Synthetic Data and Model Dynamics based Performance Analysis for Assamese-Bodo Low Resource NMT

Kuwali Talukdar, Shikhar Kumar Sarma and **Kishore Kashyap**
Department of Information Technology, Gauhati University Assam, India
{kuwalitalukdar, sks001, kb.guwahati}@gmail.com

Abstract

This paper presents details of modelling and performance analysis of Neural Machine Translation (NMT) for the low-resource Assamese-Bodo language pair, focusing on model tuning and the use of synthetic data. Given the scarcity of parallel corpora for these languages, synthetic data generation techniques, such as back-translation, were employed to enhance translation performance. The NMT architecture was used along with necessary preprocessing steps as per the NMT pipeline. Experimentation across varying model parameters have been performed and scores are recorded. The model's performance was evaluated using the BLEU score, which showed significant improvement when synthetic data was incorporated into the training process. While a base model with gold standard data of relatively smaller size yielded Overall BLEU of 11.35, optimized tuned model with synthetic data has resulted considerable improvement in BLEU scores across the domains, with overall BLEU upto 14.74. Challenges related to data scarcity and model optimization are also discussed, along with potential future improvements.

1 Introduction

Neural Machine Translation (NMT) has become the leading approach for automatic translation between languages. However, for low-resource language pairs, such as Assamese and Bodo, there are significant challenges due to the lack of large parallel corpora, which are necessary for training high quality NMT models. Assamese and Bodo are spoken mainly in the Northeastern region of India, but due to their limited global usage, resources for these languages are scarce. This study focuses on building an NMT system for the Assamese-Bodo language pair by utilizing synthetic data and tuning the NMT model to achieve better performance. Synthetic data generation, particularly back-translation, is often used in low-resource language translation tasks to artificially create additional training data. By using this method, we can augment the available parallel data and improve the translation quality.

The NMT architecture used in this research follows a standard pipeline that includes tokenization, sentence splitting, and other preprocessing steps necessary for effective training. We specifically evaluate the model's performance using the BLEU score, which is a common metric for measuring translation quality.

In this study, we aim to address the following research questions:

i. How can synthetic data improve translation performance for Assamese-Bodo, a low-resource language pair?

ii. What are the effects of model tuning on translation quality in an NMT setup for this language pair?

iii. What challenges and limitations arise when working with low-resource language pairs, and how can they be mitigated?

This work contributes to the growing field of NMT for low-resource languages, with a specific focus on Assamese and Bodo. The results of this study provide valuable insights that can be applied to similar low-resource language pairs, where data scarcity is a significant challenge.

Next section of the paper provides a review of related work in NMT for low-resource languages. Then the methodology is presented, including the synthetic data generation process and the NMT architecture. Results and performance analysis using BLEU scores are discussed, and challenges and future directions are mentioned concluding the paper.

2 Literature Review

Neural Machine Translation (NMT) has gained significant attention in recent years, primarily due to its ability to achieve state-of-the-art performance for a variety of language pairs. However, for low-resource language pairs such as Assamese and Bodo, the lack of sufficient parallel corpora presents a major obstacle. Several research efforts have explored strategies to overcome this challenge, focusing on synthetic data generation, model optimization, and leveraging related language pairs.

One of the pioneering works in Neural Machine Translation (NMT) for low-resource languages is the use of synthetic data, primarily through back-translation, as introduced by Sennrich et al. (2016). Their approach demonstrated that creating synthetic target-side data improves translation quality when real parallel data is scarce. Subsequent research, including Edunov et al. (2018), expanded this technique by exploring various ways of generating synthetic data and evaluating its impact on translation models. Research by Koehn and Knowles (2017) provided a comprehensive analysis of NMT performance across different resource settings, showing that while NMT performs well with large datasets, its effectiveness diminishes in low-resource conditions. To mitigate this, works like that of Nguyen and Chiang (2017) employed transfer learning, leveraging high-resource language pairs to improve translation models for low-resource languages. The use of synthetic data has proven to be a powerful tool for improving translation models, especially in low-resource settings. Currey et al. (2017) demonstrated the efficacy of synthetic data through copying monolingual data and aligning it with translated pairs, further enhancing the translation model's performance. Similarly, Hoang et al. (2018) employed unsupervised NMT to generate synthetic parallel data, which proved particularly beneficial for languages with minimal training data.

