English to Bodo : Neural Machine Translation

A Project Report Submitted in Partial Fulfillment of the requirements for the Degree of

Bachelor of Technology (B.Tech.)

in

Computer Science and Engineering

by

Subhash Kumar Wary (Roll No. 202002022110)
Akher Uddin Ahmed (Roll No. 202102022100)
Birhang Borgoyary (Roll No. 202102023126)
Mohanji Prasad Sah (Roll No. 202102022111)

Under the Supervision of
Dr. Apurbalal Senapati
Assistant Professor

Department of Computer Science and Engineering

केन्द्रीय प्रौद्योगिकी संस्थान कोकराझार

CENTRAL INSTITUTE OF TECHNOLOGY KOKRAJHAR
(A Centrally Funded Institute under Ministry of HRD, Govt. of India)
BODOLAND TERRITORIAL AREAS DISTRICTS :: KOKRAJHAR ::
ASSAM :: 783370
Website: www.cit.ac.in
May 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

केन्द्रीय प्रौद्योगिकी संस्थान कोकराझार

CENTRAL INSTITUTE OF TECHNOLOGY KOKRAJHAR

(An Autonomous Institute under MHRD)

Kokrajhar-783370, BTR, Assam, India

CERTIFICATE OF APPROVAL

This is to certify that the work embodied in this project entitled **English to Bodo : Neural Machine Translation**, submitted by **Subhash Kumar Wary**, **Akher Uddin Ahmed**, **Birhang Borgoyary** and **Mohanji Prasad Sah (Roll Numbers: 202002022110**, **202102022100**, **202102023126**, **202102022111)** to the **Department of Computer Science and Engineering** is carried out under our direct supervisions and guidance.

The project work has been prepared as per the regulations of Central Institute of Technology Kokrajhar and we strongly recommend that this project work be accepted in partial fulfillment of the requirement for the degree of **Bachelor of Technology (B.Tech.)**.

Supervisor

Dr. Apurbalal Senapati
Assistant Professor (Department of Computer Science and Engineering)
Central Institute of Technology
Kokrajhar

Head of Department

Dr. Apurbalal Senapati
Department of Computer Science and Engineering
Central Institute of Technology
Kokrajhar

Date:
Place: Kokrajhar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

केन्द्रीय प्रौद्योगिकी संस्थान कोकराझार

CENTRAL INSTITUTE OF TECHNOLOGY KOKRAJHAR

(An Autonomous Institute under MHRD)

Kokrajhar-783370, BTR, Assam, India

CERTIFICATE BY THE BOARD OF EXAMINERS

This is to certify that the project work entitled **English to Bodo : Neural Machine Translation**, submitted by **Subhash Kumar Wary (Roll Number: 202002022110)**, **Akher Uddin Ahmed (Roll Number: 202102022100)**, **Birhang Borgoyary (Roll Number: 202102023126)** and **Mohanji Prasad Sah (Roll Number: 202102022111)**to the **Department of Computer Science and Engineering** of Central Institute of Technology Kokrajhar has been examined and evaluated.

The project work has been prepared as per the regulations of Central Institute of Technology Kokrajhar and qualifies to be accepted in fulfillment of the requirement for the degree of **Bachelor of Technology (B.Tech.)**.

Project Co-ordinator

External Examiner

ACKNOWLEDGEMENTS

We, the final-year B.Tech students of Computer Science and Engineering, sincerely express our gratitude to the esteemed faculty of our institute for their guidance and support throughout our project. We are especially thankful to Dr. Apurbalal Senapati and Mr. Sanjib Narzary, Assistant Professors, Department of Computer Science and Engineering, for their valuable mentorship, encouragement, and for providing us with essential resources that greatly contributed to our work. We also would like to acknowledge Mr. Mithun Karmakar, Assistant Professor and Project Coordinator, for maintaining coordination and discipline among all project groups. Our heartfelt thanks goes to our classmates, seniors, friends, and family, whose support, encouragement, and belief in us served as a constant source of motivation. The successful completion of this project is the result of the collective efforts and contributions of all these individuals, for which we are deeply grateful.

