An Attention-Based Neural Translation System for English to Bodo

Subhash Kumar Wary[1], **Birhang Borgoyary**[2], **Akher Uddin Ahmed**[3], **Mohanji Prasad Sah**[4], **Apurbalal Senapati**[5]
[1,2,3,4,5]Central Institute of Technology Kokrajhar
BTR Assam, India, Pin - 783370
subhashkumarwary@gmail.com, bborgoyary021@gmail.com, akheruddinahmedcse@gmail.com, mohanjiprasadsah80@gmail.com, a.senapati@cit.ac.in

Abstract

Bodo is a resource scarce, the indigenous language belongs to the Tibeto-Burman family. It is mainly spoken in the north-east region of India. It has both linguistic and cultural significance in the region. Only a limited number of resources and tools are available in this language. This paper presents a study of neural machine translation for the English-Bodo language pair. The system is developed on a relatively small parallel corpus provided by the Low-Resource Indic Language Translation as a part of WMT-2025 [1]. The system is evaluated by the WMT-2025 organizers with the evaluation matrices like BLUE, METEOR, ROUGE-L, chrF and TER. The result is not promising but it will help for the further improvement. The result is not encouraging, but it provides a foundation for further improvement.

1 Introduction

The Bodo language, which belongs to the Sino-Tibetan language family, is one of the widely spoken languages in Assam and several other parts of the North-Eastern states of India. It is used predominantly in the Bodoland Territorial Region (BTR), which includes the districts of Kokrajhar, Chirang, Baksa, and Udalguri, as well as in other districts such as Kamrup, Sonitpur, Lakhimpur, Nagaon, Morigaon, and Karbi Anglong. Bodo is one of the 22 languages listed in the Eighth Schedule of the Indian Constitution and is officially recognized by the Government of India (Census, 2011). According to the 2011 Census, it is spoken by more than a million people, primarily members of the Bodo community (Koyel Ghosh, 2023). The number of Bodo speakers is shown in the Figure 1. The Bodo language has rich linguistic features and uses the Devanagari script for writing, similar to Hindi. Having the tonal feature. Hence, effective techniques are not developed to capture all these features (Mwnthai Narzary, 2022).

Machine translation is a core application in the field of Natural Language Processing (NLP). With advancements in computational power, the focus has shifted from rule-based methods to machine learning and deep learning approaches. However, implementing deep learning techniques requires a large volume of data (Narzary Sanjib, 2019). In this paper, we have developed an English-Bodo machine translation system using a transformer-based neural machine translation approach. Preprocessing steps such as tokenization, subword extraction, and normalization are required before feeding the data into the actual Transformer model (Guillaume et al., 2017).

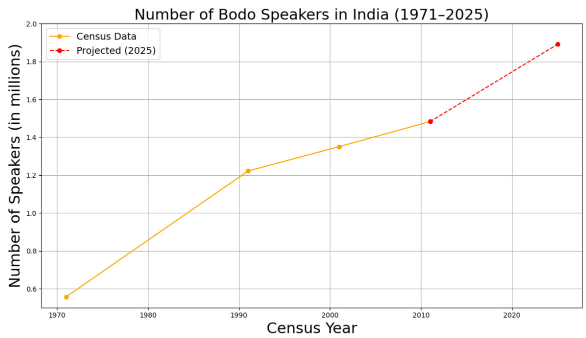


Figure 1: Number of Bodo speakers in India from 1971 to 2025 (projected).

2 Related Work

As mentioned above, Bodo is a low-resource language (Narzary et al., 2021), and development a machine translation system for it faced significant challenges due to the limited availability of parallel corpora and digital resources (Kalita et al., 2023). Most research efforts have focused on creating and expanding parallel corpora, as well as adapting machine translation techniques to function effectively

[1]https://www2.statmt.org/wmt25/indic-mt-task.html

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 968–972
November 8–9, 2025.

with scarce data. Although work in Bodo machine translation remains limited, there have been some notable efforts, particularly in building English-Bodo translation systems. A brief overview of these works is provided below.

