SPRING Lab IITM's submission to Low Resource Indic Language Translation Shared Task

Hamees Sayed, Advait Joglekar, Srinivasan Umesh
SPRING Lab,
Indian Institute of Technology Madras
hameessayed71@gmail.com, advaitjoglekar@gmail.com, umeshs@ee.iitm.ac.in

Abstract

We develop a robust translation model for four low-resource Indic languages: Khasi, Mizo, Manipuri, and Assamese. Our approach includes a comprehensive pipeline from data collection and preprocessing to training and evaluation, leveraging data from WMT task datasets, BPCC, PMIndia, and OpenLanguageData. To address the scarcity of bilingual data, we use back-translation techniques on monolingual datasets for Mizo and Khasi, significantly expanding our training corpus. We fine-tune the pre-trained NLLB 3.3B model for Assamese, Mizo, and Manipuri, achieving improved performance over the baseline. For Khasi, which is not supported by the NLLB model, we introduce special tokens and train the model on our Khasi corpus. Our training involves masked language modelling, followed by fine-tuning for English-to-Indic and Indic-to-English translations.

1 Introduction

Translation of low-resource languages poses significant challenges in natural language processing. While substantial progress has been made in developing machine translation models for high-resource languages, low-resource languages often suffer from a lack of parallel corpora and digital resources (Haddow et al., 2022). Languages like Khasi, Mizo, Manipuri, and Assamese are representative of this challenge, where limited data and unique linguistic complexities hinder the development of robust translation systems.

In recent years, efforts to bridge this gap have gained momentum, driven by initiatives such as the Bharat Parallel Corpus Collection[1] (BPCC) (Gala et al., 2023) and government-supported projects like PMIndia (Haddow and Kirefu, 2020), which aim to provide bilingual data for Indic languages. Despite these efforts, translation models for low-resource Indic languages have yet to achieve performance levels comparable to their high-resource counterparts (Suman et al., 2023), necessitating innovative approaches to model training and data utilization.

In this work, we develop a robust translation model for four low-resource Indic languages: Khasi, Mizo, Manipuri, and Assamese. Our approach involves data collection, preprocessing, training, and evaluation. We utilize datasets from WMT, BPCC, PMIndia, and OpenLanguageData[2] (Maillard et al., 2023), and enhance bilingual data through back-translation (Edunov et al., 2018) techniques, especially for Mizo and Khasi, significantly expanding our training corpus.

We follow Meta's data preprocessing standards and use LoRA (Low-Rank Adaptation) (Hu et al., 2021) fine-tuning on the NLLB (et al., 2022) 3.3B model to improve efficiency and performance with fewer parameters. Our model initially focuses on one-way translation from English to the Indic languages, then on reverse translations (Dabre et al., 2019). The results show improved performance over the baseline, particularly for Khasi, where we address gaps in pre-trained model support.

2 Dataset

In this study, we focus on four low-resource Indic languages covered in the Low Resource Indic Languages Shared Task: Khasi, Mizo, Manipuri, and Assamese. This section highlights the significance of each language, including their role in their respective regions, their linguistic and cultural importance, and the details of the datasets used. Statistics regarding language speakers are according to the 2011 Indian Census[3].

[1] https://ai4bharat.iitm.ac.in/bpcc/

[2] https://github.com/openlanguagedata/seed

[3] https://censusindia.gov.in/

Proceedings of the Ninth Conference on Machine Translation, pages 770–774
November 15-16, 2024 ©2024 Association for Computational Linguistics

Language	ISO-693-3	WMT Parallel	BPCC	PMIndia	OLD	Back-Translated	Total
Assamese	asm	50,000	35,354	9,732	0	0	95,086
Manipuri	mni	21,687	0	7,419	6,193	0	35,036
Khasi	kha	24,000	0	0	0	102,070	126,070
Mizo	lus	50,000	0	0	0	30,164	80,164

Table 1: Breakdown of data sources and volumes for each language. "OLD" refers to OpenLanguageData. The "Back-Translated" data was initially generated using Google Translate[4] for the first 500k characters from the monolingual WMT task data, with subsequent iterations increasing the data size using the trained model.

2.1 Languages

Assamese *(Asamiya)* is an Indo-Aryan language spoken primarily in the northeastern Indian state of Assam, where it serves as an official language and a regional lingua franca. With over 15 million native speakers, it is one of the most widely spoken languages in the region. Historically, Assamese was the court language of the Ahom kingdom. It is written in the Assamese script, an abugida system, known for its unique typographic ligatures.

