The core reason for selecting these two models lies in their complementarity in multilingual environments: IndicTrans2_1B has obvious advantages in the fine-grained processing of Indic languages, while NLLB_3.3B is more robust in cross-lingual structure mapping and low-resource direction generalization. Their combination helps obtain more diverse and high-quality pseudo-parallel data under extremely low-resource conditions.

3.2 Direction-Specific Fine-Tuning

In the first phase of system construction, the above two models are respectively fine-tuned in one-to-one directions on the bilingual parallel corpora provided by WMT 2025, covering four translation directions: en→as, as→en, en→mni, and mni→en.

One-way translation fine-tuning at the granularity of translation directions enables the model to focus on learning the syntactic, lexical, and domain features of that direction. Compared with directly training a multilingual multi-directional model, it can avoid cross-direction interference and achieve higher convergence speed and better direction adaptability in low-resource scenarios. The training results of this phase serve as the baseline models for subsequent back-translation augmentation.

During the fine-tuning phase for NLLB_3.3B, LoRA (Low-Rank Adaptation) parameter-efficient fine-tuning technology is adopted, with specific configurations as follows (Hu et al., 2021):

- rank: 128
- alpha: 256
- dropout: 0.1
- Fine-tuning modules: All linear layers

LoRA injects low-rank matrix parameters into the model's linear layers, keeping most original parameters frozen and only updating a small number of trainable parameters, which significantly reduces memory usage and training costs while maintaining model performance. This design is particularly effective for models at the 3.3B scale, enabling high-quality directional fine-tuning to be completed under single-card or low-resource computing power conditions (Hu et al., 2021).

3.3 Monolingual Data Back-Translation Augmentation

The second phase introduces 100,000 English monolingual sentences with a domain proportion highly consistent with the test set to improve the model's adaptability in the target domain. This monolingual data is sourced from the NLLB open corpus, BPCC open-source data, and domain-specific web crawls. It is matched and filtered with test set samples using the all-mpnet-base-v2 semantic similarity model to ensure the domain distribution proportion is consistent with the test set (65.29% healthcare, 23.56% entertainment and sports, 11.15% culture).

Based on the two fine-tuned models obtained in the first phase, dual-model back-translation is implemented, which is a widely used data augmentation technique in low-resource machine translation to generate pseudo-parallel corpora (Sennrich et al., 2016):

1. IndicTrans2_1B translates English monolingual sentences into the target language, generating pseudo-parallel corpus set D1;

2. NLLB_3.3B translates English monolingual sentences into the target language, generating pseudo-parallel corpus set D2.

D1 and D2 are respectively merged with the official parallel data to fine-tune IndicTrans2_1B and NLLB_3.3B again, forming the first-round augmented models. Taking "back-translation → merging → fine-tuning" as a cycle, the iteration is performed until the BLEU score of the development set no longer improves. In actual experiments, significant improvements can be achieved with two iterations, and the third iteration is difficult to bring additional benefits. Therefore, two iterations are finally adopted as the optimal solution.

This dual-model iterative back-translation augmentation method fully leverages the complementary advantages of the two pre-trained models in language modeling and cross-lingual generalization, significantly enriches the diversity and domain coverage of training data in low-resource directions, and thereby improves the translation performance of the final system (Sennrich et al., 2016).

4 Experimental Results and Analysis

Table 2 shows the BLEU score performance of different systems and data augmentation strategies in the four translation directions (en→as, en→mni, as→en, mni→en). The experiment compares the performance changes of IndicTrans2_1B and NLLB_3.3B under one-way fine-tuning on official data, different back-translation data augmentations, and dual-model iterative back-translation.

Strategy	**en→as**	**en→mni**
IndicTrans2-1B	16.33	10.28
+OFT-off	23.80	16.24
+OFT-off+BT-it	–	–
+OFT-off+BT-nllb	–	–
+OFT-off+BT-itnllb	25.92	24.34
NLLB-3.3B	17.04	15.01
+OFT-off	24.52	21.29
+OFT-off+BT-it	29.72	25.19
+OFT-off+BT-nllb	28.32	25.28
+OFT-off+BT-itnllb	30.61	27.71
+OFT-off+DBT (P2)	32.11	28.92
Strategy	**as→en**	**mni→en**
IndicTrans2-1B	29.20	34.74
+OFT-off	40.36	44.35
+OFT-off+BT-it	40.10	43.86
+OFT-off+BT-nllb	41.61	44.56
+OFT-off+BT-itnllb	–	–
NLLB-3.3B	30.88	30.77
+OFT-off	37.69	37.75
+OFT-off+BT-it	–	–
+OFT-off+BT-nllb	–	–
+OFT-off+BT-itnllb	–	–
+OFT-off+DBT (P2)	–	–

Table 2: BLEU Scores of Different Systems and Data Augmentation Strategies on WMT 2025 Development Set. Note: "–" indicates that this combination was not tested in this direction or the results were not included in the statistics. We use the following abbreviations: OFT denotes One-way Fine-Tuning, off denotes official data,BT denotes Back-Translation, it and nllb denote different back-translated datasets, LoRA denotes Low-Rank Adaptation, and DBT denotes Dual Back-Translation.

4.1 Significant Gains from One-Way Fine-Tuning

After one-way (one-to-one) fine-tuning using official parallel corpora, both baseline models show significant improvements in BLEU scores across all tested translation directions:

• IndicTrans2_1B increases from 29.20 to 40.36 (+11.16 BLEU) in the as→en direction, and from 34.74 to 44.35 (+9.61 BLEU) in the mni→en direction;

• NLLB_3.3B (LoRA fine-tuning) rises from 17.04 to 24.52 (+7.49 BLEU) in the en→as direction, and from 15.01 to 21.29 (+6.28 BLEU) in the en→mni direction.