Bahdanau et al. (Dzmitry Bahdanau, 2015) observed that traditional neural machine translation and statistical machine translation models have an alignment problem that affects the performance. The common encoder-decoder architecture, which uses an encoder to compress a source sentence into a single, fixed-length vector, faced a critical bottleneck: this fixed-length vector was often insufficient to capture all the information from a long sentence. To solve this, they proposed a new model that allows the decoder to automatically search for and focus on the most relevant parts of the source sentence when predicting each word of the translation. This mechanism, known as attention, significantly improved translation quality by overcoming the limitations of a single context vector. Verma et al. (Verma and Bhattacharyya, 2017) conducted a literature survey on Neural Machine Translation (NMT), highlighting the advantages of the NMT architecture. Vaswani et al. (Vaswani et al., 2017) proposed the attention-based Transformer model, which gained significant popularity in machine translation due to its novel architecture and promising performance. Parvez et al. (Parvez et al., 2023) attempted neural machine translation for the pair of low-resource languages of English and Bodo. They utilized a relatively small English-Bodo parallel corpus and implemented their system using the OpenNMT-py framework. Their model achieved a highest BLEU score of 11.01. Islam et al. (Islam and Purkayastha, 2019) worked on Bodo-to-English machine translation using a phrase-based statistical machine translation (PB-SMT) approach. They applied this technique to Bodo-English parallel corpora in both the general and news domains, reporting a highest BLEU score of 30.13. Narzary et al. (Narzary Sanjib, 2019) developed an attention-based English-Bodo neural machine translation system using data from the tourism domain. Their baseline model achieved a BLEU score of 11.8. By incorporating an attention mechanism, they improved the model's performance, reaching a BLEU score of 16.71. Talukdar et al. (Talukdar et al., 2023) focused on Assamese-Bodo neural machine translation and investigated the impact of data quality and quantity on translation performance. They iteratively augmented the dataset and evaluated the outcomes at each stage. The experiments were conducted using the OpenNMT-py framework. Gaikwad et al. (Gaikwad et al., 2024) suggested that the use of a high-resource language as a pivot can improve translation into related low-resource languages. They conducted experiments on machine translation of the English to Indian language - specifically translating English into Konkani, Manipuri, Sanskrit, and Bodo - employing Hindi, Marathi, and Bengali as pivot languages.

3 System Description

We have implemented the Attention-Based Neural Machine Translation (NMT) system. It is a deep learning model specifically designed for sequence-to-sequence tasks, such as translating text from one language to another. In traditional sequence-to-sequence models, the entire input sentence is encoded into a single, fixed-length vector that captures its relevant contexts. That single vector cannot effectively capture all the rich context of long or complex sentences. On the other hand, attention-based architecture allows the model dynamically to focus on relevant parts of the input sentence while generating each word in the output. The system typically consists of an encoder-decoder architecture along with an attention mechanism. The attention mechanism dynamically modifies the context vector for each output word. This allows the decoder to "attend" to different parts of the input sentence at each step. The basic encoder-decoder architecture, along with the attention mechanism, is depicted in Figure 2. The figure is configured for the English-Bodo translation, which is influenced by the Tato et al. (Tato and Nkambou, 2022) diagram.

The attention mechanism used in the decoder to decide which parts of the input sequence to focus on while generating an output. Calculating a context vector by taking a weighted sum of the encoder's hidden states. The weights for this sum are dynamically adjusted, giving more importance to the input words that are most relevant to the current output being generated. Based on this mechanism, the model can capture long-range dependencies and produce higher-quality translations. This is particularly effective for long or complex sentences or when translating between languages with different word orders.