Manipuri *(Meiteilon)* is a key Tibeto-Burman language spoken mainly in Manipur, India, where it is an official language and it is one of the constitutionally scheduled official languages of the Indian Republic. With 1.76 million speakers, it is the most widely spoken Tibeto-Burman language in India and holds the third place among the fastest-growing languages of India, following Hindi and Kashmiri. It is written in its own Meitei script as well as the Bengali script.

Khasi *(Ka Ktien Khasi)* is an Austroasiatic language primarily spoken by the Khasi people in Meghalaya, India, with approximately 1 million native speakers as of the 2011 census. The language holds an associate official status in certain districts of Meghalaya. Khasi is written in the Latin script. The closest relatives of Khasi are other languages in the Khasic group, such as Pnar and War.

Mizo *(Mizo ṭawng)* belonging to the Sino-Tibetan language family, is primarily spoken in the state of Mizoram, India, with around 800 thousand speakers. The Mizo language, also known as Lushai, has a rich oral history and was first written using the Latin script in the late 19th century. Mizo is recognized as the official language of Mizoram and is used in education, government, and media.

3 Methodology

This section covers the preprocessing steps and training methods used, including dataset preparation and the fine-tuning of Meta's multilingual NLLB 3.3B base pre-trained model. Detailed statistics on data distribution are presented in Table 1.

3.1 Preprocessing

In the preprocessing phase, we followed a series of steps to ensure the text data was clean and consistent before model training. We began by normalizing punctuation using Moses (Koehn et al., 2007), an open-source toolkit designed for preprocessing, training, and testing translation models. This step helps maintain consistency in text data, which is crucial for training robust models.

Non-printable characters, which often interfere with text processing, were replaced with a space. This choice ensures that any invisible or non-standard characters do not disrupt the tokenization process and ensures that the text is composed of standard printable characters.

We also applied Unicode normalization (NFKC) to transform characters into their canonical forms, making the text more uniform across different Unicode representations.

These preprocessing steps are aligned with those outlined by Meta for their multilingual models, and further details can be found on their GitHub[5]. This approach ensures that the text data used for training is clean, consistent, and compatible with the modelling requirements.

3.2 Training

For model training, we employed Meta's NLLB (No Language Left Behind) 3.3B parameter model,

[4] https://google.translate.com/

[5] https://github.com/facebookresearch/stopes/blob/main/stopes/pipelines/monolingual/monolingual_line_processor.py

a state-of-the-art multilingual machine translation model built to support over 200 languages, making it ideal for tasks involving low-resource languages (Tran et al., 2021; Yang et al., 2021). The NLLB 3.3B model is based on a Transformer (Vaswani et al., 2023) architecture with 3.3 billion parameters, featuring a dense encoder-decoder design. It includes the following hyperparameters:

Hyperparams	
embed size	2048
ffn size	8192
attn heads	16
enc/dec layers	48

Table 2: Hyperparameters for the baseline pre-trained model. 24 Encoder and 24 Decoder Layers.

To fine-tune the model, we employed LoRA, a technique that significantly reduces computational demands and training time by adapting only a small subset of the model's parameters. LoRA has been shown to match the performance of traditional fine-tuning methods while reducing the number of trainable parameters by a factor of 50 (Alves et al., 2023). This approach is especially effective for large-scale models like Meta's NLLB 3.3B, allowing efficient adaptation without significantly compromising on performance.

3.3 Parameters

The training process was conducted in three stages: first, the model was trained on masked language modelling (Devlin et al., 2019) to enhance its understanding of the target language by leveraging monolingual data. Next, it was fine-tuned for English-to-Indic translations, followed by further fine-tuning for Indic-to-English translations. In the case of Khasi, which was not natively supported by the NLLB model, special tokens were added to the tokenizer's vocabulary to accommodate the Khasi language. The model was subsequently trained on the Khasi corpus to ensure proper handling and integration of this language.

The training was performed across 4 Nvidia A6000 GPUs. These settings allowed us to optimize the model's performance while managing computational efficiency.

3.4 Inference

For inference, the trained adapter was loaded onto the NLLB 3.3B model. The model generated predictions using a beam search strategy with 10 beams and a repetition penalty of 2.5 to improve the diversity and quality of the translations. We experimented with various beam and penalty configurations, ultimately finding that this particular setup produced the most accurate and linguistically coherent outputs.

