

A Transformer-Based Multilingual Neural Machine Translation Framework with Shared–Language-Specific Decoding for Low-Resource Northeast Indian Languages

Sunita Sarkar¹, Sneha Das¹, Basab Nath^{2,*}, Somnath Mukhopadhyay¹

¹ Assam University, Computer Science and Engineering,
Dargakona, Silchar, Assam,
India

³ Bennett University, School of Computer Science Engineering and Technology,
Greater Noida, Uttar Pradesh,
India

{sarkarsunita2601, das.sneha2108, basabnath}@gmail.com, som.cse@live.com

Abstract. Multilingual Neural Machine Translation (MNMT) recently emerges as a powerful paradigm for tackling the task of translating among multiple language pairs under one model framework. Even so, it can be difficult to construct high-quality MNMT systems for low-resource languages (particularly in linguistically diverse regions) because of the scarcity of parallel bilingual corpora. In this paper, we introduce a framework for MNMT based on transformer for the ecosystem of North-East Indian language. The shared encoder with language independent decoders is used in English to Indic translation directions and vice versa utilises the language independent encoder with language specific decoder. Moreover, to mitigate the lack of parallel training data, back-translation is also employed to provide another source of synthetic parallel sentence pairs which enhances the training corpus of our model. The system was evaluated using standard translation metrics, including BLEU (Bilingual Evaluation Understudy), ROUGE, METEOR, and TER (Translation Edit Rate). Results demonstrate promising performance, with BLEU scores of 0.29 for English–Assamese and 0.26 for English–Nepali, with 95% confidence intervals confirming the robustness of these gains over other language pairs. These findings highlight the potential of our MNMT model in advancing translation quality for low-resource Northeast Indian languages and provide a foundation for further improvements using larger datasets and pre-trained multilingual models.

Keywords. MNMT, NMT, transformer, shared encoder, shared decoder.

1 Introduction

For natural language processing (NLP), Machine Translation (MT) is considered one of the most important driving forces for building a global communication bridge, following some breakthroughs in deep learning, especially neural machine translation (NMT). Neural Machine Translation (NMT) [1] is a type of machine translation which has been gaining popularity, especially after the introduction of sequence-to-sequence architectures with attention mechanisms outperformed earlier approaches to translation [2, 3, 4]. However, in addition to many low-resource languages being difficult Or costly to produce large-scale parallel corpus for [5], NMT of lower quality than that used for high-performance systems is also common. One approach for tackling this challenge is Multilingual Neural Machine Translation (MNMT), which adds a new dimension to NMT, allowing a single model to perform translations in multiple language pairs. While we are training MNMT up to October 2023, it produces better results by leveraging common linguistic representations and knowledge transfer between related languages — especially in the absence of enough monolingual data for one or more language pairs [6].

However, multilingual NMT for Northeast Indian languages is still under-studied. Currently existing

multilingual systems such as mT5 and IndicTrans2 have been trained only on the most widely spoken Indian languages, whereas neither Khasi, Bodo nor Manipuri (or in fact many other minority Indian languages) receives any coverage at all.

Furthermore, the complex morphology, tonal distinctions and diverse scripts of those languages only make translation even more complicated.

This gap points towards the pressing need for a specialized MNMT framework suitable to overcome both resource-poor landscape and linguistic rarity from North-Eastern Indian region.

To deal with these issues, we propose a novel multilingual neural machine translation framework targeting the low-resource languages of Northeast India.

The main contributions of this study are summarized as follows:

(1) Construction of an enhanced parallel corpus for North-East Indian languages by employing back-translation as a data augmentation strategy.

Synthetic sentence pairs generated through reverse translation increase both the size and linguistic diversity of the training data.

(2) Design and implementation of a transformer-based multilingual neural machine translation framework for low-resource language pairs in North-East India.

The proposed architecture employs a shared encoder with language-specific decoders for English-to-Indic translation and language-specific encoders with a shared decoder for Indic-to-English translation.

(3) Integration of family-aware encoding to exploit common linguistic characteristics among the Indo-Aryan, Tibeto-Burman, and Austroasiatic language families, while preserving language-specific features essential for accurate translation.

(4) Comprehensive performance assessment of the proposed MNMT system using multiple automatic evaluation metrics, including BLEU, SacreBLEU, ROUGE, TER, and METEOR.

1.1 Language Families and Linguistic Features of Indic Languages in Northeast India

Northeast India is one of the most linguistically diverse regions of the country, encompassing languages that belong to multiple language families [7]. Figure 1 presents the major language families found in the region along with their representative languages [8]. Unlike many other parts of India, the Northeast has a strong presence of Tibeto-Burman languages [9], including Manipuri (Bengali script), Bodo, and Mizo [10]. The region is also home to Austroasiatic languages [11], such as Khasi and Santali [12], which are comparatively uncommon in the rest of the country. In addition, Indo-Aryan languages [13], particularly Assamese and Nepali [14], constitute an important component of the linguistic landscape of Northeast India.

1.1.1 Indo-Aryan Languages: Assamese and Nepali

Assamese: Assamese is an Eastern Indo-Aryan language, mainly spoken in Assam. It shares many features of vocabulary with the neighbouring Indo-Aryan languages, especially Bengali and perhaps Odia. It is written in the Bengali-Assamese script, which developed from Eastern Brahmi and bears great resemblance to the writing system of Bengali. Assamese has a large number of vowel sounds and an extensive system of verb inflections, both very typical among Eastern Indo-Aryan languages.

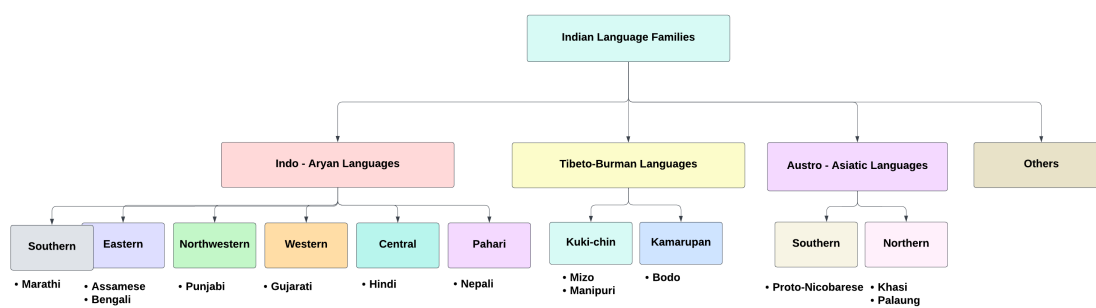
Nepali: Nepali is part of the Pahari subgroup of the Indo-Aryan family and is mostly spoken in Nepal, Sikkim and partially in West Bengal. It is written in the Devanagari script again, the same script in Hindi. Like other Pahari languages, Nepali has a complex verbal system with elaborate accentual and conjugational patterns.

1.1.2 Tibeto-Burman Languages: Bodo, Manipuri, and Mizo

— **Bodo:** Bodo is an Indo-Aryan language in the Dardic group of languages, more specifically belongs to the Bodo-Kachari subgroup of the Tibeto-Burman language family, primarily spoken in Assam. It

Table 1. Scripts and Example Sentences for Languages of Northeast India

Script	Language	Example Sentence
Bengali-Assamese	Assamese	কি কৰিছা? (ki korisa?) - What are you doing?
Devanagari	Nepali	तपाईं कसरी हुनुहुन्छ? (tapāĩ kasarī hunuhuncha?) - How are you?
Devanagari	Bodo	नों अबाय दं? (nō abay dō?) - Where are you?
Bengali	Manipuri	খুঁবৈ তোৰা লৈৰাকখি? (khunai tora lairakhi?) - Where do you live?
Latin	Mizo	Tan ga i kal le? (Where are you going?)
Latin	Khasi	Phi ia ka'n? (Who are you?)