In the context of Assamese and Bodo, language processing research is still in its early stages. However, some efforts have been made to build resources and tools for these languages. Sarma et al. (2023) developed a parallel Assamese-Bodo dataset, which marked a crucial step toward advancing NMT for this language pair. Another previous work (Sarma et al., 2024) focused on the development of a baseline NMT system for Assamese-Bodo, using traditional NMT pipelines with necessary preprocessing. This study introduced the first benchmark results for Assamese-Bodo translation using BLEU scores, offering insights into the specific challenges faced in low-resource translation tasks.

Neural Machine Translation (NMT) has become a major focus of research for low-resource languages, such as Assamese and Bodo, due to the limited availability of parallel corpora. The majority of recent research has explored techniques to enhance the translation quality through data augmentation, hyperparameter tuning, and hybrid approaches (Talukdar et al., 2023). These approaches often use tokenization methods like Byte Pair Encoding (BPE), SentencePiece, and WordPiece, which reduce vocabulary size and help manage out-of-vocabulary issues (Kanchan Baruah et al., 2014).

One significant challenge in Assamese-Bodo translation is the scarcity of high-quality parallel corpora. Studies have shown that adding synthetic data, such as that generated through back-translation or created by expert linguists, significantly improves the performance of NMT models. Talukdar et al. (2023) reported experiments with 70,000 parallel sentences, along with additional gold standard datasets, achieving BLEU scores up to 17 through careful hyperparameter tuning and cross-validation (Talukdar et al., 2023).

The influence of data preprocessing methods and the quality of linguistic resources has also been highlighted as critical factors in translation performance. For Assamese-Bodo translation, tokenization using tools such as IndicNLP Suite and OpenNMT's preprocessing techniques were found to be effective in managing linguistic complexities. (Kuwali Talukdar, Shikhar Kumar Sarma, 2023).

Several studies on Assamese-English translation have also contributed to the development of resources and methodologies applicable to Assamese-Bodo. Kanchan Baruah et al. (2014) introduced a bilingual translation system, integrating transliteration to improve translation performance for low-resource settings (Kanchan

Baruah et al., 2014). Similarly, Pathak & Pakray (2018) have demonstrated that the use of neural machine translation models can yield significant improvements even in low-resource scenarios, leveraging effective training techniques (Pathak & Pakray, 2018).

Furthermore, phrase-based statistical machine translation has been explored to address the limitations of purely neural approaches. The development of phrase-based systems for English-Bodo has provided insights into domain-specific translation, achieving promising BLEU scores in areas like tourism (Islam & Purkayastha, 2018).

The primary challenge in NMT for the Assamese-Bodo language pair, and low-resource languages in general, lies in the limited availability of parallel data. To overcome this, synthetic data generation techniques have been widely used. In our study, we adopt back-translation to augment the Assamese-Bodo dataset, following the successful methods of previous works. Another challenge is model tuning, as low-resource NMT models are sensitive to hyperparameters and preprocessing techniques. Previous work has emphasized the importance of careful preprocessing, including tokenization, sentence splitting, and aligning the data for optimal performance.

In summary, the related previous literature indicates that while NMT models can be effectively applied to low-resource languages, they require a combination of synthetic data generation and careful model tuning. Our contribution builds on these ideas, focusing on Assamese-Bodo NMT using synthetic data and BLEU score evaluation.

3 Methodology

In this chapter, we describe the methodology followed for developing the Neural Machine Translation (NMT) model for the Assamese-Bodo language pair. Given the low-resource nature of these languages, we rely on synthetic data generation and proper tuning of the NMT model to improve translation performance. The approach is structured around data preprocessing, synthetic data generation through back-translation, and training the NMT model with the appropriate configurations.

3.1 Data collection and preprocessing

The first step in any NMT pipeline is data collection. For Assamese-Bodo, we began with a parallel corpus created through prior research. This dataset contains sentence pairs that have been manually aligned for translation purposes. However, the dataset is small and insufficient for training a robust NMT model. Therefore, we needed to augment the data using synthetic methods. This base dataset comprises of 110541 parallel sentence pairs. This is a gold standard parallel dataset, as it includes validated, manually created, and aligned parallel sentences.

