DELAB-IIITM WMT25: Enhancing Low-Resource Machine Translation for Manipuri and Assamese

Dingku Singh Oinam and Navanath Saharia
India Institute of Information Technology, Manipur
dingkuoinam@ieee.org

Abstract

This paper describes DELAB-IIITM's submission system for the WMT25 machine translation shared task. We participated in two subtasks of the Indic Translation Task, en↔as and en↔mn i.e., Assamese (Indo-Aryan language) and Manipuri (Tibeto-Burman language) with a total of six translation directions, including mn→en, mn←en, en→as, en←as, mn→as, mn←as. Our fine-tuning process aims to leverage the pretrained multilingual NLLB-200-Distilled-600M model, a machine translation model developed by Meta AI as part of the No Language Left Behind (NLLB) project, through two main developments: Synthetic parallel corpus creation and Strategic Fine-tuning. The Fine-tuning process involves strict data cleaning protocols, Adafactor optimizer with low learning rate (2e-5), 2 training epochs, train-test data splits to prevent overfitting, and Seq2SeqTrainer framework. The official test data was used to generate the target language with our fine-tuned model. Experimental results show that our method improves the BLEU scores for translation of these two language pairs. These findings confirm that back-translation remains challenging, largely due to morphological complexity and limited data availability.

1 Introduction

Meiteilon (Manipuri) is a Tibeto-Burman language spoken primarily in Manipur, while Assamese is an Indo-Aryan language spoken mainly in Assam. Both Manipuri and Assamese are recognized as official languages of India. There is a severe lack of parallel corpora and standardized digital resources. The data scarcity hinders the development of robust neural machine translation (NMT) models, as they typically require large-scale bilingual datasets for training (Sennrich and Zhang, 2019). The morphological complexity and syntactic diversity of Tibeto-Burman languages such as Meiteilon pose major challenges for MT systems, especially within low-resource scenarios (Singh and Singh, 2022b). The neglect of low-resource languages in machine translation is exacerbated by the overwhelming focus on high-resource languages, but this problem can be mitigated through transfer learning from massively multilingual pre-trained models such as mBERT (Conneau et al., 2020) and NLLB (Team et al., 2022). For such languages, MT systems risk perpetuating datasets and language-specific architectures (Joshi et al., 2021). Despite the challenges, researchers are actively working on improving MT for low-resource languages such as Manipuri and Assamese through various techniques like transfer learning, multilingual models, and back-translation (Singh and Singh, 2022b; Wei et al., 2023; Singh and Singh, 2022a). Recent WMT Shared Tasks on Low-Resource Indic Languages Translation have been significantly advancing in the field (Pal et al., 2023; Pakray et al., 2024, 2025).

This paper describes the fine-tuning of a pretrained multilingual NLLB-200-Distilled-600M model for translating Manipuri to English, English to Assamese, Manipuri to Assamese and back-translation. Back-translation, here, refers to the translation in which the translation direction is opposite to which the model is trained to perform the translation task.

The layout of the subsequent paper is as follows. Section 2 highlights some of the related works. Section 3 describes the implementation of the proposed translation systems. Finally, the conclusion and future work is drawn in Section 4.

2 Related Works

Transformer-based models have formed the backbone of many modern machine translation systems for low-resource Indic languages (Pal et al., 2023). These architectures, often enhanced with monolingual pre-training, language-specific fine-

Proceedings of the Tenth Conference on Machine Translation, pages 1222–1226

tuning, and inference-time strategies like kNN-MT, have demonstrated notable improvements in translation quality. For instance, (Ju et al., 2024) observed that such enhancements consistently improved BLEU scores, reinforcing the value of augmenting training with back-translation and model averaging techniques. mBART (Chipman et al., 2022) and mBART-large-50 (Tang et al., 2020) are used in multilingual setups, where fine-tuning on filtered corpora, using semantic tools like LaBSE (Feng et al., 2020) embeddings, showed limited gains due to poor back-translation quality and the morphological complexity of the target languages (M et al., 2024). IndicBART (Dabre et al., 2021) is a pre-trained BART model for Indic languages, specifically trained for Assamese, Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Kannada, Malayalam, Tamil, Telugu and English. Recently transformer-based models specialized for machine translation of Indic languages like IndicTrans (Ramesh et al., 2022) and IndicTrans2 (Gala et al., 2023) are available, which are trained on largest available Indic language parallel corpora namely Samanantar and BPCC respectively. IndicTrans model was trained for 11 Indic languages whereas IndicTrans2 was trained for all the 22 scheduled Indian languages. NLLB (Costa-jussa et al., 2022), a massively multilingual machine translation model has proven to be a breakthrough in the high-quality translation of around 200 languages across the world. MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), is a multilingual Language Model specifically built for Indic languages supporting around 17 languages. MuRIL outperforms multilingual BERT on all NLP tasks.