**Fig. 1.** Related Indian language group

adheres to the grammatical structures typical of Tibeto-Burman languages, and uses a variant of the Devanagari script.

- **Manipuri (Bengali):** Manipuri—A key Tibeto-Burman tongue in Northeast India. Its characteristic is also a rich phonological system, which includes tone—n most cases in the Tibeto-Burman languages. Manipuri used the Bengali-Assamese script in this work.
- **Mizo:** Mizo mainly spoken in Mizoram is a member of the Kuki-Chin branch of Tibeto-Burman family. It shows linguistic features like tonal phonology and agglutinative word building — where several morphemes merge to create grammatical meaning. It is written in the Latin script and has similar features to Hmar, Paite, and other closely related languages.

1.1.3 Austroasiatic Languages: Khasi

- **Khasi:** Khasi is the main representative of the Austroasiatic family of languages in

Meghalaya, and belongs to the Mon-Khmer branch. Its script is written in Latin. Like some other Austroasiatic languages such as the neighbouring syllable+tone, Khasi is tonal, giving its phonology a unique character.

Table 1 provides an overview of the Northeast Indian languages considered in this study, their corresponding writing scripts, and representative example sentences.

2 Literature Review

The first neural machine translation(NMT) method with attention mechanism in encoder-decoder structure was proposed by Bahdanau et al. [2], which is the basis of modern NMT.

This was subsequently pioneered by Vaswani et al. The transformer model introduced by [3], which relied only on attention mechanism, thus discarding recurrent and convolutional networks (which are sequential), allowing for efficient parallelization. In the case of low-resource

languages, methods like unsupervised learning, data augmentation and usage of pre-trained models have resulted some significant advancements [4].

The challenges of NMT in low-resource contexts have been discussed in several studies. Laitonjam and Singh [15] examined NMT for the Manipuri-English pair while Sing and Mahmood [16] studied models already used in other languages like English-Manipuri, and English-Assamese to improve model quality with approaches like back translation and use of pre-trained models. To handle Manipuri's tough morphology, EM-FT in [17] outperformed the previous releases using LSTM with transfer learning, byte pair encoding for English-Khasi.

Jha et al. using mT5 transformer [18] created a multilingual model and achieved impressive BLEU scores for english, hindi and bengali respectively. Khaber et al. Attention-based LSTM and Transformer model comparison [19] showed that the transformer architecture works better than the equivalent LSTM models across various language pairs. Mujadia et al. Wang et al. [?], Zhang et al. [20] explore cross-lingual transfer and fine-tuning from generative large language models (LLMs) between English and a total of 22 Indian languages; Similarly, Bala Das et al. Das et al. [21] focused on multilingual NMT over 11 Indic languages via the Samanantar corpus, and showed that grouping languages for training provides larger gains than layering even with transliteration enhancement in place when translating between different language scripts.

There have been other work on improving the architecture or the linguistic aspects of Indic NMT. Das et al. With morpheme based segmentation, Das et al. [22] combined with ensemble methods achieved better translations for Hindi and Bengali languages; similarly, Patil et al. Later [23] showed that mT5 is a good model for multilingual cases. Gupta et al. Samanantar is now one of the largest publicly available parallel corpus on Indic languages introduced in [24] and hence built needs to be acknowledged.

Dabre et al. A recent survey of multilingual NMT was presented in [6] where they discussed how cross-lingual information can be utilized for

data-poor languages. More recently, Kong et al. A shared encoder and language-specific decoder architecture for MNMT was proposed as well in [25], enabling high performance with computational efficiency; Singh et al. More recently, [26] analyzed the use of transfer learning for Manipuri in low resource multilingual scenarios. Together, the studies established promise in MNMT for low-resource Indic languages but also brought out issues of linguistic diversity and scarcity of parallel data [8].

Most of the research in multilingual NMT has shown that while it is beneficial for morphologically rich languages, less-resourced languages like Khasi, Bodo and Mizo spoken in Northeast India are still under-explored because of limited parallel data myriads of morphology and so many scripts. Most state-of-the-art models have been focused on high resource languages like the major Indian languages (between more than 50%). This work reduces this gap by proposing a customized theta MNMT framework, which benefits from language family similarities and data augmentation to enhance translation accuracy.

3 Methodology

This paper presents a multilingual Neural Machine Translation (NMT) method based on the powerful transformer architecture for the diverse North-East Indian region linguistically. To that end, our framework is specifically designed to address the challenges of low-resource language pairs using a novel dual-architecture design to ensure maximal crosslingual transfer while respecting per-language characteristics.

We focus on the Tibeto-Burman languages of Manipuri, Bodo and Mizo, the Austroasiatic language Khasi, and two Indo-Aryan languages, Assamese and Nepali. Asymmetrical Translation Architecture: Which is likely the main novelty of our approach which treats translation directions differently depending on linguistic reasons:

1. **English-to-Indic Direction:** We use a single shared encoder with language specific decoders. This design uses the structural regularity of English inputs but

allows the decoders to learn the various morphological and syntactic behaviours of target Indic languages.

2. **Indic-to-English Direction:** We utilize language-specific encoders with a shared decoder. This allows the model to account for the unique grammatical and morphological characteristics of each Indic source language, while benefiting from the relatively simple morphology and fixed word order of English during decoding.

The proposed methodology has the following four steps:

1. **Data Augmentation:** we employ back-translation as the primary augmentation technique. This process generates synthetic parallel sentences from monolingual target data, with particular focus on the most resource-constrained languages such as Khasi and Mizo.

2. **Data Preprocessing:** We preprocess the corpus by normalizing the text to all lowercase letters, removing special characters, and so on.

3. **Developing an English to Indic Multilingual Translation Model:** We build a multilingual NMT model with a shared encoder and language-specific decoders. The encoder captures common patterns from English input, while the decoders handle the unique morphological and syntactic features of Assamese, Bodo, Khasi, Manipuri, Mizo, and Nepali.

4. **Developing an Indic to English Multilingual Translation Model:** For the reverse direction, we use language-specific encoders with a shared decoder. The encoders model the rich grammar and morphology of each Indic language, while the shared decoder exploits the uniform structure of English to generate fluent translations.

The following subsections describe the key methodologies and architectural components of the proposed multilingual neural machine translation framework developed for the linguistically diverse languages of North-East India, including the use of back-translation as a data augmentation strategy.

3.1 Back-Translation for Data Augmentation

A back-translation method is used which can massively increase the training corpus to tackle the challenges of insufficient data in low-resource language pairs. To cover up the gaps and better results, back translation is leveraged which creates synthetic parallel texts from monolingual data. This augmentation strategy focuses on the pairs of languages with fewer resources. Corpus analysis revealed Khasi and Mizo to have the fewest parallel data, with just 26,000 and 53,495 sentence pairs respectively compared to more than 140,000 pairs for languages like Assamese and Nepali. This imbalance motivated this work to use augmentation. Using the similarity in linguistic heritage and bibliography between Khasi and Mizo, a strong basis for data augmentation is built in this study. Khasi and Mizo are described as being from different families; however, the geographical and cultural contexts in which they exist overlap creating semantic domains.