Date:

Subhash Kumar Wary
Roll Number: 202002022110

Akher Uddin Ahmed
Roll Number: 202102022100

Birhang Borgoyary
Roll Number: 202102023126

Mohanji Prasad Sah
Roll Number: 202102022111

ABSTRACT

Our work presents a Neural Machine Translation (NMT) system for English-to-Bodo translation which aimed at addressing the overall challenges of low resource languages. It is built using a custom encoder-decoder architecture enhanced with a Bahdanau style attention system. A bidirectional-GRU based encoder captures contextual representations of input sequences and also the GRU based decoder generates the output with dynamic attention over the source tokens. The model was trained on a real world English-to-Bodo parallel corpus and was evaluated using the BLEU metrics which achieved a BLEU-1 score of 14.64%. Despite with the limited data, our system demonstrates meaningful baseline performance and with further analysis which includes translation examples, loss trends, and output patterns offers insight into the model's behavior and limitations. Our work lays the groundwork for scalable improvements in Bodo language translation and contributes to ongoing research in low resource neural translation systems.

Keywords: • Neural Machine Translation (NMT) • Attention Mechanism • Low Resource Languages • Encoder-Decoder Architecture • Bodo Language Translation

Contents

1 Introduction **1**
- 1.1 Problem Statement . . . 2
- 1.2 Objectives . . . 2
- 1.3 Motivation . . . 2

2 Literature Review **3**
- 2.1 Overview of Machine Translation Paradigms . . . 3
- 2.2 Encoder–Decoder Architecture in NMT . . . 4
- 2.3 Attention Mechanism in Neural MT . . . 4
- 2.4 Neural Architectures for Low-Resource Translation . . . 5
- 2.5 Related Works on English–Bodo Translation . . . 5
- 2.6 Challenges in Low-Resource NMT . . . 6
- 2.7 Research Gap and Motivation . . . 6

3 Proposed System **8**
- 3.1 System Overview . . . 8
- 3.2 Architecture . . . 9
- 3.3 Methodology . . . 10
  - 3.3.1 Data Preparation . . . 10
  - 3.3.2 Model Building . . . 11
  - 3.3.3 Training Strategy . . . 12
  - 3.3.4 Evaluation Metrics . . . 13
- 3.4 Implementation . . . 14

4 Results and Analysis **16**
- 4.1 Training and Validation Loss . . . 16
- 4.2 BLEU Score Evaluation . . . 17
- 4.3 Translation Examples . . . 20
- 4.4 Analysis . . . 21

5 Conclusion & Future Works **23**
5.1 Conclusion 23
5.2 Future Works 24

References **25**

List of Figures

3.1 System Architecture of English-to-Bodo Neural Machine Translation . 9
3.2 Datasets - English and Bodo 10

4.1 Training and Validation Loss Curve 17
4.2 Word Level Evaluation of a Sentence 18
4.3 Word Level Evaluation of a Sentence 19
4.4 Sample Translation Outputs of the English–Bodo NMT Model 20
4.5 Interactive Translation Interface 21

List of Tables

2.1 Comparison of Related Works on Neural Machine Translation 6

4.1 BLEU Scores for English–Bodo Translation 18

Chapter 1

Introduction

Language serves as a vital medium for human communication and cultural expression. In an increasingly globalized world, the ability to automatically translate between languages has become essential for enabling cross-cultural interaction, access to information, and inclusive technology. Machine Translation (MT) systems are designed to bridge the linguistic gap by converting text or speech from one language into another. Over the past decade, MT has undergone a paradigm shift from traditional rule-based and statistical approaches to end-to-end neural models, yielding significant improvements in translation quality for high-resource languages such as English, Spanish, and Chinese.

However, low-resource languages—those with limited linguistic datasets, tools, and research support—continue to lag behind in terms of automated translation capabilities. One such language is Bodo, a Sino-Tibetan language primarily spoken in the Indian state of Assam. Despite being recognized as one of the 22 scheduled languages of India, Bodo remains underrepresented in the digital and computational linguistic landscape. Due to the lack of parallel corpora of the linguistic annotation tools, and the standardized orthography presents considerable challenges for developing effective translation systems for Bodo.