The above results demonstrate that one-way fine-tuning can effectively reduce multi-task interference and improve translation quality in specific directions under low-resource conditions.

4.2 Single-Model Back-Translation Augmentation Effect

When introducing the first round of back-translation augmentation, model performance continues to improve, but the effect depends on the source of back-translated data:

• IndicTrans2_1B: Using back-translated data from NLLB_3.3B (41.61 BLEU in the as→en direction) is superior to using its own back-translated data (40.10 BLEU); a slight gain is also maintained in the mni→en direction (44.56 vs. 43.86);

• NLLB_3.3B (LoRA): Using back-translated data from IndicTrans2_1B (29.72 BLEU in the en→as direction) is better than self-back-translated data (28.32 BLEU); the performance in the en→mni direction is close (25.19 vs. 25.28).

This indicates that pseudo-parallel data generated across models has complementarity in syntactic and lexical distributions, which can reduce noise accumulation in self-back-translation.

4.3 Dual-Model Back-Translation and Iterative Optimization

After adding the back-translated data of both models to the training simultaneously (dual-model back-translation), NLLB_3.3B (LoRA) achieves a BLEU score of 30.61 in the en→as direction and 27.71 in the en→mni direction; further performing the second round of dual back-translation iteration results in 32.11 BLEU in the en→as direction (+1.50 compared to the previous stage) and 28.92 BLEU in the en→mni direction (+1.21). The results show that multiple rounds of back-translation can bring additional benefits, but the marginal gain diminishes.

4.4 Value of LoRA Fine-Tuning

Considering the 3.3B parameter size of NLLB_3.3B, this study adopts LoRA (rank=128, alpha=256, dropout=0.1, injected into fully connected layers) for efficient one-way fine-tuning. Under the premise of low memory usage, a significant BLEU improvement is still achieved, making multi-stage data augmentation possible under limited computing power conditions (Hu et al., 2021).

5 Conclusion

This study addresses the low-resource translation task by combining two complementary multilingual pre-trained models, IndicTrans2_1B and NLLB_3.3B, and proposes a system construction method of one-way fine-tuning for specific translation directions and dual-model iterative back-translation augmentation. The introduction of LoRA parameter-efficient fine-tuning technology on NLLB_3.3B significantly reduces memory and computational costs, enabling multi-stage data augmentation under limited computing power conditions.

Experimental results show that:

1. One-way fine-tuning can significantly improve BLEU scores in low-resource translation directions (up to +11.16 BLEU), effectively reducing multilingual multi-directional interference;

2. Cross-model back-translation data augmentation is superior to single-model self-back-translation, proving that pseudo-parallel data generated by different models has complementarity in syntactic and lexical distributions;

3. Dual-model back-translation + multi-round iteration can further improve model performance, although the gain tends to converge after the second round;

4. LoRA technology balances efficiency and effectiveness in the directional fine-tuning of ultra-large-scale models, enabling the performance of low-resource translation directions to approach the improvement range of full fine-tuning (Hu et al., 2021).

Overall, the system method in this study fully leverages the complementary advantages of the two pre-trained models, combines parameter-efficient fine-tuning and dual-model iterative back-translation, and achieves significant BLEU improvements in the WMT 2025 low-resource task, providing a feasible and efficient reference scheme for the construction of low-resource machine translation systems. Future work will further explore adaptive back-translation data screening for multi-model collaboration and the introduction of multi-modal auxiliary information in low-resource scenarios to break through performance bottlenecks.

6 Future Work

On the basis of improving the low-resource translation performance achieved in this study, future work will continue to expand in the following two directions:

6.1 In-depth Utilization of Monolingual Data

Although parallel corpora for low-resource languages are limited, monolingual texts are often relatively abundant. Future work will consider:

1. Continual Monolingual Pretraining: Conducting continuous training on existing models (such as IndicTrans2_1B, NLLB_3.3B) using a large amount of Indic monolingual data to improve language fluency and localized expression ability;

2. Denoising Self-Supervised Training: Drawing on methods such as mBART and MASS, enabling the model to better grasp contextual dependencies and syntactic structures through tasks such as Masked Span Prediction and Noising & Reconstruction;

3. Combining Monolingual Back-Translation and Forward Translation: Constructing bidirectional pseudo-parallel data by combining monolingual data, that is, adding forward translation data generated from the target language to the source language on the basis of back-translation, to further improve the model's generalization ability.

6.2 Application of Large Language Models in Translation

With the development of multilingual Large Language Model (LLM) capabilities, introducing them into low-resource translation tasks has potential. Future work will consider:

1. LLM-as-Translator: Using general-purpose LLMs (such as Qwen, LLaMA, Mixtral, mT5) for direct translation or back-translation to generate higher-quality pseudo-parallel data that is more contextually appropriate;

2. Parameter-Efficient Fine-Tuning (PEFT) for Small Languages: Quickly adapting LLMs to specific small languages and domains through methods such as LoRA, Prefix Tuning, and Adapters, reducing computational costs while improving performance in low-resource scenarios;

3. Multi-Task Learning and Instruction-Tuning: Simultaneously training tasks such as translation, question answering, and paraphrasing on LLMs, and improving their ability to understand and generate low-resource languages through multi-task transfer effects.

References

Angela Fan and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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. *arXiv preprint arXiv:2106.09685*.

Anoop Kunchukuttan and 1 others. 2023. Indictrans2: Towards high-quality and low-resource machine translation for indic languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.