Preprocessing: The raw data was preprocessed to ensure it is in the correct format for the NMT model. The following preprocessing steps were applied:

Tokenization: Sentences in both Assamese and Bodo were tokenized using sentencepiece, which helps convert text into subword units. This ensures that the model can handle out-of-vocabulary words during training.

Data Cleaning: Noisy sentence pairs (mistranslations, misalignments) were filtered out to maintain data quality. Sentences that were too long were removed to make the data more manageable for the model.

These steps were essential for ensuring that the data was properly formatted and ready for training.

3.2 Synthetic data generation

Given the limited availability of parallel Assamese-Bodo data, synthetic data generation was employed using back-translation. This technique has proven effective for low-resource languages, as demonstrated by a few previous works. In back-translation, we first train a model to translate from Bodo to Assamese using the available parallel data. Then, we use the trained model to translate monolingual Bodo sentences into Assamese, thereby generating synthetic parallel sentence pairs. This method helps create additional training data that boosts the performance of the primary NMT model. However the quality depends on both the quality of monolingual raw data, and the baseline model through which such data are synthesized.

The process of synthetic data generation can be outlined as follows:

i. Train a reverse translation model (Bodo-to-Assamese) using the available parallel corpus.
ii. Select a set of monolingual Bodo sentences.
iii. Use the reverse model to generate Assamese translations of these Bodo sentences.
iv. Combine these generated sentence pairs with the original parallel data to form a larger dataset for training.

Back-translation adds diversity to the training data, helping the model generalize better to unseen text. However the overall quality of the generated synthetic data is not of gold standard, and requires checking, validation, and filtering. It has been observed that quality of translation of longer sentences are poor across the sentences, and need to be filtered out. Shorter sentences have been observed to be good quality, and adds to the size of the training corpus for next iteration of model training. A thorough examination on different aspects of synthetic data generated shall through light on the overall quality, which is planned to be performed in continuation of the current research.

3.3 NMT architecture and training

For this study, we used a standard NMT architecture based on the Transformer model. The Transformer has become the preferred choice for machine translation due to its ability to handle long-range dependencies in text and its parallel processing capabilities. The architecture was chosen because of its superior performance compared to earlier sequence-to-sequence models like LSTM and GRU.

Model Components:
Encoder-Decoder Architecture: The Transformer is based on the encoder-decoder structure, where the encoder reads the input sentence (Assamese), and the decoder generates the corresponding output sentence (Bodo).
Attention Mechanism: The model leverages a self-attention mechanism, which helps focus on different parts of the input sentence when generating the translation.
The model was trained with the following configurations:

- Number of Encoding Layers: 3
- Number of Decoding Layers: 3
- Attention Heads: 4
- Learning Rate: 2
- Optimizer: Adam optimizer-2
- Batch Size: 256
- Save Checkpoint Steps: 10000
- Report_every: 10000
- Dropout: 0.1
- rnn_size: 256

Training was carried out over multiple epochs until the model reached satisfactory convergence, monitored using validation BLEU scores.

3.4 Evaluation and testing

The performance of the NMT model was evaluated using the BLEU (Bilingual Evaluation Understudy) score, a commonly used metric for assessing the quality of machine-translated text. BLEU measures the overlap between machine-generated translations and a set of reference translations.

Domain	No of Sentences	Percent distribution
Administration	79	15.80%
Agriculture	53	10.60%
Education	92	18.40%
Healthcare	54	10.80%
Law	124	24.80%
Science & Climate	16	3.20%
Tourism	82	16.40%
	500	100%

Table 1: Test Dataset.

To test the model:

- A separate test set of 500 Assamese-Bodo sentence pairs was used.
- The model's translations were compared to the human-generated translations, and BLEU scores were calculated to assess translation quality.
- In addition to BLEU scores, qualitative evaluations were performed to analyze the adequacy and fluency of the translated sentences.

3.5 Hyperparameter tuning

Hyperparameter tuning is crucial in achieving optimal performance in any NMT model. In this study, several hyperparameters were tuned to improve the model's performance:

- Learning Rate: Experimentation was conducted with different learning rates to find the one that ensures smooth convergence.