In the attentional model (Dzmitry Bahdanau,

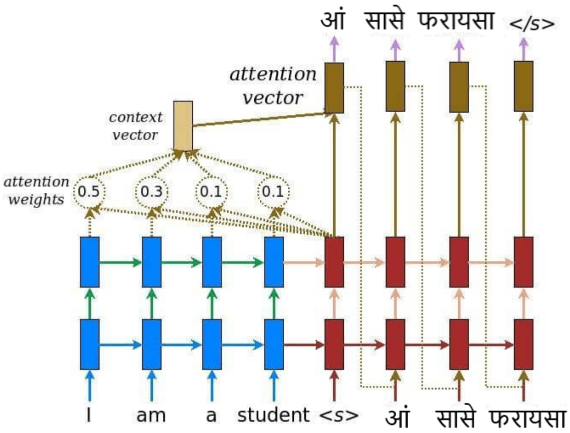


Figure 2: Example of attention mechanism in the translation from English to Bodo

2015), all hidden states h_i of the encoder are utilized to compute the context vector c_t. This model generates a variable-length alignment vector a_t, whose size corresponds to the number of time steps on the source side. The alignment vector is obtained by comparing the current hidden state h_i with each encoder hidden state $\bar{h_s}$. Where:

$$\alpha_{ts} = \frac{exp(score(h_t, \bar{h_s}))}{\sum_{s'} exp(score(h_t, \bar{h_{s'}}))} \quad (1)$$

$$c_t = \sum \alpha_{ts} \bar{h_s} \quad (2)$$

$$a_t = f(c_t.h_t) = tanh(W_c[c_t; h_t]) \quad (3)$$

$$a_t = f(c_t.h_t) = tanh(W_c[c_t; h_t]) \quad (4)$$

The $score(h_t, \bar{h_s})$ is calculated as

$$\begin{cases} {h_t}^T W \bar{h_s} & \text{[Luong]} \\ {v_a}^T tanh(W_1 h_t + W_2 h_s) & \text{[Bahdanau]} \end{cases}$$

Here, both the multiplicative and additive (Luong et al., 2015) (Dzmitry Bahdanau, 2015) attention mechanisms have normalized variants, known respectively as Scaled Luong and Normed Bahdanau (Raffel et al., 2017). The main idea behind the attention vector is to determine how much emphasis should be focused on each source word at a given time step. A higher value in the attention weight at indicates that the corresponding source word has a greater influence on predicting the next word in the output sentence. Particularly, the model the model performs translation based on the conditional probability $p(y|x)$, which represents the likelihood of translating a source sentence $x_1; ....; x_n$, into a target sentence $y_1; ....; y_m$. This is achieved using an encoder-decoder framework.

4 Dataset used

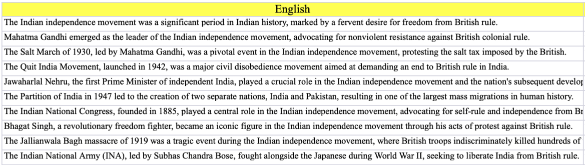
The data set we used in this work for the translation task is provided in the Very Limited Training Data setting of the WMT25 Indic Multilingual Machine Translation Shared Task [2], a snapshot of the data is shown in Figure 3 and Figure 4 respectively. Bodo is a low-resource language, and the availability of high-quality parallel corpora is significantly constrained. The official data set provided by the WMT25 organizers consists of a small number of English-to-Bodo sentence pairs curated for the purpose of benchmarking machine translation systems under low resource conditions. In the dataset, contains training, development, and test splits, with the training set including only a few thousand sentence pairs. These sentences span general purpose domains such as basic conversational language. The development set was used for validation and tuning, while the test set was reserved for blind evaluation by the task organizers.