Training Args	
optimizer	adafactor
learning Rate	1e-5
epochs	8
precision	bf16
p_{mask}	0.15
peft type	lora
rank	128
lora alpha	256
lora dropout	0.1
target modules	all linear

Table 3: Training parameters and LoRA configuration used for fine-tuning the NLLB 3.3B model.

4 Results

The evaluation of our translation model across various language pairs and directions is shown in Table 4, with performance assessed using BLEU (Papineni et al., 2002), Translation Error Rate (Snover et al., 2006), RIBES (Isozaki et al., 2010), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015) metrics. We found that the scores in English-to-Manipuri and English-to-Mizo direction suffered from the poor quality of backtranslated data used in our training.

English-Assamese The model performed relatively well, with BLEU scores of 27.26 for English-to-Assamese and 26.69 for Assamese-to-English.

English-Manipuri The model showed lower BLEU scores for English-to-Manipuri (2.7) compared to Manipuri-to-English (20.88). The TER score was higher for English-to-Manipuri, reflecting greater translation errors in this direction.

English-Khasi For Khasi, the BLEU score was 12.12 for English-to-Khasi and 10.47 for Khasi-to-English.

English-Mizo The performance was mixed, with a BLEU score of 6.6 for English-to-Mizo and 18.49 for Mizo-to-English. The TER score indicates a higher error rate for English-to-Mizo, while the METEOR and ChrF scores were relatively balanced across both directions.

Language Pairs	Test Set	BLEU	TER	RIBES	METEOR	ChrF
English-Assamese	en_to_as_contrastive	27.26	52.79	0.3032	0.513	65.2
	as_to_en_contrastive	26.69	39.08	0.3308	0.7066	60.48
English-Manipuri	en_to_mn_contrastive	2.7	84.6	0.1185	0.1567	44.28
	mn_to_en_contrastive	20.88	48.77	0.3031	0.61	53.64
English-Khasi	en_to_kh_contrastive	12.12	63.31	0.1864	0.4453	44.55
	kh_to_en_contrastive	10.47	61.43	0.2172	0.5042	42.71
English-Mizo	en_to_mz_contrastive	6.6	66.06	0.1746	0.495	49.79
	mz_to_en_contrastive	18.49	53.19	0.2684	0.588	50.44

Table 4: Translation performance metrics of our MT System reported in the final evaluation.

5 Conclusion

In this work, we utilized Meta's NLLB 3.3B model, a large-scale multilingual transformer with 3.3 billion parameters, to enhance translation between low-resource Indic languages and English. The training process included masked language modelling, followed by English-to-Indic and Indic-to-English translations. Special tokens were added for Khasi, and LoRA (Low-Rank Adaptation) was employed to optimize computational efficiency and reduce training time.

Conducted on 4 NVIDIA A6000 GPUs, our approach demonstrates that large-scale multilingual models, when combined with LoRA, effectively capture diverse linguistic patterns and advance translation capabilities.

6 Limitations

In this study, we encountered several limitations that impacted the overall effectiveness of our translation system. One major challenge was the constrained size of our dataset due to computational resource limitations. The limited dataset size may have hindered the model's ability to generalize, particularly for low-resource languages where larger and more diverse datasets would have been advantageous.

Another issue we faced was the quality of back-translated data. The process of augmenting the dataset through machine translation often resulted in lower-quality data, which negatively influenced the model's performance. This highlights the need for more robust data generation techniques in future work.

We also observed a noticeable performance gap between translations where English was the target language and those where an Indic language was the target. This suggests that while the model may understand the morphological aspects of Indic languages, it struggles to generate accurate translations in these languages. This limitation underscores the need for further refinement in handling the complexities of Indic language generation.

Finally, the potential domain mismatch between our training data and real-world applications posed a significant challenge. The training data may not fully capture the linguistic and contextual nuances found in practical scenarios, leading to reduced performance in actual use cases. Addressing this issue in future work will be crucial for improving the model's real-world applicability.

References

Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. *Preprint*, arXiv:2310.13448.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *Preprint*, arXiv:1808.09381.

NLLB Team et al. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india. *Preprint*, arXiv:2001.09907.

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.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Mohana Krishna, Arnab Kumar Maji, Sandeep Kumar Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

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*, 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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Dhairya Suman, Atanu Mandal, Santanu Pal, and Sudip Naskar. 2023. IACS-LRILT: Machine translation for low-resource Indic languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 972–977, Singapore. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai wmt21 news translation task submission. *Preprint*, arXiv:2108.03265.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online. Association for Computational Linguistics.