-
- Vocab Size: Different Vocab Size were used for recording optimum performance
- Batch Size: We experimented with various batch sizes to find a balance between training time and model accuracy.
-
- Dropout Rate: Dropout was used to prevent overfitting. Various dropout rates were tested to determine the most effective rate for our dataset.
-
- Model Steps: Performances were recorded against changing Model Steps

These hyperparameter adjustments helped improve the BLEU scores and ensure the model was well-suited to the Assamese-Bodo language pair.

3.6 Challenges faced

Several challenges were encountered during this study, primarily related to the low-resource nature of the Assamese-Bodo language pair:

- Limited Data: The lack of a large parallel dataset was a significant challenge, which was mitigated by synthetic data generation.

- Domain-Specific Translations: Some words and phrases were domain-specific, making it difficult for the model to generalize across different contexts.

- Preprocessing Variability: The need for consistent and accurate preprocessing was critical, as variations in tokenization or sentence splitting could affect model performance.

The methodology outlined in this chapter provides a comprehensive approach to building an NMT system for the Assamese-Bodo language pair. By leveraging synthetic data through back-translation and tuning the NMT model, we were able to enhance the model's performance, as reflected in improved BLEU scores. The next chapter will present the results of these experiments and further analyze the performance of the model.

4 Experimentation

This chapter provides an overview of the experimental setup, the different tests conducted, and the results obtained from the Neural Machine Translation (NMT) model for the Assamese-Bodo language pair. Given the low-resource nature of these languages, our experiments were designed to assess how synthetic data and model tuning affect translation quality.

Experimental Setup

The experiments were conducted using the following hardware and software configurations:

- Hardware: Training was done on a GPU-enabled system to speed up the training process. It was a 64 bit Intel Xeon CPU, with 16 GB main memory, and NVIDIA Quadro P1000 GPU. The setup has 640 CUDA Cores and 4096 MB of GPU memory. System's graphics clock speed was: min-136 MHz, max-5010 MHz, Graphics RAM 4 GB, with system storage: 256 GB SSB and 1 TB HDD.

- Software: The experiments were implemented using PyTorch and OpenNMT, which are widely used frameworks for NMT research. Python was used for preprocessing and data handling. IndicNLP was used for tokenization. Subword_nmt was used for byte pair encoding. For BLEU calculations, SacreBLEU was used.

Model Architecture:

We used the Transformer model for our experiments. The Transformer, with its encoder-decoder structure and self-attention mechanism, was well-suited for this task due to its efficiency in handling longer sequences, even with relatively small datasets.

Data Preparation

The data used for training consisted of both real and synthetic parallel sentences. The real data comprised 110541 parallel sentence pairs. This dataset was augmented using back-translation to generate synthetic data, as described in the methodology chapter. Monolingual Bodo sentences were back-translated into Assamese to create additional synthetic sentence pairs. The reverse translation model (Bodo-to-Assamese) was trained using the real parallel dataset and then used used to generate total 330000 synthetic

Assamese translations of the monolingual Bodo sentences.

Training Process

The model was trained using the Real Parallel Data, as well as combined real and synthetic dataset, with a focus on the following key aspects:

- Training Duration: Each model was trained for 10 epochs, with early stopping implemented to prevent overfitting.

- Batch Size: A batch size of 256 was used.

- Optimizer: The Adam optimizer was employed with a learning rate of 2. A learning rate scheduler was used to adjust the learning rate during training to ensure stable convergence.

Final Dataset Composition:
- Real Parallel Data: 110541 sentence pairs

Domain	Dataset size: 110541 Vocab size: 16000	
	BLEU Training Step at 40K	BLEU Training Step at 80K
Administration	14.34	15.48
Agriculture	14.88	8.06
Education	10.88	11.33
Healthcare	11.79	15.55
Law	4.69	6.73
Science & Climate	9.61	4.11
Tourism	13.38	12.00
Overall	11.27	11.48

Table 2: BLEU of Baseline Model (Without Synthetic Data).

- Synthetic Data: 330000 sentence pairs generated through back-translation
- Total Dataset Size: 440541 sentence pairs

For testing, a Test Dataset of 500 manually crafted and curated pairs of Assamese-English sentences were used (Table 1). Throughout the training process, the validation set BLEU score was monitored to track performance and make adjustments as needed.