We have also taken a parallel data set focused on tourism [3] augmented with WMT25 data for training purposes shown in Table 1. This data set contains English and Bodo corpus, which we trained and tested in a different model for the contrastive output element. The data set is divided into six subsets to facilitate training, validation, and testing for both languages involved in translation. Specifically, they have provided *train_brx, val_brx*, and *test_brx* contain Bodo sentences for training, validation, and testing respectively, while *train_eng, val_eng*, and *test_eng* contain the corresponding English sentences. Each Bodo sentence in a given split aligns with its English counterpart, enabling parallel corpus training for machine translation models.

Sl. No.	**Corpus name**	**# Files**	**# Sentences**
1	WMT25 (en-bodo)	1	15,216
2	Tourism (en-bodo)	6	33,258

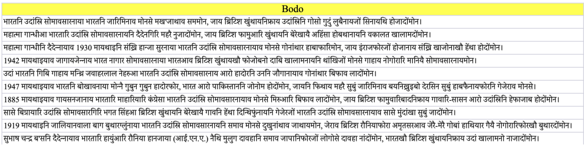
Table 1: English-Bodo training data set

[2] https://www2.statmt.org/wmt25/indic-mt-task.html
[3] https://get.alayaran.com/parallel-data/



English
The Indian independence movement was a significant period in Indian history, marked by a fervent desire for freedom from British rule.
Mahatma Gandhi emerged as the leader of the Indian independence movement, advocating for nonviolent resistance against British colonial rule.
The Salt March of 1930, led by Mahatma Gandhi, was a pivotal event in the Indian independence movement, protesting the salt tax imposed by the British.
The Quit India Movement, launched in 1942, was a major civil disobedience movement aimed at demanding an end to British rule in India.
Jawaharlal Nehru, the first Prime Minister of independent India, played a crucial role in the Indian independence movement and the nation's subsequent develop
The Partition of India in 1947 led to the creation of two separate nations, India and Pakistan, resulting in one of the largest mass migrations in human history.
The Indian National Congress, founded in 1885, played a central role in the Indian independence movement, advocating for self-rule and independence from Br
Bhagat Singh, a revolutionary freedom fighter, became an iconic figure in the Indian independence movement through his acts of protest against British rule.
The Jallianwala Bagh massacre of 1919 was a tragic event during the Indian independence movement, where British troops indiscriminately killed hundreds of
The Indian National Army (INA), led by Subhas Chandra Bose, fought alongside the Japanese during World War II, seeking to liberate India from British rule.

Figure 3: WMT25 Data set - English (eng)