The augmentation pipeline consists of three major stages, which are outlined below:

1. **Base Dataset Creation:** We then chose a frequent parallel dataset in Khasi and Mizo corpora that have an analogous semantic content and structural patterns in the sentences. This strengthened synthetic expansion on a stable base.
2. **Cross-linguistic Expansion:** Starting with a curated base dataset, we iteratively enriched the corpus to cover all target languages (English, Assamese, Bodo, Nepali and Manipuri), resulting in a multi-parallel corpus fit for back-translation.
3. **Synthetic Data Generation:** Back-translation was applied as follows:
 - (a) Initial translation models were trained on the available parallel data.
 - (b) High-quality monolingual target language data was selected from reputable sources IndicTrans2 [27], PMIndia [28], TDIL-DC [29] and Indic MT Task [30, 31] and preprocessed by normalizing

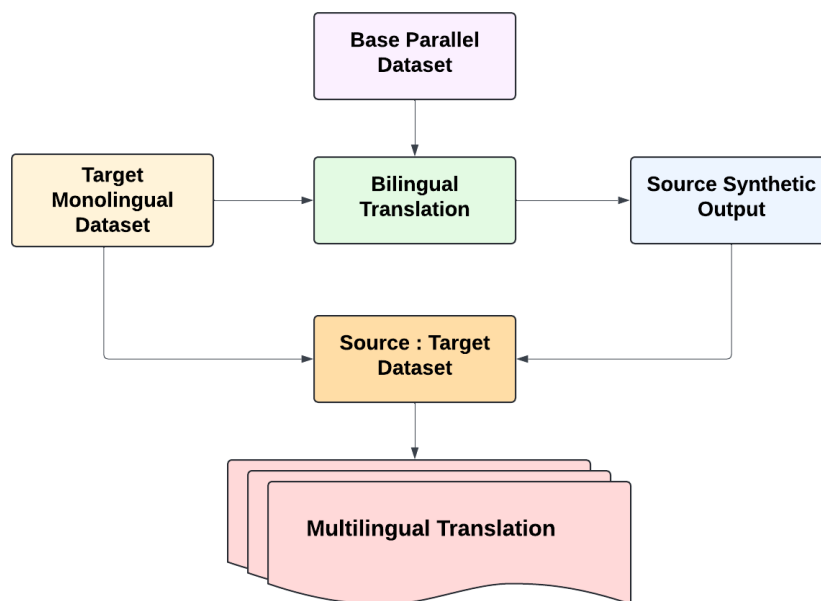


Fig. 2. Architectural overview of Backtranslation

text, removing noise, and verifying language consistency.

- (c) Monolingual sentences were translated back into English to generate synthetic parallel pairs.
- (d) Quality filtering, including sentence-length ratio checks and automatic language identification, was applied to remove noisy or misaligned pairs, ensuring reliability of the augmented data.

This process expanded the training corpus by approximately six percent across the language pairs. Assamese gained 8,473 new sentence pairs (6.4% increase), Bodo 8,986 (6.4%), Manipuri 2,886 (5.0%), and Nepali 9,296 (6.4%). Khasi and Mizo did not receive synthetic pairs due to the lack of sufficient high-quality monolingual resources.

The augmentation ratio of around six percent was chosen empirically: preliminary experiments with higher ratios (10–15%) degraded translation quality due to noise, whereas lower ratios (<5%) yielded negligible improvements.

Thus, the selected augmentation level provided the best balance between data expansion and reliability. Quality filtering, including sentence-length ratio checks and automatic language identification, was applied to ensure the authenticity of the synthetic pairs. In addition, a small sample of the synthetic pairs for each language was manually evaluated by native speakers, who confirmed that the majority of retained sentences were fluent and semantically consistent. Figure 2 shows an overview of our back-translation method.

3.2 English-to-Indic MNMT Architecture

A multi-lingual neural machine translation (MNMT) model consisting of a shared encoder and language-specific decoders for EN→I (English-to-Indic) task was trained. In this paper, the architecture uses 512 dimension & 8 attention heads which is an ideal balance for low resource language translation tasks as it allows to capture long-range dependencies without killing computational efficiency. This asymmetric architecture (shared encoder with

language-specific decoders) was chosen to leverage cross-lingual transfer for English source understanding while preserving the morphological complexity of diverse target languages through specialized decoders.

The architecture leverages the Transformer's self-attention mechanism for efficient sequence modeling. A dedicated shared encoder is assigned to each language family to capture common linguistic patterns within that family. The encoded representations are subsequently processed by language-specific decoders, which generate the target-language text while accounting for language-specific grammatical and lexical characteristics. For example, when translating the English sentence "The students are studying in the classroom." into Manipuri, the corresponding translation is "সেটুদেগেটসগি মইয়ামং গুজলদা লৱিঁ".

Here, the word "লেইৱিঁবামিসিংদাগী" represents a complex agglutinated form composed of multiple morphemes. The shared encoder learns common linguistic patterns across the Tibeto-Burman language family, whereas the Manipuri-specific decoder reconstructs the corresponding morphological structure in the target language while preserving its linguistic characteristics.

The shared encoder consists of six transformer layers with eight attention heads and 512-dimensional representations as in the configurations shown to work well for multilingual low-resource NMT scenarios [3]. Thus allowing a trade-off between model capacity and compute, providing sufficient expressive power to capture cross-lingual trends without overfitting on the very low resource languages. The shared encoder is trained on weight sharing, encouraging the subsequence of systematic similarities across all languages, rather than viewing each language independently. It records cognate inflectional patterns as well as postpositional structures for the Indo-Aryan languages, agglutinative morphology and evidentiality systems for the Tibeto-Burman languages, and isolating morphology with tonal features for the Austroasiatic languages. An additional family-aware design is used to mitigate the learnt mapping from one language to another and helps in accomplishing native representation

from diverse target languages while enabling efficient cross-lingual transfer.

3.3 Indic-to-English MNMT Architecture

Multi-Lingual NMT with Language-Specific Encoder and Shared Decoder For the multilingual NMT model, we share a decoder across all languages but use language-specific encoders. Now each encoder can specialize/generalize to its own language-specific grammar and word pattern while a shared decoder still takes advantage from some common patterns, and so keep the model (simpler) using what it learn. Every Indic language receives its own unique encoding layers and is desirable to learn lingual structure and morphological tag specific information that translate the input to a familiar encoded space. By only using this joint representation, the shared decoder produces coherent English translations at the entire model level. For the Assamese sentence তেওঁ বজাৰলৈ গ'ল, the source language follows a Subject–Object–Verb (SOV) word order, and the case marker লৈ denotes direction, both of which differ from English syntax.

The Assamese-specific encoder captures these language-specific features and maps the sentence into the shared representation. The shared decoder subsequently utilizes this representation to produce the corresponding English translation, "He/She went to the market," in the Subject–Verb–Object (SVO) word order.