Evaluation Metrics

The primary metric used for evaluating the performance of the translation model was the BLEU (Bilingual Evaluation Understudy) score. BLEU measures the overlap between machine-generated translations and a set of human reference translations. Higher BLEU scores indicate better translation quality.

Domain	Dataset size: 220541	
	BLEU Vocab size: 8000	BLEU Vocab size: 16000
Administration	17.26	19.24
Agriculture	11.01	11.04
Education	10.61	10.87
Healthcare	7.94	8.59
Law	10.01	9.31
Science & Climate	10.64	9.93
Tourism	12.79	13.74
Overall	13.98	14.74

Table 3: BLEU of Augmented Model 1 (With Synthetic Data).

5 Experimental Results

The experiments yielded the following results:

When trained using only the real parallel dataset (110541 sentence pairs), the model achieved overall BLEU score of 11.48 with training steps at 80000.

The baseline results indicate that while the model was able to generate reasonable translations, the performance was limited by the size of the dataset.

After augmenting the dataset with synthetic data through back-translation (an additional 330000 sentence pairs), the model's performance improved significantly. The results were as follows:

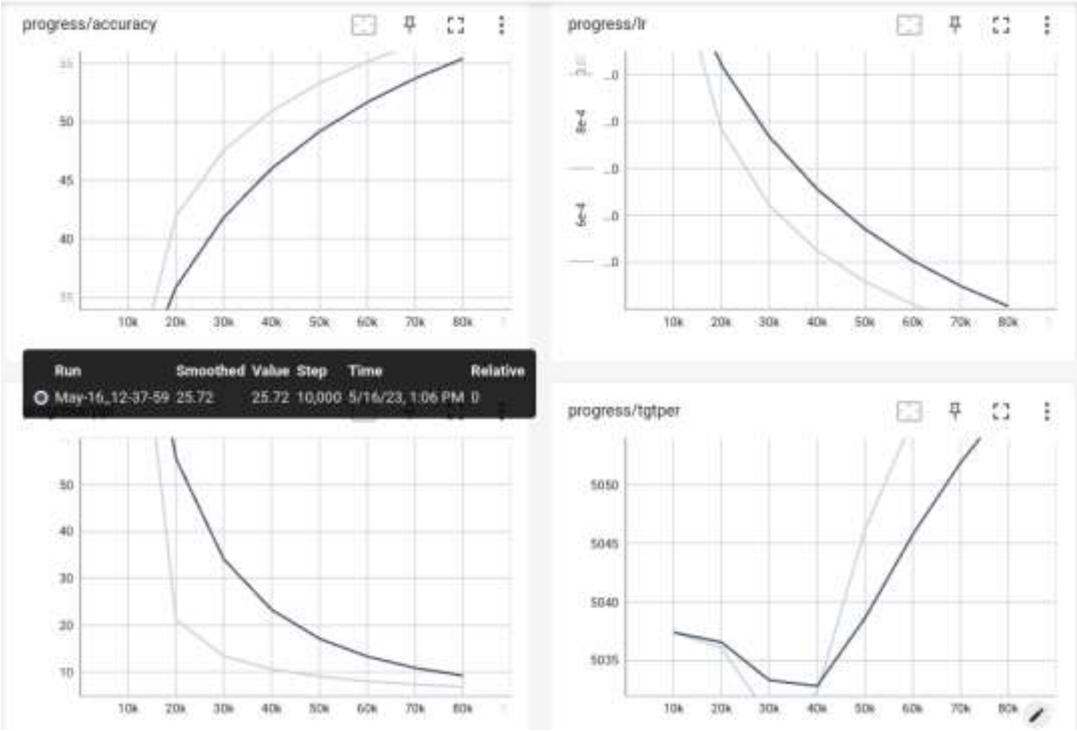
Domain	Dataset size: 440541	
	BLEU Vocab size: 8000	BLEU Vocab size: 16000
Administration	19.24	18.95
Agriculture	13.34	12.27
Education	9.83	9.51
Healthcare	8.28	7.65
Law	8.59	8.86
Science & Climate	6.37	8.62
Tourism	12.28	5.00
Overall	13.92	13.5

Table 4: BLEU of Augmented Model 2 (With Synthetic Data).