Bodo
भारतनि उदांस्रि सोमावसारनाया भारतनि जारिमिनाव मोनसे मखजाथाव सममोन, जाय ब्रिटिश खुंथायनिफ्राय उदांस्रिनि गोसो गुदुं लुबैनायजों सिनायथि होजादोंमोन।
महात्मा गान्धीआ भारतारि उदांस्रि सोमावसारनायनि दैदेनगिरि महरै नुजादोंमोन, जाय ब्रिटिश फामुआरि खुंथायनि बेरेखायै अहिंसा होबथानायनि वकालत खालामदोंमोन।
महात्मा गान्धीनि दैदेन्नायाव 1930 मायथाइनि संख्रि हान्जा सुरनाया भारतनि उदांस्रि सोमावसारनायाव मोनसे गोनांथार हाबाफारिमोन, जाय इंराजफोरजों होजानाय संख्रि खाजोनाखौ हेंथा होदोंमोन।
1942 मायथाइयाव जागायजेन्नाय भारत नागार सोमावसारनाया भारतआव ब्रिटिश खुंथायखौ फोजोबनो दाबि खालामनायनि थांखिजों मोनसे गाहाय नोगोरारि मानियै सोमावसारनायमोन।
उदां भारतनि गिबि गाहाय मन्थ्रि जवाहरलाल नेहरुआ भारतनि उदांस्रि सोमावसारनाय आरो हादोरनि उननि जौगानायाव गोनांथार बिफाव लादोंमोन।
1947 मायथाइयाव भारतनि बोखावनाया मोन्नै गुबुन गुबुन हादोरफोर, भारत आरो पाकिस्ताननि जोनोम होदोंमोन, जायनि फिथाय महरै सुबुं जारिमिनाव बयनिख्रुइबो देरसिन सुबुं हाबफैनायफोरनि गेजेराव मोनसे।
1885 मायथाइयाव गायसनजानाय भारतारि माहारियारि कंग्रेसा भारतनि उदांस्रि सोमावसारनायाव मोनसे मिरुआरि बिफाव लादोंमोन, जाय ब्रिटिश फामुवारिबादनिफ्राय गावारि-सासन आरो उदांस्रिनि हेफाजाब होदोंमोन।
सासे बिग्रायारि उदांस्रि सोमावसारगिरि भगत सिंहआ ब्रिटिश खुंथायनि बेरेखायै गावनि हेंथा दिन्थिफुंनायनि गेजेरजों भारतनि उदांस्रि सोमावसारनायाव सासे मुंदांखा सुबुं जादोंमोन।
1919 मायथाइनि जालियानवाला बाग बुथारग्लुंनाया भारतनि उदांस्रि सोमावसारनायनि समाव मोनसे दुखुनांथाव जाथायमोन, जेराव ब्रिटिश रौनियाफोरा अमृतसरआव जैरै-मैरै गोबां हाथियार गैयै नोगोरारिफोरखौ बुथारदोंमोन।
सुभाष चन्द्र ब'सनि दैदेन्नायाव भारतारि हायुंआरि रौनिया हानजाया (आई.एन.ए.) नैथि मुलुग दावहानि समाव जापानिफोरजों लोगोसे दावहा नांदोंमोन, भारतखौ ब्रिटिश खुंथायनिफ्राय उदां खालामनो नाजादोंमोन।

Figure 4: WMT25 Data set - Bodo (brx)

5 Implementation

All model training and experimentation were conducted using Google Colab, a cloud-based development platform. We utilized the NVIDIA A100 GPU available through Google Colab for training our NMT model. The A100, based on NVIDIA's Ampere architecture, offers high memory bandwidth and massive parallel processing capabilities, making it a well suited for deep learning tasks. Its support for mixed precision training and large batch processing significantly accelerated model training and testing. The GPU enabled efficient handling of our encoder to decoder architecture with attention which allowed us to train and test on the English to Bodo dataset with reduced computation time and improved performance.

The model is developed and trained using PyTorch within a Google Colab environment. The data set is cleaned by removing null values and then shuffled to eliminate order bias. We load a test set provided as plain text files, ensuring that sentence alignment is preserved across the splits of training and validation sets. For preprocessing, we utilize a custom LanguageProcessor class that tokenizes the data and constructs a vocabulary with special tokens <PAD>, <SOS>, <EOS>, and <UNK>. It maps words to indices and vice versa and computes the maximum sentence length for padding.

We configured a sequence-to-sequence encoder-decoder model with attention using carefully selected hyperparameters to ensure efficient training and optimal performance. The embedding dimension for both the encoder and decoder was set to 256, while the hidden state dimensions were set to 512 units. The dropout regularization was applied with a rate of 0.5 in both the encoder and decoder. Training was conducted using a batch size of 64 over 15 epochs.

6 Result

The system is evaluated by the WMT25 organiser, which provided a test dataset of 1,000 sentences. A total of nine runs were assessed for the task. Five evaluation metrics were used: BLEU, METEOR, ROUGE-L, chrF, and TER[4]. The result of our system is shown on Table 2. For comparison, the highest and lowest scores for the track are presented in Table 3 and Table 4, respectively.