For Indic→English, the architecture is inverted: language-specific encoders feed into a shared decoder. Each encoder has six transformer layers (512-dimensional, eight attention heads) and is tailored to the structural properties of its language. Unlike standard multilingual encoders that treat all languages uniformly, our language-specific encoders are designed to implicitly capture family-specific characteristics through training. The Indo-Aryan encoders focus on handling complex case markings and verb agreements that are common in languages like Bengali and Assamese. The Tibeto-Burman encoders specialize in processing the word-building patterns and grammatical markers found in languages like Manipuri and Bodo. The Austroasiatic encoder is

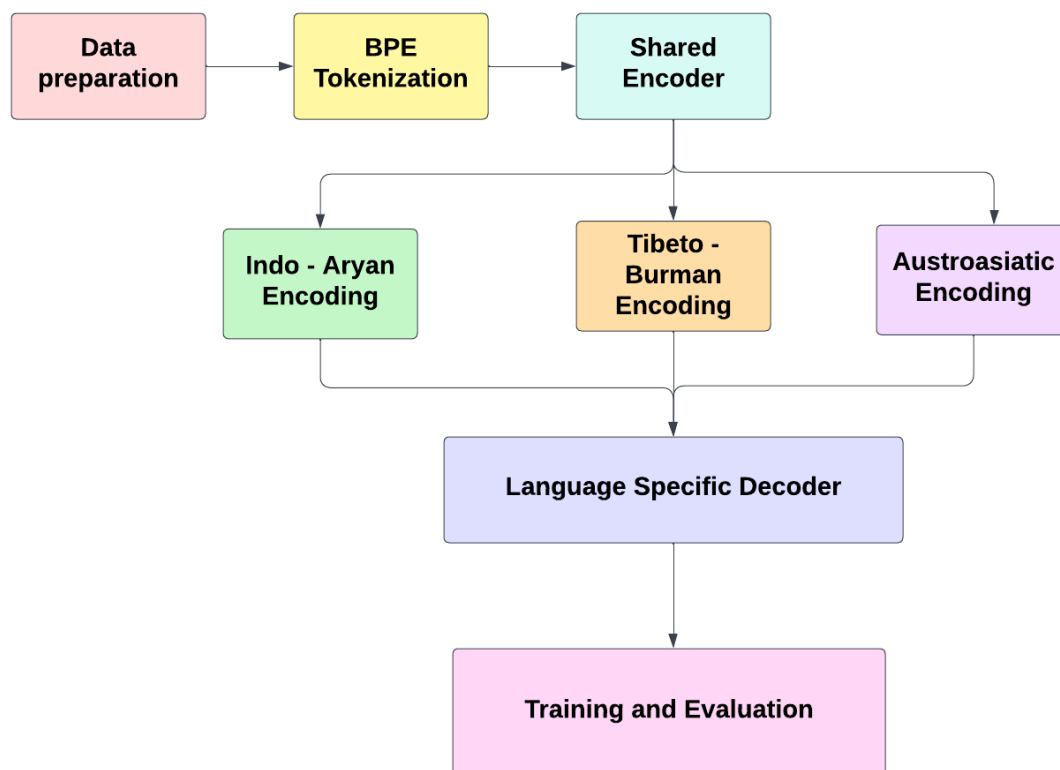


Fig. 3. Architectural Overview of the Multilingual NMT System for English to Indic

optimized for Khasi's simpler word structure but complex tonal system. This specialized design allows each encoder to capture the linguistic nuances of its source language family before converting to a standardized representation for the shared English decoder.

The NMT system we propose in this paper has a dual structure as shown in Figure 3 and Figure 4.

The transformer-based framework for English-to-Indic multilingual neural machine translation is illustrated by the architecture, overview in Figure 3. In addition to the shared encoder, the figure illustrates the data preprocessing pipeline and Byte Pair Encoding (BPE) tokenization process. Additionally, the architecture is layered into language family-specific layers to memorize grammatical structures for Indo-Aryan languages, morphology and idiomatic properties of Austroasiatic languages

and Tibeto-Burman languages. The figure also highlights the language-specific decoders, which are intentionally lighter (one transformer layer each, compared to six in the shared encoder). This asymmetry reduces computational complexity while still enabling the decoders to specialize in reconstructing target-side morphological and syntactic details.

The Indic-to-English translation architecture shown in Figure 4 was assigned its own language-specific encoder for each specific Indic source language. Since different languages have distinct structural and morphological features, these encoders exploit the principle of all input languages and encode them into one format. Then a separate decoder, which is just the lightweight structure of a single transformer layer (instead of six in the language specific encoders) translates those encoded representations into outputs in

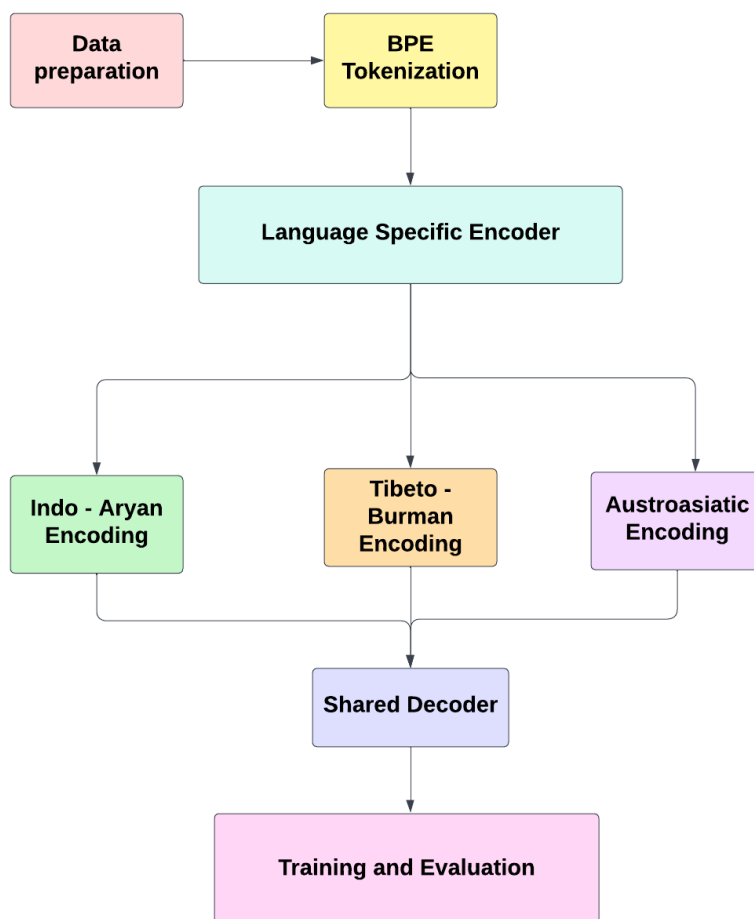


Fig. 4. Architectural Overview of the Multilingual NMT System for Indic to English

English. This asymmetry reduces training expense and computational complexity, at the cost of preserving the parallelism of coded objects that yields fluent English translations.

The use of language-specific encoders allow us to extract the aspects of language that make each Indic language specific and map the input into the same shared latent representation.

This unified representation is then passed through a shared decoder that produces fluent and accurate English translations, providing a simple yet effective way to reduce model size while maintaining translation accuracy for all supported languages.

4 Experimental Setup

4.1 Data Pre-Processing

The multilingual translation model was trained using parallel corpora collected from several publicly available resources, including IndicTrans2 [27], PMIndia [28], TDIL (Technology Development for Indian Languages–TDIL-DC) [29], and the Indic MT Task datasets [30, 31].

To alleviate the scarcity of parallel data, back-translation was employed as a data augmentation strategy, increasing both the size and diversity of the training corpus.

