MTNLP-IIITH: Machine Translation for Low-Resource Indic Languages

Abhinav P M*, **Ketaki Shetye***, **Parameswari Krishnamurthy**
Machine Translation and NLP Lab, LTRC
International Institute of Information Technology, Hyderabad, India
pmabhinav20@gmail.com, ketaki.shetye@research.iiit.ac.in, param.krishna@iiit.ac.in

Abstract

Machine Translation for low-resource languages poses significant challenges, primarily due to the limited availability of data.The WMT24 Low-Resource Indic Neural Machine Translation task challenges us to employ innovative techniques to improve machine translation for low-resource Indian languages. Our proposed solution leverages advancements in neural machine translation, focusing on methodologies such as back-translation and fine-tuning. By fine-tuning pretrained models like mBART, we achieved significant progress in translating languages such as Manipuri and Khasi. The best score was achieved for the English-to-Khasi (en-kh) primary model, with the highest BLEU score of 0.0492, chrF score of 0.3316, and METEOR score of 0.2589 (on scale of 0 to 1) and comparable scores for other language pairs.

1 Introduction

Machine translation is a sub-field of computational linguistics that focuses on developing systems capable of automatically translating text or speech from one language to another. The WMT24 task enables us to perform machine translation on those languages which are considered low-resource that is with limited data availability due to their lesser prevalence or documentation. Our work focuses on translating the 'En-X' language pair in both directions, where 'En' stands for English and 'X' includes Manipuri, a Tibeto-Burman language, and Khasi, which belongs to the Austroasiatic language family.

In recent years, Neural Machine Translation (NMT) has emerged as a powerful approach within machine translation, leveraging deep learning to achieve state-of-the-art results. Although, the NMT models being the data-hungry lead to peformance degradation when it comes to low resource languages. To tackle this problem, we employed fine-tuning and utilised the mBART (mbart-large-50-many-to-many-mmt) (Tang et al., 2020) experimenting with different configurations and settings for both preprocessing and training. mBART (Liu et al., 2020) is a multilingual sequence-to-sequence model trained on extensive monolingual datasets using a denoising autoencoder approach. It builds on the BART framework (Lewis et al., 2019) by combining a bidirectional encoder with a left-to-right autoregressive decoder, making it suitable for various translation tasks across multiple languages. Even if most of our final systems did not reach a satisfactory or competitive performance, we argue that our experiments brought up some interesting points that deserve more attention.

2 Related Work

In a comprehensive study, (Gaikwad et al., 2023) examined fine-tuning-based techniques to improve translation capabilities for low-resource languages by harnessing the multilingual IndicTrans2 model and achieved significant results.

In 2023, (Suman et al., 2023) utilized IndicBART (Dabre et al., 2022) and mBART-large-50, adapting them to specific language pairs and this method led to substantial performance gains for the Assamese and Manipuri languages.

Another 2023 study, (Jha et al., 2023), proposed and evaluated a multilingual neural machine translation system for Indian languages using the mT5 transformer. This system, trained on the modified Asian Language Treebank (ALT) dataset, demonstrated strong performance in translations between English, Hindi, and Bengali, achieving BLEU scores above 20 for five out of the six language pairs.

(Saini and Vidhyarthi, 2023) evaluated various pretrained models for English-to-Marathi translation, developing a bidirectional system. The

[0]* These authors contributed equally to this work.

Proceedings of the Ninth Conference on Machine Translation, pages 751–755
November 15-16, 2024

findings indicated that fine-tuning significantly enhanced the mBART model's performance.

(Signoroni and Rychly, 2023) addressed the challenges of neural machine translation (NMT) for low-resource language pairs by using supervised NMT systems. They experimented with different configurations and settings for both preprocessing and training, delving into the complexities of translating these languages.

3 Dataset

3.1 Languages

Manipuri, also known as Meitei or Meiteilon, is predominantly spoken in the northeastern Indian state of Manipur and is one of India's 22 scheduled languages having about 1.8 million native speakers (Signoroni and Rychly, 2023). It is distinguished by its rich morphological features, including a complex phonological system with tones, an agglutinative structure, and a Subject-Object-Verb (SOV) word order, as shown in Figure1. As a tonal language, Manipuri uses various tones and pronunciations to convey meaning and employs primarily two scripts: Meitei and Bengali. Despite being a scheduled language, Manipuri is often considered a low-resource language in natural language processing, presenting valuable opportunities for transfer learning and the development of multilingual models.

Khasi primarily spoken in the northeastern Indian state of Meghalaya by the Khasi people, is one of the major languages of the region spoken by over 1 million individuals (Signoroni and Rychly, 2023). Belonging to the Austroasiatic language family, Khasi is more commonly written using the Latin alphabet. Structurally, Khasi typically follows a subject-verb-object (SVO) order, similar to English, but differs from most Indian languages, which generally use a subject-object-verb (SOV) order, as shown in Figure 1.

Language	Sentence
Manipuri (Subject-Object-Verb)	লুহুঙবা চাক চবা।
Khasi (Subject-Verb-Object)	U khynnah u bam ïa ka soh.