Sl. No.	Metric	Score
1	BLEU	0.3106045292
2	METEOR	0.01875594452
3	ROUGE-L	0.002595238095
4	chrF	7.235394682
5	TER	808.9101286

Table 2: Result of the system (en-bodo)

Sl. No.	Metric	Score
1	BLEU	24.44868688
2	METEOR	0.5126346512
3	ROUGE-L	0.1684904762
4	chrF	67.70727358
5	TER	51.84296487

Table 3: Highest score of the track (en-bodo)

Sl. No.	Metric	Score
1	BLEU	0.2047914219
2	METEOR	0.006037416908
3	ROUGE-L	0.02716098904
4	chrF	0.8138721309
5	TER	131.9585726

Table 4: Lowest score of the track (en-bodo)

7 Conclusion

It is observed that your result (Table 2) is not much surprising. While the score is higher than the lowest score (Table 4), it is still lower compared to the highest (Table 3). To investigate its weaknesses, a granular-level error analysis is needed. At a glance, we found that the system performs poorly on complex sentences compared to simple ones. The data

[4]https://www2.statmt.org/wmt25/mteval-subtask.html

provided by the track, along with the various systems presented here, will be valuable for future research.

Limitations

The training dataset is insufficient for developing a sophisticated machine translation system.

Ethics Statement

Not Applicable

Acknowledgements

We sincerely thank Mr. Sanjib Narzary for his valuable guidance in implementing the system.

References

Census. 2011. Abstract of speakers' strength of languages and mother tongues – census 2011. *Office of the Registrar General Census Commissioner, India. 2018*, New Delhi: Ministry of Home Affairs, Government of India.

Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Pranav Gaikwad, Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. How effective is multi-source pivoting for translation of low resource indian languages? *ArXiv*, abs/2406.13332.

Klein Guillaume, Kim Yoon, Yuntian Deng, Senellart Jean, and Rush Alexander. 2017. OpenNMT: Open-source toolkit for neural machine translation. pages 67–72.

Saiful Islam and Bipul Syam Purkayastha. 2019. Bodo to english machine translation through transliteration. *International Journal of Innovative Technology and Exploring Engineering*.

S. Kalita, P. Boruah, Kishore Kashyap, and Shikhar Kr Sarma. 2023. Nmt for a low resource language bodo: Preprocessing and resource modelling. *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–5.

Mwnthai Narzary Maharaj Brahma Koyel Ghosh, Apurbalal Senapati. 2023. Hate speech detection in low-resource bodo and assamese texts with ml-dl and bert models. *Scalable Computing: Practice and Experience*, 24(4).

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Sanjib Narzary Apurbalal Senapati Pranav Kumar Singh Mwnthai Narzary, Maharaj Brahma. 2022. A computational approach for the tonal identification in bodo language. *Bhattacharjee, R., Neog, D.R., Mopuri, K.R., Vipparthi, S.K. (eds) Artificial Intelligence and Data Science Based RD Interventions. NERC 2022*.

Mwnthai Narzary, Gwmsrang Muchahary, Maharaj Brahma, Sanjib Narzary, P. Singh, and Apurbalal Senapati. 2021. Bodo resources for nlp - an overview of existing primary resources for bodo. *Proceedings of Intelligent Computing and Technologies Conference*.

Singha Bobita Brahma Rangjali Dibragede Bonali Barman Sunita Nandi Sukumar Som Bidisha Narzary Sanjib, Brahma Maharaj. 2019. Attention based english-bodo neural machine translation system for tourism domain. pages 335–343.

Boruah Parvez, Talukdar Kuwali, Ahmed Mazida, and Kashyap Kishore. 2023. Neural machine translation for a low resource language pair: English-bodo.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2837–2846.

Kuwali Talukdar, Shikhar Kumar Sarma, Farha Naznin, and Kishore Kashyap. 2023. Influence of data quality and quantity on assamese-bodo neural machine translation. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Ange Tato and Roger Nkambou. 2022. Infusing expert knowledge into a deep neural network using attention mechanism for personalized learning environments. *Front. Artif. Intell*, 5:921476.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Ajay Anand Verma and Pushpak Bhattacharyya. 2017. Literature survey: Neural machine translation. *CFILT, Indian Institute of Technology Bombay, India*.