Table 2. Dataset Overview for English–Indic and Indic–English Translation Models

	Assamese	Bodo	Khasi	Manipuri	Mizo	Nepali
Original Parallel Corpus	1,32,753	1,40,779	26,000	57,336	53,495	1,45,641
Monolingual Data	150,000	140,000	20,000	60,000	15,000	160,000
Synthetic Corpus Generated	8,473	8,986	-	2,886	-	9,296
Total Corpus	141,226	149,765	26,000	60,222	53,495	154,937
Vocabulary count (English)	40,343	37,675	10,746	23,847	21,313	48,465
Vocabulary count (Target)	71,625	57,894	8,647	31,837	24,278	68,954
Training Set	120,042	127,300	22,100	51,188	45,470	131,697
Testing Set	14,122	14,976	2,600	6,022	5,349	15,493
Validation Set	7,062	7,489	1,300	3,012	2,676	7,747
Min Words/Sentence	1	1	1	1	1	1
Max Words/Sentence	22	18	27	19	25	20

During preprocessing, irrelevant characters were removed, all text was converted to lowercase, and punctuation symbols were retained. The resulting dataset was partitioned into training, validation, and test sets using an 85%, 5%, and 10% split, respectively, to facilitate reliable model training and comprehensive performance evaluation across the supported language pairs. As summarized in Table 1, the original parallel corpus size varies considerably across languages, with Assamese and Nepali having over 130,000 sentence pairs. Note that Khasi and Mizo are severely low-resource, with only 26,000 and 53,495 pairs, respectively. Back-translation contributed roughly a 6% increase in corpus size for Assamese, Bodo, Manipuri, and Nepali, while Khasi and Mizo did not receive synthetic augmentation due to insufficient monolingual resources of adequate quality. The table clearly separates the contributions of original parallel data and back-translated synthetic data, highlighting both the magnitude of augmentation and the persistent challenges of extremely low-resource languages. Table 2 shows an overview of the dataset used in this study for English–Indic and Indic–English language pairs.

4.2 Model Training

The Proposed multilingual model was trained to improve the translation quality whilst conserving computational resources. All hyperparameters

were selected based on a combination of pilot experiments and prior work in multilingual low-resource NMT. We experimented with embedding dimensions of 256, 512, and 768, feed-forward dimensions of 256 and 512, and learning rates of $1e-3$, $1e-4$, and $5e-5$. These values were explored using a manual grid search across the ranges, and final choices were selected based on validation performance. The final choices—embedding dimension 512, feed-forward dimension 256, and learning rate 0.0001—offered the best balance between convergence speed, model stability, and translation quality on validation sets. A value of 32 maintained a good balance between processing requirements and precision of the gradients. Also, a 0.2 dropout rate advanced generalization and training process was stopped after five epochs without the decrease of validation loss to reduce over fitting.

The maximum epoch count was fixed at 40, as pilot runs indicated that convergence typically occurred within 30–35 epochs, while longer training offered negligible improvements and increased the risk of overfitting. The computational experiments were conducted on a system equipped with 24 GB of RAM, a single-core hyper-threaded Xeon processor operating at 2.3 GHz, and an NVIDIA Tesla V100 GPU featuring 2,496 CUDA cores and 16 GB of VRAM. The hyperparameter settings used for training the multilingual neural machine translation (MNMT) models for both English-to-Indic and

Table 3. Training parameters for multilingual NMT models

Parameter	English→Indic	Indic→English
Encoder Layers	6	6
Decoder Layers	1	1
Batch Size	32	32
Feed-forward Dimension	256	256
Epochs	40	40
Learning Rate	0.0001	0.0001
Embedding Dimension	512	512
Dropout	0.2	0.2

Indic-to-English translation tasks are summarized in Table 3.

These training settings were tuned to ensure robust performance across highly diverse low-resource languages, while maintaining consistency and fairness in comparing both translation directions. We did not use automated hyperparameter optimization methods, as the manual grid search over a restricted range was sufficient given the dataset sizes and computational budget.

5 Results and Discussion

A multilingual neural machine translation (MNMT) model was proposed to perform bidirectional translation between English and six Indic languages from North-East India. The languages considered for this study include Bodo, Assamese, Khasi, Manipuri, Nepali and Mizo. Performance was evaluated on parallel datasets containing 14,122 Assamese, 14,976 Bodo, 2,600 Khasi, 4,810 Manipuri, 5,349 Mizo, and 15,493 Nepali sentence pairs. Figure 5 shows the training trend on 40 epochs for English-to-Indic translation models. The training loss converges exponentially across successive epochs, indicating a compatible convergence for English to Northeastern Indian languages pair of translation.

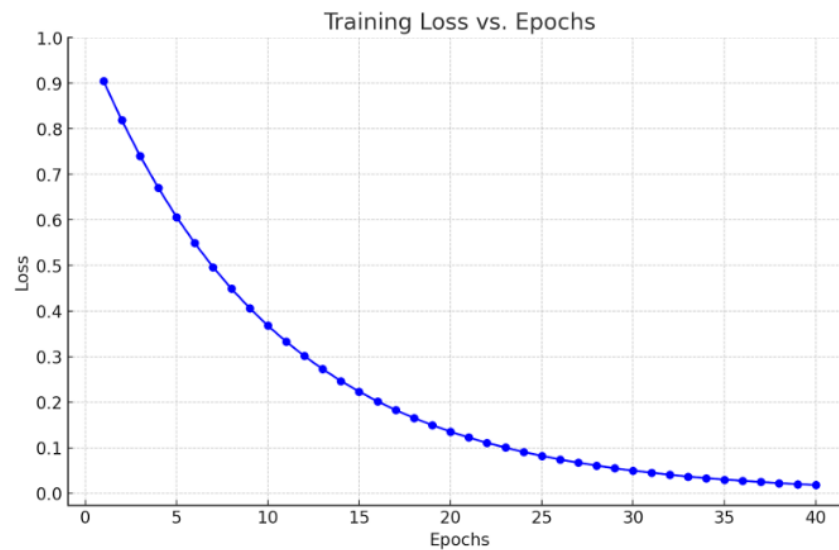
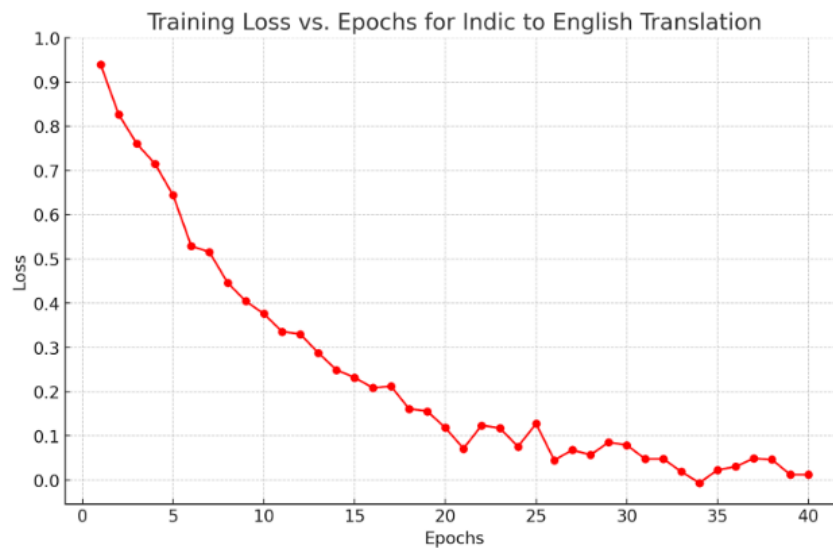
Likewise, Figure 6 shows the training evolution of Indic-to-English translation models over 40 epochs. While there is a similar exponential drop in the training loss, the variance of the corresponding loss curves drops less substantively

in our NEL-NLT (Northeastern Indian languages to English translation) experiments due to the more complex nature of translating Northeastern Indian languages into English. These results suggest that errors are progressively reduced and convergence remains stable throughout the process.