Recent advances in low-resource machine translation for Indic languages were explored in the WMT 2024 shared task (Pakray et al., 2024).

3 Method

We participate in two sub-tasks en↔as and en↔mn with a total of six translation directions, including mn→en, mn←en, en→as, en←as, mn→as, mn←as. We generate synthetic parallel data using the pretrained model NLLB-200-Distilled-600M. The proposed technique includes data preparation, pre-training, fine-tuning, and model evaluation to develop the machine translation systems.

3.1 Data Preparation

We used mn↔en (23,688 sentences), en↔as (54,000 sentences) parallel data provided by WMT25 (Kakum et al., 2023; Pakray et al., 2024; Pal et al., 2023). Since the organizers did not provide bilingual parallel data for mn↔as, we generate synthetic parallel data by translating to the target-language using the pretrained model NLLB-200-Distilled-600M. Specifically, we used the Manipuri (mn) side from the bilingual data (mn↔en) and Assamese (as) side from the bilingual data (en↔as) to generate target language data i.e., Assamese (as) and Manipuri (mn) respectively. Both the synthetic parallel data is then combined to get a total of 77,688 sentences. After removing empty sentences, we finally have 77,571 sentences.

For the translation directions that include English, we used English side from both mn↔en, en↔as parallel data provided by WMT25 to generate the target language data i.e., as and mn respectively. Combining the synthetic parallel data we get 77,688 sentences and after removing empty sentences, we have 77,681 sentences.

Language Pair	Sentences
mn↔as	77,581
mn↔en	77,681
en↔as	77,681

Table 1: No. of sentences for each language pair

Lang.	Token	Unique Token	Avg. word length
mn	1,614,626	93,910	5.70
en	1,272,380	46,404	4.66
as	1,131,164	85,098	5.10

Table 2: Token statistics for each language corpus

3.2 Pre-training

Starting with NLLB-200-Distilled-600M, a pretrained multilingual model as the base architecture. We perform additional training with the synthetic data, adapting the model's parameters to each specific language pair.

3.3 Fine-Tuning

The AutoTokenizer from the NLLB-200-Distilled-600M model was used to tokenize the inputs. We took the training data and fine-tuned it on NLLB-200-Distilled-600M for the translation settings from Manipuri to Assamese, Manipuri to English

and English to Assamese. To train (fine-tune) the NLLB-200-Distilled-600M model, 2 epochs with a learning rate of 2e-5 is set. The 2 epochs will help the model to pass through the entire training dataset 2 times and the learning rate (2e-5) is used to specialize the translation and retain its general knowledge. Using the same training parameters, we trained three fine-tuned models: Manipuri-English, English-Assamese, and Manipuri-Assamese model.

3.4 Model Evaluation

BLEU (Papineni et al., 2002) has been a standard and widely used metric for evaluating translation quality and ChrF (Popović, 2015) represents a promising metric for automatic evaluation of machine translation output. Table 3 shows the evaluation scores of our fine-tuned model while Table 4 shows the evaluation scores of the base NLLB-200-Distilled-600M model. The comparison shows that the fine-tuned model achieves better results than the base model in certain metrics.

		mn-as	mn-en	en-as
Translation	BLEU	45.6	70.3	54.1
	ChrF	37.4	3.1	76.1
Back-Translation	BLEU	41.1	8.0	27.1
	ChrF	29.0	3.1	55.0

Table 3: Evaluation scores for the Fine-tuned model

		mn-as	mn-en	en-as
Translation	BLEU	10.0	23.0	40.0
	ChrF	45.0	55.0	74.0
Back-Translation	BLEU	5.0	8.0	28.0
	ChrF	35.0	43.0	58.0

Table 4: Evaluation scores for the base NLLB-200-Distilled-600M model

Metric	mn→en	en→as
BLEU	7.346	16.105
METEOR	0.463	0.406
ROUGE-L	0.479	0.003
ChrF	48.783	55.702
TER	103.197	68.324

Table 5: WMT25 evaluation scores for normal translation direction

Table 5 and 6 give the WMT25 evaluation scores using the fine-tuned model for the translation and back-translation respectively, as released by the organizers.