Domain	Dataset size: 440541 Vocab size: 8000	
	BLEU After Averaging	BLEU After length penalty
Overall	14.11	14.24

Table 5: Effect of Model Averaging and Length Penalty on (With Synthetic Data).

This improvement demonstrates the effectiveness of using synthetic data to enhance translation performance, particularly for this low-resource language pair. Experimental results for modeling with different sets of augmented data as well as with varying training setup are shown through Table 3 to 5.



Run	Smoothed	Value	Step	Time	Relative
○ May-16_12-37-59	25.72	25.72	10,000	5/16/23, 1:06 PM	0

Figure 1: Different plots during training (without synthetic data)

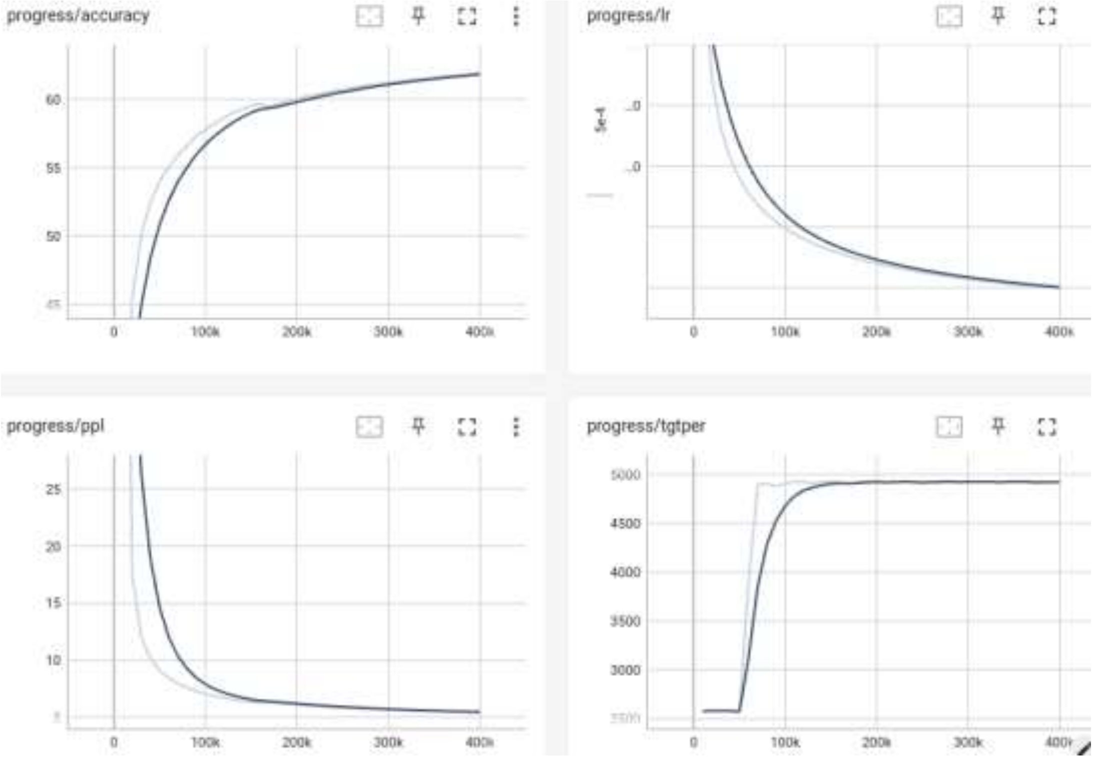


Figure 2: Different plots during training (with synthetic data)

6 Analysis of Results

The results clearly show that the inclusion of synthetic data significantly boosts the model's performance with overall score, as well as scores across all domains.

The higher scores indicate improved word-level accuracy and coherence, likely due to the increased vocabulary and exposure to diverse sentence structures provided by the synthetic data.

In terms of model tuning:

- Learning Rate: A learning rate of 2 provided the best balance between convergence speed and stability. Higher learning rates led to instability in training, while lower rates caused slow convergence.
- Batch Size: A batch size of 256 was optimal in terms of GPU memory usage and convergence time.