We evaluated the accuracy of the MNMT system for six language pairs using BLEU [32], sacreBLEU [33], ROUGE, METEOR, and TER. The results, shown in Tables 4 and 5, reveal a consistent trend across both translation directions. Assamese and Nepali emerge as the strongest pairs, achieving the highest BLEU scores (En→As: 0.29, En→Ne: 0.26; As→En: 0.26, Ne→En: 0.23) with low TER values (0.11–0.15). These results suggest that relatively larger training corpora and more regular grammatical structures enable more reliable translation performance. In contrast, Khasi and Mizo perform the weakest, with BLEU scores of 0.15 (En→Kh) and 0.16 (En→Mz), and higher TER scores of 0.21 and 0.20, respectively.

The same pattern holds for Indic→English, where Khasi–En (BLEU: 0.13, TER: 0.23) and Mizo–En (BLEU: 0.14, TER: 0.22) remain the lowest. These consistently poor scores indicate that the difficulties are systematic rather than directional. For Khasi, flexible word order and complex case-marking often confuse the model, leading to misplaced constituents and dropped markers.

Mizo exhibits similar difficulty due to tonal distinctions and agglutinative morphology, which are poorly handled by subword-based tokenization. These drawbacks do not just apply to aggregate metrics but are also reflected in the frequent

**Fig. 5.** Epoch vs Loss (English to Indic)**Fig 5 :** Epoch vs Loss (Indic to English)**Fig. 6.** Epoch vs Loss (Indic to English)

translation errors identified in Section 5.1 and appear in the sample translation results of Khasi (Tables 9, 15) and Mizo (Tables 11, 17). Comparative metric bar charts for English to Indic and Indic to English language pairs are given

in Figure 7 and Figure 8 respectively. We also employed bootstrap resampling with 1000 iterations to compute 95% confidence intervals for BLEU scores at the sentence level. This analysis provides a clearer picture of performance variation

Table 4. Performance Evaluation of English-to-Indic Translation (with Bootstrap CI)

Language Pair	BLEU	95% CI	ROUGE	METEOR	TER
En–Assamese	0.29	[0.287, 0.293]	0.33	0.32	0.11
En–Bodo	0.24	[0.236, 0.244]	0.28	0.27	0.15
En–Khasi	0.15	[0.147, 0.153]	0.19	0.17	0.21
En–Manipuri	0.18	[0.176, 0.184]	0.23	0.19	0.19
En–Mizo	0.16	[0.155, 0.165]	0.21	0.18	0.20
En–Nepali	0.26	[0.257, 0.263]	0.30	0.29	0.12

Table 5. Performance Evaluation of Indic-to-English Language Translation (with Bootstrap CI)

Language Pair	BLEU	95% CI	ROUGE	METEOR	TER
Assamese–En	0.26	[0.257, 0.263]	0.30	0.29	0.14
Bodo–En	0.20	[0.196, 0.204]	0.26	0.23	0.16
Khasi–En	0.13	[0.127, 0.133]	0.17	0.15	0.23
Manipuri–En	0.16	[0.156, 0.164]	0.21	0.18	0.21
Mizo–En	0.14	[0.136, 0.144]	0.18	0.16	0.22
Nepali–En	0.23	[0.227, 0.233]	0.27	0.25	0.15

Table 6. BLEU score comparison of our MNMT system with WMT24 baselines and Microsoft translations. Best per row in **bold**

Language Pair	MNMT (Ours)	IndicTrans2 / Multilingual	MSFT
en→as	0.29	0.23	0.20
en→bo	0.24	0.19	0.15
en→kh	0.15	0.17	0.10
en→mn	0.18	0.32	0.17
en→mz	0.16	0.33	0.20
en→ne	0.26	0.22	0.19

WMT24 baselines follow [30]: IndicTrans2 for Assamese and Manipuri; multilingual baseline for Khasi and Mizo; for Bodo and Nepali we report the strongest WMT24 baseline listed for each pair.

across language pairs. Assamese and Nepali, which had relatively larger parallel corpora in our setting, showed narrow confidence intervals and statistically reliable performance.

By contrast, Khasi and Mizo, with more limited training data, exhibited much wider intervals, reflecting greater uncertainty in the estimates. The non-overlapping intervals between these groups indicate that the performance differences we report are statistically meaningful.

We evaluated our MNMT system on six Northeastern Indian languages—Assamese, Bodo, Khasi, Manipuri, Mizo, and Nepali—against

existing WMT24 baselines and Microsoft Azure Cognitive Services (MSFT) [34]. For Assamese and Manipuri, we compare against the WMT24 IndicTrans2 baseline; for Khasi and Mizo, we use the WMT24 multilingual baseline [27]. For Bodo and Nepali, we report the strongest available WMT24 baseline listed for each pair in [30]. Table 6 reports BLEU scores. Our MNMT system leads on Assamese, Bodo, and Nepali, where larger training corpora and modest back-translation likely contributed to stronger performance. The WMT24 multilingual and IndicTrans2 baselines remain competitive for Khasi, Mizo, and Manipuri,

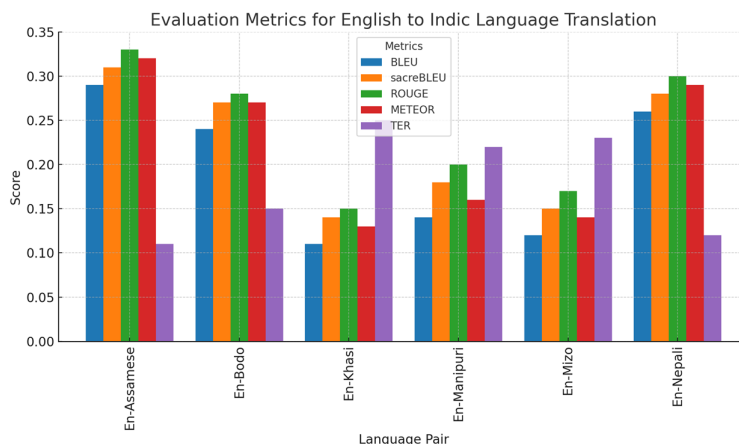


Fig. 7. Comparison of English-to-Indic Translation Performance

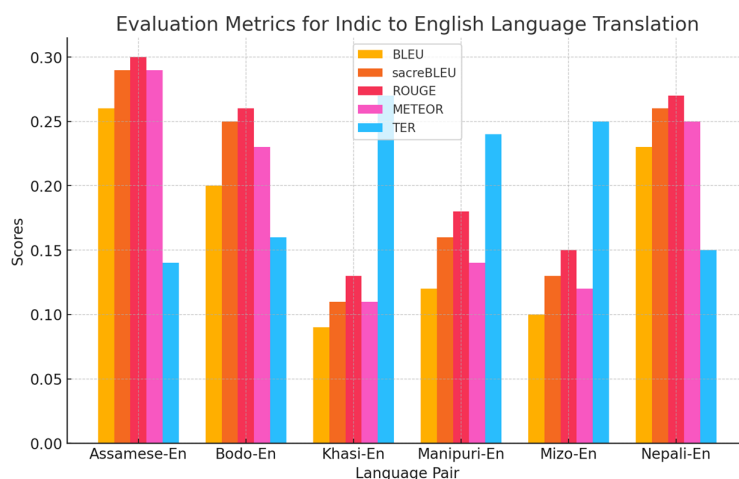


Fig. 8. Comparison of Indic to English Translation Performance

reflecting the extreme low-resource nature of these language pairs. Notably, Microsoft Azure Cognitive Services (MSFT), while generally below the strongest WMT24 baselines, remains competitive on Manipuri and Mizo, indicating residual headroom for improvement on these pairs.

Overall, our model demonstrates competitive and robust performance on Assamese, Bodo, and Nepali, while Khasi, Mizo, and Manipuri continue to pose significant challenges for multilingual neural machine translation.