Metric	mn←en	en←as
BLEU	3.151	15.020
METEOR	0.113	0.603
ROUGE-L	0.008	0.605
ChrF	37.512	59.374
TER	132.054	75.247

Table 6: WMT25 evaluation scores for back-translation direction

3.5 Model Output

We use the three fine-tuned models (Manipuri-English, English-Assamese, and Manipuri-Assamese) to perform translation and back-translation testing. The following results are observed.

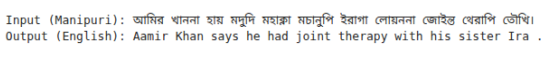

Figure 1: Translation mn→en

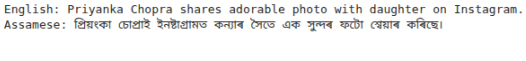

Figure 2: Translation en→as

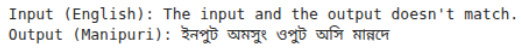

Figure 3: Back Translation en→mn

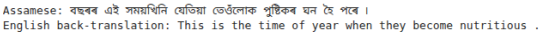

Figure 4: Back Translation as→en

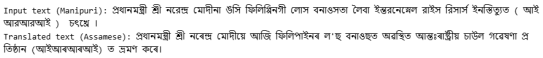

Figure 5: Translation mn→as

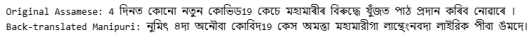

Figure 6: Back Translation as→mn

From the observation, Figure 1, 2, and 5 show that the fine-tuned model works well for translation. Similarly, Figure 3, 4, and 6 show test results for back-translations using the three fine-tuned models.

4 Conclusion and Future Work

In this paper, we describe low-resource Indic language translation shared task. We participated in two sub-tasks with a total of six translation directions. Experimental results show that our method improves over the base pretrained model. The fine-tuned model (mn-en, en-as) achieved BLEU (7.346) in mn→en while BLEU (16.105) is achieved in en→as. But for back translation (en→mn, en→as), the model achieved BLEU (3.151) and BLEU (15.020) respectively. From the Model Output Section, we can see how the model performs. Our experiment confirms that back translation for low resource languages still remains challenging due to the morphological complexity and data scarcity.

In future, we can explore semantic filtering techniques and ensemble NLLB with other pre-trained models.

References

Hugh A. Chipman, Edward I. George, Robert E. McCulloch, and Thomas S. Shively. 2022. mbart: Multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marta R. Costa-jussa, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Alahe Kalabassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1849–1863.

Fuliang Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Bing Pang. 2020. Language-agnostic bert sentence embedding.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 491–515.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the NLP world. 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 6282–6293.

Chenfu Ju, Junpeng Liu, Kaiyu Huang, and Degen Huang. 2024. Dlut-nlp machine translation systems for wmt24 low-resource indic language translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 742–746.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources English-Nyishi pair. *Sādhanā*, 48:237.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8199–8213.

Abhinav P. M, Ketaki Shetye, and Parameswari Krishnamurthy. 2024. Mtnlp-iiith: Machine translation for low-resource indic languages. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 751–755.

Partha Pakray, Rajen Chatterjee, Somnath Pal, Sunita S, Karthik Puranik, Shantipriya Parida, Satya Prakash, Sudipta Kar, Subhadarshi Panda, Md Hasan, Santanu Pal, and Ondrej Bojar. 2024. Findings of the WMT 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*.

Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, EMNLP 2025, Suzhou, China.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 682–694.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 145–162.

Rico Sennrich and Biao Zhang. 2019. Improving neural machine translation models with monolingual data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–10.

Salam Michael Singh and Thoudam Doren Singh. 2022a. An empirical study of low resource neural machine translation of manipuri in multilingual settings. *Neural Computing and Applications*, 34.

Salam Michael Singh and Thoudam Doren Singh. 2022b. Low resource machine translation of english manipuri: A semi supervised approch. *Expert systems with applications*, 209:118187.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3458–3473.

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, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Bin Wei, Jiawei Zhen, Zongyao Li, Zhanglin Wu, Daimeng Wei, Jiaxin Guo, Zhiqiang Rao, Shajun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie, and Hao Yang. 2023. Machine translation advancements of low-resource indian languages by transfer learning. *Huawei Translation Service Center*.