Qualitative Evaluation

In addition to quantitative metrics, qualitative analysis of the translations was also conducted. The following observations were made:

- Fluency: Translations from the model trained with synthetic data were notably more fluent, especially for longer sentences. The model exhibited fewer instances of repetitive or broken sentence structures.
- Adequacy: The model was able to translate common phrases and domain-specific terms more accurately, particularly in cases where the baseline model struggled.
- Errors: Some common errors included incorrect word order and occasional mistranslations of rare words. However,

these errors were less frequent in the model trained with synthetic data.

	Testset 500	IndicTrans Testset:1000
Adequacy (5 point scale)	4.04	3.2
Fluency (5 point scale)	4.08	3.23

Table 6: Human evaluation for different Testsets.

Limitations

While the synthetic data improved translation quality, there were some challenges:

- Domain-Specific Data: The model struggled with sentences from highly specialized domains that were not well-represented in the training data. This is reflected with lower performance changes in domains like Technical and Climate etc.
- Overfitting: Although early stopping and dropout were employed, there were signs of overfitting in later epochs, particularly when training on the smaller real dataset.

Conclusion:

Based on the research conducted in this paper, it can be concluded that incorporating synthetic data significantly enhances the performance of Neural Machine Translation (NMT) for the Assamese-Bodo low-resource language pair. The use of back-translation to generate synthetic data has proved effective in augmenting the limited available parallel corpus, resulting in substantial improvements in BLEU scores across all domains. The experiments also demonstrated that careful tuning of hyperparameters and preprocessing steps is crucial for optimal model performance. Despite challenges such as data scarcity and domain-specific variability, the findings suggest that synthetic data generation, combined with effective model optimization, can mitigate many limitations associated with low-resource language translation tasks. This work highlights the potential of synthetic data to bridge the resource gap and improve translation quality, providing a valuable approach for other low-resource language pairs. This may be applicable for many Indian Languages machine translation efforts as many of such languages are resource poor. Gold standard and sizable parallel dataset generation is a time consuming effort, and also demands substantial financial involvement. Synthetic Data generation with tuned and moderately performed Model shall contribute to size of the corpus, which is critical for NMT training.

Acknowledgments

Authors acknowledge Ministry of Electronics and Information Technology, Government of India for funding the ISHAAN Bidirectional MT Project, under which the development of Assamese-Bodo NMT was carried out.

References

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 86-96.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 489-500.

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, 28-39.

Nguyen, T. Q., & Chiang, D. (2017). Transfer Learning Across Low-Resource Related Languages for Neural Machine Translation. Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 296-301.

Currey, A., Miceli Barone, A. V., & Heafield, K. (2017). Copied Monolingual Data Improves Low-Resource Neural Machine Translation. Proceedings of the Second Conference on Machine Translation, 148-156.

Hoang, H., Haffari, G., & Cohn, T. (2018). Improving Neural Translation Models with Synthetic Data. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4265-4275.

Talukdar, K., Sarma, S.K., Naznin, F., Kashyap, K., Ahmed, M., & Boruah, P.A. (2023). Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair. Proceedings of the 20th International Conference on Natural Language Processing (ICON). NLP Association of India.

Neural Machine Translation for a Low Resource Language Pair: English-Bodo (2023). ACL Anthology.

Talukdar, K., Sarma, S.K., Naznin, F., Kashyap, K., Ahmed, M., & Boruah, P.A. (2023). Influence of Data Quality and Quantity on Assamese-Bodo Neural Machine Translation. IEEE Xplore.

Kanchan Baruah, K., Das, P., Hannan, A., & Sarma, S.K. (2014). Assamese-English Bilingual Machine Translation. International Journal of Natural Language Computing (IJNLC).

Pathak, A., & Pakray, P. (2018). Neural Machine Translation for Indian Languages. Journal of Intelligent Systems.

Islam, M.S., & Purkayastha, B.S. (2018). English to Bodo Phrase-Based Statistical Machine Translation. Advances in Intelligent Systems and Computing, Springer.

Pathak, A., & Pakray, P. (2019). Neural Machine Translation for Indian Languages. IEEE Conference on Information and Communication Technology.

Kuwali Talukdar, Shikhar Kumar Sarma, Farha Naznin, Kishore Kashyap, Mazida Akhtara Ahmed, and Parvez Aziz Boruah. 2023. Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair. In Proceedings of the 20th International Conference on Natural Language Processing (ICON), pages 714–719, Goa University, Goa, India. NLP Association of India (NLPAI).