Figure 1: Translations of "The boy eats an apple" showing word order in Manipuri and Khasi.

3.2 Composition

In this study, we used WMT 2024 (Pal et al., 2023) (Kakum et al., 2023) to fine-tune which includes both bilingual and monolingual data. For the bilingual data, we used the language pairs English-Khasi (en ↔ kh) and English-Manipuri (en ↔ mn). The compositions of these datasets are presented in Table 1 and 2.

Lang. Pair	Train	Test	Validation	Monolingual
en ↔ kh	24,000	1000	1000	182,737
en ↔ mn	21,687	1000	1000	2,144,897

Table 1: Number of lines in the dataset for the language pairs used in the task. The Monolingual column refers to the size of the non-English side.

Lang. Pair	Type:Token Ratio	Avg. Sentence Length
en ↔ kh	0.019 (en) 0.0093 (kh)	30.41 tokens (en) 36.48 tokens (kh)
en ↔ mn	0.0573 (en) 0.0083 (mn)	18.02 tokens (en) 15.23 tokens(mn)

Table 2: Training Dataset Statistics for Language Pairs: Type-Token Ratio and Average Sentence Length

4 System Overview

4.1 Initial Fine-Tuning

We begin by fine-tuning the mBART model (mbart-large-50-many-to-many-mmt) for the language pairs: English to Khasi (en → kh), Khasi to English (kh → en), English to Manipuri (en → mn), and Manipuri to English (mn → en). To ensure the quality and consistency of the bilingual data, we perform several preprocessing steps, including the removal of HTML tags, invisible characters, newline tabs, and duplicate entries.

For machine translation preparation, we tokenize both the input and the target texts. Truncating techniques are applied to standardize the texts by setting the maximum length of the tokenized sequences to 512 tokens. This ensures uniformity across all the examples in the dataset. This serves as our baseline model.

4.2 Data Augmentation

4.2.1 Backtranslation

A back-translation strategy (Sennrich et al., 2016) is employed to augment the training dataset with more data. Specifically, we back-translate 100,000 monolingual Khasi and Manipuri sentences into English using the baseline model. However, it is likely that the backtranslated data contains a significant portion of low-quality translations. To remove

Lang. Pair	Filtered Data
kh ↔ en	534
mn ↔ en	662

Table 3: Count of high-quality sentence pairs after cosine similarity filtering (threshold 0.84) for Khasi-English (kh ↔ en) and Manipuri-English (mn ↔ en).

low-quality data and ensure high-quality translation pairs, we employ a filtering process using the LaBSE model and cosine similarity.

4.2.2 Data Filtering

LaBSE Fine-tuning The Language-agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022) is not originally trained in the Khasi or Manipuri languages. Therefore, to generate accurate sentence embeddings for these language pairs, we fine-tune the LaBSE model specifically for en ↔ kh and en ↔ mn pairs, despite the limited size of available bilingual data. This fine-tuned model is then employed to produce sentence embeddings for the back-translated Khasi-English and Manipuri-English pairs.

Embedding and Similarity Calculation To ensure the accuracy of the back-translated data, we use cosine similarity, a metric that measures the cosine of the angle between two vectors in multidimensional space, to compare sentence embeddings. We apply a threshold of 0.84, effectively filtering out low-quality translations, and retaining only those pairs that meet our quality standards. Consequently, only a small portion of the original 100,000 back-translated sentences remain after filtering using this threshold. The data retained after filtering are presented in Table 3.

4.2.3 Further Filtering and cleaning

Despite filtering, some sentences with continuous symbols or non-English characters remain. To address this, we conduct an additional data cleaning round, removing sentences with continuous symbols and residual Manipuri or Khasi words in the English translations. The cleaned data is then combined with the original training set to create the augmented dataset as the final dataset.

4.3 Training with Augmented Data

Subsequently, we fine-tune the mBART model (mbart-large-50-many-to-many-mmt) using the augmented dataset, which includes both the original training data and the filtered back-translated data. The same data-preprocessing steps are employed for the augmented dataset as applied for the baseline model to maintain uniformity. The fine-tuning process incorporates this augmented data to enhance the model's performance and robustness.

5 Results and Analysis

Table 4 shows WMT24 evaluation results, highlighting a more challenging test set compared to last year. The low average semantic similarity score of 0.0253 as found using the (Reimers and Gurevych, 2019) sentence-transformer model that maps sentences to a 384 dimensional dense vector space to calculate semantic similarity between train and test data indicating reduced model performance too.

Among the primary models, the English-to-Khasi (en → kh) model, trained on both original and filtered back-translated data, performed the best across most metrics. It achieved the highest BLEU score of 0.0492, a chrF score of 0.3316, and a METEOR score of 0.2589, indicating strong performance for this language pair. The high chrF score shows effective capture of character-level nuances, while the lowest TER score of 84.79 reflects fewer required edits to match reference translations.

It is important to note that the Khasi-to-English (kh-en) primary model is excluded from this evaluation because of issues encountered during the evaluation process. Meanwhile, the English-to-Manipuri (en → mn) model shows a BLEU score of 0, highlighting the difficulty in translating from English to Manipuri. This may be partly due to the smaller size of the training data for this pair compared to the en-kh pair. Despite some fluctuations, the overall performance of both models is comparable in the test data.