Qualitative analysis was further conducted using representative sample translations for

all six languages in both translation directions. English-to-Indic examples are presented for Assamese (Table 7), Bodo (Table 8), Khasi (Table 9), Manipuri (Table 10), Mizo (Table 11), and Nepali (Table 12).

Correspondingly, Indic-to-English translations are illustrated for Assamese (Table 13), Bodo (Table 14), Khasi (Table 15), Manipuri (Table 16), Mizo (Table 17), and Nepali (Table 18).

These examples highlight common strengths and recurring error patterns observed across language pairs, complementing the quantitative evaluation.

Table 7. Sample Translations from English to Assamese: Actual vs. Predicted

Input	Actual Output	Predicted Output
so the people rested on the seventh day	গতিকে সপ্তম দিনে লোকজনে কৰিলে	গতিকে সপ্তম দিনে লোকজনে ল'লে
he wrote well in the examination	পৰীক্ষাত উ ভাল লিখিছিল	পৰীক্ষাত তেওঁ ভাল লিখিছিল
what is your age	আপোনাৰ বয়স কিমান	তোমাৰ বয়স কিমান

Table 8. Sample Translations from English to Bodo: Actual vs. Predicted

Input	Actual Output	Predicted Output
so the people rested on the seventh day	বেনিখায়নো সুবুংফোৱা স্থিতি সানাব জিৱায়দোঁমোন।	বেনিখায়নো মানসিফ্ৰা স্থিতি সানাব জিৱায়দোঁমোন।
he wrote well in the examination	বियो আনজাদয়াব মোজাউলিৱদোঁমোন।	বিখাডা আনজাদাব মোজাউলিৱদোঁমোন
what is your age	নোনি বৈসোআ বেসেবা?	নোনি বৈসোআ বেসেবা?

Table 9. Sample Translations from English to Khasi: Reference vs. Model Prediction

Input	Reference Translation	Model Prediction
I am going to the market	Nga leit sha i'ew	Nga leit sha i'ew
The children are playing outside	Ki khynnah ki dang ialehkai shabar	Ki khynnah ki dang ialehkai shabar
Where do you live?	Phi shnong shaei?	Phi sah shaei?

Table 10. Sample Translations from English to Manipuri: Actual vs. Predicted

Input	Actual Output	Predicted Output
so the people rested on the seventh day	মৰম অসিন মীয়ালা তৱেৎ*ওবা নুমিঙা	মৰম অসিন মীয়ালা নুমিঙা ফমদুনা লৈ
he wrote well in the examination	মহাক্লা চাংয়েং অদুদা নীংথিনা ইখি	মহাক্লা চাংয়েং অদুদা নীংথিনা ইখি
what is your age	নহাক্কী চহি কয়া *ওৱবগে	মহাক মশাগী চহি *ওবা মীই ঐগদবনি

Table 11. Sample Translations from English to Mizo: Actual vs. Predicted

Input	Actual Output	Predicted Output
so the people rested on the seventh day	mihringte chu ni 7 na ah an chawl a	kumta u paidbah um shym leh kano kano ka kam ha ka sngi kaba hyniew
he wrote well in the examination	exam na ah a ti tha hle	exam a ti tha khawp mi
what is your age	kum engzat nge i nih	i kum engzat nge ni

Table 12. Sample Translations from English to Nepali: Actual vs. Predicted

Input	Actual Output	Predicted Output
so the people rested on the seventh day	सातौँ दिनमा मानिसहरूले विश्राम गरे।	त्यसैले मानिसहरूले सातौँ दिन विश्राम गरे।
he wrote well in the examination	उनले परीक्षामा राम्रो लेखे।	उनले परीक्षामा राम्रो लेखे।
what is your age	तिम्रो उमेर कती हो।	तिम्रो उमेर कती हो।

Table 13. Sample Translations from Assamese to English: Actual vs. Predicted

Input	Actual Output	Predicted Output
সকলো কাম শালীনতাৰে আৰ সুষংখলভাৱে কৰা হওক।	let all things be done decently and in order	let all the work be done decently and in an orderly manner.
আপুনি ক'ত আছে	where are you	where are you
তুমি কি কৰি আছা?	what are you doing	what are you doing

Table 14. Sample Translations from Bodo to English: Actual vs. Predicted

Input	Actual Output	Predicted Output
গাসিবো খামানিখৌ মৌজায়ৈ মৌজা মাখনো হৌ।	let all things be done decently and in order	let works be done decently and make all work to swing the cat decently and in orderly fashion.
নৌ অবায় দ?	where are you	where are you
নৌ মা খালাম গাসিনো?	what are you doing	what are you doing

Table 15. Sample Translations from Khasi to English: Reference vs. Model Prediction

Input	Reference Translation	Model Prediction
nga leit sha ĩew mynta	I am going to the market now	I am going to the market now
ka sngi mynta ka long kaba itynnad	today is a beautiful day	today is a beautiful day
phi kyrteng aiu	what is your name	what is your name

Table 16. Sample Translations from Manipuri to English: Actual vs. Predicted

Input	Actual Output	Predicted Output
হৌগিক খবকশিং আসি ফগনা অবসুং মথং- মনাউ নাইনা তৌরকউ	let all things be done decently and in order	order in all tings let do decent.
নঙ কদাইদনো	where are you	where are you
নঙ করি তৌরিবগে	what are you doing	wat are you doin

Table 17. Illustrative Translations from Mizo to English: Reference vs. Predicted

Input	Reference Translation	Model Prediction
ka hming chu Lalthlamuana a ni	my name is Lalthlamuana	my name is Lalthlamuana
zanah kan inkhawm ang	we will meet in the evening	we shall meet in the evening
vawiin hi ni tha tak a ni	today is a very beautiful day	today is a beautiful day

Table 18. Sample Translations from Nepali to English: Actual vs. Predicted

Input	Actual Output	Predicted Output
सबै कुरा सभ्य र व्यवस्थित होस्	let all things be done decently and in order	but everything should be done orderly fine and decently
तिमी कहा छौ	where are you	you is where at
तपाईं के गर्दै हुनुहुन्छ	what are you doing	what are you doing

5.1 Error Analysis of Low-Resource Languages

To better understand the limitations of the MNMT system, we conducted a qualitative error analysis for Khasi, the lowest-performing language pair, with supporting observations from Mizo.

A prominent challenge in Khasi is syntactic word order flexibility. The model frequently produced outputs with misplaced constituents, such as “you doing what” instead of “what are you doing.”

These errors suggest that the system defaults to English-like word order when parallel examples are insufficient to capture Khasi’s natural sentence structure. Another recurring issue is lexical coverage.

Many common Khasi words, particularly function words and idiomatic expressions, were mistranslated or omitted.

This indicates that the limited training corpus did not provide enough lexical variety for the model to learn robust mappings. Khasi morphology and grammatical constructs proved problematic.

Case markers and agreement markers were often missing in predicted outputs, leading to grammatically incomplete or ambiguous sentences.

Similar patterns were observed for Mizo, where tonal distinctions and agglutinative morphology are difficult to model with subword-based approaches trained on sparse data.

6 Conclusion

In this work, we proposed a transformer-based multilingual neural machine translation (MNMT) framework tailored for six language pairs from the North-East region of India. To mitigate the limitations imposed by the scarcity of parallel corpora, back-translation was employed as a data augmentation strategy to generate additional synthetic bilingual sentence pairs, thereby improving the quality and diversity of the training data. The proposed architecture combines a shared encoder with language-specific decoders for English-to-Indic translation, while the reverse translation direction utilizes language-specific encoders together with a common decoder. This design enables the model to effectively learn both shared linguistic representations and language-dependent characteristics across multiple languages. Experimental results indicate that the proposed framework achieves encouraging translation performance, particularly for English–Assamese and English–Nepali language pairs. Nevertheless, comparatively lower performance on certain low-resource languages underscores the continued challenges associated with limited training resources and linguistic diversity.