When considering all metrics, the primary model shows slight improvements over the baseline model, indicating that the additional filtered back-translated data enhanced translation quality. However, the baseline model also performs competitively. The filtered back-translated data includes only 534 sentences for the kh ↔ en pair and 662 for the mn ↔ en pair (Table 3), with the small dataset likely due to stringent filtering.

Table 5 shows that the primary model slightly outperforms the baseline model in BLEU, chrF, and TER metrics for the English-to-Khasi (en → kh) and English-to-Manipuri (en → mn) models, sug-

Lang. Pair	BLEU (↑)	chrF (↑)	TER (↓)	METEOR (↑)	RIBES (↑)
Baseline					
en-kh	0.0359	0.2333	103.49	0.1649	0.1106
kh-en	0.0060	0.1731	106.60	0.1020	0.0487
en-mn	0.0064	0.3191	96.46	0.0724	0.0628
mn-en	0.0484	0.2662	101.76	0.1940	0.1087
Primary					
en-kh	0.0492	0.3316	84.79	0.2589	0.1595
en-mn	0.0000	0.3325	94.77	0.0822	0.0737
mn-en	0.0362	0.2777	94.79	0.1873	0.1136

Table 4: Results of Primary and Baseline models evaluated on WMT24 evaluation test data (scores calculated on the scale from 0 to 1)

Lang. Pair	BLEU (↑)	chrF (↑)	TER (↓)
Baseline			
en-kh	0.1748	0.3964	0.75699
kh-en	0.1274	0.3566	0.8791
en-mn	0.2089	0.5676	0.6537
mn-en	0.3265	0.5709	0.6522
Primary			
en-kh	0.1867	0.4126	0.7275
kh-en	0.1234	0.3570	0.8845
en-mn	0.2097	0.5726	0.6495
mn-en	0.3224	0.5698	0.6483

Table 5: Results of Baseline and Primary models evaluated on WMT23 validation data (scores calculated on the scale from 0 to 1)

gesting that filtered back-translated data improves translation quality. However, for the Khasi-to-English (kh → en) and Manipuri-to-English (mn → en) models, the primary model experiences a slight performance drop. This decrease is likely due to the filtered back-translated English sentences lacking coherence and contextual appropriateness, which affects the model's effectiveness. Despite these variations, the primary model still performs slightly better than the baseline model when considering all metrics.

6 Conclusion

Improving machine translation for low-resource languages remains a critical focus in the field. In this paper, we develop a system for translating low-resource Indic languages, specifically Manipuri and Khasi, in both English-to-Indic and Indic-to-English language pairs. We use back-translation and then apply cosine similarity for data filtering. While effective, their success depends on the quality of the back-translation and fine-tuned LaBSE models.

The morphological complexity of the Indic languages along with inability of capturing cultural and context specific meanings also poses as a challenge which the model could not solve. We further encounter challenges including data scarcity and high computational requirements which we believe can help produce better results.

7 Future Work

For future work, we aim to focus on enhancing machine translation for low-resource languages by leveraging language-specific properties such as part-of-speech (POS) tags and dependency parsing. By integrating POS tagging one can enable the model to better understand the syntactic roles of words, leading to more accurate and contextually appropriate translations. Dependency parsing can also capture the grammatical structure and relationships between words, allowing the model to manage complex sentence structures more effectively. Additionally, the use of more filtered backtranslated data can provide a richer training dataset, further improving translation quality. Combining these linguistic techniques with extensive backtranslation, so that we can capture the nuances of the individual low-resource languages, we can significantly address the current challenges in machine translation.

References

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic

natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Pranav Gaikwad, Meet Doshi, Sourabh Deoghare, and Pushpak Bhattacharyya. 2023. Machine translation advancements for low-resource Indian languages in WMT23: CFILT-IITB's effort for bridging the gap. In *Proceedings of the Eighth Conference on Machine Translation*, pages 950–953, Singapore. Association for Computational Linguistics.

Abhinav Jha, Hemprasad Yashwant Patil, Sumit Kumar Jindal, and Sardar M N Islam. 2023. Multilingual indian language neural machine translation system using mt5 transformer. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5.

N. Kakum, S.R. Laskar, K. Sambyo, and et al. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48:237.

Ivana Kvapilíková and Ondřej Bojar. 2023. Low-resource machine translation systems for Indic languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 954–958, Singapore. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

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, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Lekhraj Saini and Deepti Vidhyarthi. 2023. Bidirectional english-marathi translation using pretrained models: A comparative study of different pre-trained models. pages 1–8.

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, Berlin, Germany. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychly. 2023. MUNI-NLP systems for low-resource Indic machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 959–966, Singapore. Association for Computational Linguistics.

Dhairya Suman, Atanu Mandal2, Santanu Pal3, and Sudip Kumar Naskar. 2023. Iacs-lrilt: Machine translation for low-resource indic languages.

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.