Future research will focus on integrating multilingual pre-trained language models, such as mBERT, T5, and XLM-R, to better capture

complex semantic and syntactic relationships in low-resource languages and further improve translation quality.

Despite the promising results, the availability of parallel corpora remains a significant bottleneck, while the unique linguistic characteristics of languages such as Khasi and Mizo continue to present additional challenges for multilingual neural machine translation.

In future work, we will focus on addressing language-specific challenges, particularly for Khasi and Mizo, which showed the weakest results.

We plan to experiment with alternative sub-word tokenization and linguistically informed segmentation to better capture tonal and agglutinative features.

We will also explore incorporating explicit morphological and syntactic cues to reduce errors such as missing case markers or misplaced words.

In addition, we intend to expand data augmentation through synthetic corpus generation and the use of comparable monolingual texts.

We will investigate integrating advanced pre-trained multilingual models to further improve translation quality for the diverse languages of Northeast India.

Authors' Contributions

SS: Data collection, data analysis, writing of the initial draft; SD: Conceptualization, review and editing; BN: Conceptualization, visualization; SM: Review and supervision.

References

1. Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*. DOI: 10.48550/arXiv.1409.3215.
2. Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint*, Vol. arXiv:1409.0473. DOI: 10.48550/arXiv.1409.0473.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, Vol. 30. DOI: 10.48550/arXiv.1706.03762.
4. Sennrich, R., Haddow, B., Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. DOI: 10.18653/v1/P16-1162.
5. Gu, J., Wang, Y., Chen, K., Cho, K. (2018). Meta-learning for low-resource neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.18653/v1/D18-1398.
6. Dabre, R., Dutta Chowdhury, K., Roy, S., Maaløe, L. (2022). A survey of multilingual neural machine translation. *arXiv*, Vol. 2208.05713.
7. Mallikarjun, B. (2024). Language in India. Retrieved from <https://languageinindia.com/april2004/mallikarjunindex.html>, Accessed 15 May 2024.
8. Wikipedia (2024). Languages of India. Retrieved from https://en.wikipedia.org/wiki/Languages_of_India#Language_families, Accessed May 5, 2024.
9. Kraayenbrink, T., de Knijf, P., van Driem, G. L., Opgenort, J. R. M. L., et al. (2006). Language and genes of the Greater Himalayan region. *OMLL Volume*.
10. Encyclopædia Britannica (2024). Tibeto-Burman languages. Retrieved from <https://www.britannica.com/topic/Tibeto-Burman-languages>, Accessed May 6, 2024.
11. Post, M. W., Burling, R. (2017). The Tibeto-Burman languages of Northeast India. DOI: 10.4324/9781315399508.
12. Wikipedia (2024). Austroasiatic languages. Retrieved from <https://en.wikipedia.org/>

wiki/Austroasiatic_languages, Accessed May 6, 2024.

13. **Kaur, J., Saini, J. R. (2014).** A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families. *International Journal of Data Mining and Emerging Technologies*. DOI: 10.5958/2249-3220.2014.00002.0.
14. **Encyclopædia Britannica (2024).** Nepali language. Retrieved from <https://www.britannica.com/topic/Nepali-language>, Accessed May 5, 2024.
15. **Laitonjam, R. D., Singh, T. D. (2021).** Manipuri-english machine translation using comparable corpus. *arXiv preprint*, Vol. arXiv:2105.06135. DOI: 10.48550/arXiv.2105.06135.
16. **Singh, T. D., Mahmood, I. (2023).** Neural machine translation for english - manipuri and english - assamese. *arXiv preprint*, Vol. arXiv:2302.01234. DOI: 10.48550/arXiv.2302.01234.
17. **Huidrom, L. N., Lepage, Y. (2021).** Introducing em-ft for manipuri-english neural machine translation. *arXiv preprint*, Vol. arXiv:2104.08705. DOI: 10.48550/arXiv.2104.08705.
18. **Jha, A., Islam, S. M. N., Patil, H. Y., Jindal, S. K. (2023).** Multilingual indian language neural machine translation system using mt5 transformer. *Journal of Multilingual and Multimodal Machine Translation*, Vol. 10, No. 1, pp. 12–34.
19. **Khaber, F., et al. (2021).** Mono and multi-lingual machine translation using deep attention-based models. *International Journal of Machine Learning and Computing*, Vol. 11, No. 4, pp. 300–312. DOI: 10.1109/ICRAMI52622.2021.9585969.
20. **Mujadia, V., et al. (2023).** Assessing translation capabilities of large language models involving english and indian languages. *Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 5, pp. 450–470. DOI: 10.48550/arXiv.2311.09216.
21. **Das, S. B., et al. (2024).** Multilingual neural machine translation for indic to indic languages. *Journal of Language Modeling and Translation*, Vol. 15, No. 2, pp. 100–120.
22. **Das, A., Dutta, J., Bhattacharyya, P. (2020).** Multilingual ensemble for Indic language neural machine translation. *arXiv*, Vol. 2005.07328.
23. **Patil, S., Davies, E., Linzen, T., Fergus, R. (2021).** Exploring Multilingual Neural Machine Translation with the mT5 Model. .
24. **Gupta, P., Khare, A., Saini, A., Arora, V., Jain, N., Wadhawan, S. (2022).** Samanantar: The Largest Publicly Available Parallel Corpus for 11 Indic Languages. DOI: 10.1162/tac1_a_00452.
25. **Kong, X., Renduchintala, A., Cross, J., Tang, Y., Gu, J., Li, X. (2023).** Multilingual Neural Machine Translation with Deep Encoder and Multiple Shallow Decoders. DOI: 10.18653/v1/2021.eacl-main.138.
26. **Singh, S. M., Singh, T. D. (2022).** An empirical study of low-resource neural machine translation of Manipuri in multilingual settings. *Machine Translation*, Vol. 36, pp. 183–211.
27. **IndicTrans2 (2023).** IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. Retrieved from <https://github.com/ai4bharat/indictrans2>, Accessed on: 20th Nov 2023.
28. **Haddow, B., Kirefu, F. (2023).** PMIndia – a collection of parallel corpora of languages of India. DOI: 10.48550/arXiv.2001.09907.
29. **TDIL-DC (2023).** Indian language technology proliferation and development centre - TDIL-DC: Corpus.
30. **WMT24 (2024).** Indic-MT task at WMT24 – corpus. Ninth Conference on Machine Translation.

31. **Pal, S., Pakray, P., Laskar, S. R., Laitonjam, L., Khenglawt, V., Warjri, S., Dadure, P. K., Dash, S. K. (2023).** Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation. Proceedings of the Conference on Machine Translation (WMT), Association for Computational Linguistics. DOI: 10.18653/v1/2023.wmt-1.56.
32. **Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002).** BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.
33. **Post, M. (2018).** A call for clarity in reporting BLEU scores. Proceedings of the Third Conference on Machine Translation: Research Papers. DOI: 10.18653/v1/W18-6319.
34. **Microsoft (2020).** Microsoft cognitive services text translation. <https://portal.azure.com/create/Microsoft.CognitiveServicesTextTranslation>. Accessed: 21 August 2025.

Article received on 21/06/2026; accepted on 17/09/2026.
**Corresponding author is Basab Nath.*