Our work addresses these challenges by proposing a Neural Machine Translation (NMT) model specifically for the English to Bodo language pair. Our model is based on the Sequence-to-Sequence (Seq2Seq) architecture which is enhanced with an attention mechanism (Bahdanau Attention) to improve alignment between source and target sequences. Compared to traditional MT methods, neural approaches offer superior flexibility and accuracy by learning rich contextual representations directly from data. This makes them especially valuable in low-resource settings, where handcrafted rules and statistical alignments often fail due to data sparsity.

The primary contributions of this work include the development of a custom encoder-decoder model using gated recurrent units (GRUs), integration of a Bahdanau-style

attention mechanism, and a detailed evaluation of the model's performance using real-world English–Bodo parallel corpora. The goal is to establish a foundational framework that not only demonstrates baseline translation capabilities but also paves the way for future advancements in low-resource MT research.

1.1 Problem Statement

The Bodo language lacks effective machine translation tools due to limited linguistic resources and parallel corpora. Traditional rule-based and statistical methods are inadequate in such low-resource settings. This creates a communication barrier between Bodo and English speakers, restricting digital access and information exchange. A neural machine translation system is essential to bridge this gap and promote language inclusion for Bodo-speaking communities.

1.2 Objectives

1. To construct a Neural Machine Translation model to translate English text into Bodo.
2. To train and validate the model using real-world parallel corpora.
3. To implement a Seq2Seq architecture enhanced by an attention mechanism for better performance.
4. To evaluate the model's performance using BLEU scores.

1.3 Motivation

Bodo, spoken by over 1.5 million people in Northeast India, is a constitutionally recognized language but remains underrepresented in modern language technologies. The lack of digital translation tools limits access to information, education, and communication for Bodo-speaking communities. Developing an accurate English–Bodo Neural Machine Translation (NMT) system is a crucial step toward promoting linguistic inclusivity and preserving cultural identity. Attention-based NMT models have shown promising results in other low-resource scenarios, making them a strong candidate for improving translation quality even with limited data. This work aims to explore such techniques to support the advancement of Bodo in the digital era.

Chapter 2

Literature Review

2.1 Overview of Machine Translation Paradigms

Machine Translation (MT) is a core subfield of Natural Language Processing (NLP) that focuses on automatically translating text between languages. The evolution of MT has progressed through three major paradigms: Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT).

Rule-Based Machine Translation (RBMT)

RBMT systems operate using a large set of handcrafted linguistic rules and bilingual dictionaries. These systems attempt to analyze the syntactic and semantic structure of the source language and apply rules for translating into the target language. While RBMT offers interpretability and deterministic behavior, it is labor-intensive to build and maintain, and often lacks generalization across domains or linguistic nuances.

Statistical Machine Translation (SMT)

SMT introduced a data-driven alternative by using aligned bilingual corpora to learn translation probabilities. A foundational model is the noisy channel framework, which estimates the most probable translation $\hat{t}$ of a source sentence s as:

$$\hat{t} = \arg\max_t P(t|s) = \arg\max_t P(s|t)P(t)$$

where $P(s|t)$ is the translation model and $P(t)$ is the target language model. Despite being a significant advancement, SMT struggled with phrase fragmentation, long-range dependencies, and required complex feature engineering.

Neural Machine Translation (NMT)

NMT revolutionized MT by utilizing deep neural networks to model translation as an end-to-end sequence learning problem. Introduced by Sutskever et al. (2014), early NMT systems used a sequence-to-sequence (Seq2Seq) architecture comprising recurrent neural networks (RNNs), where the encoder processes the input sentence into a vector representation, and the decoder generates the translation. NMT offers superior fluency and context understanding, particularly for high-resource languages.

2.2 Encoder–Decoder Architecture in NMT

The encoder–decoder framework underpins most NMT systems. The encoder transforms the input sentence into hidden states that encode contextual meaning, while the decoder generates the output sequence token by token, conditioned on the encoder output and previously generated tokens.

However, early Seq2Seq models used a fixed-length context vector that became a bottleneck for long or complex sentences. This issue was addressed by the introduction of attention mechanisms, enabling the decoder to dynamically attend to different parts of the input sequence during translation.

2.3 Attention Mechanism in Neural MT

The attention mechanism, first proposed by Bahdanau et al. (2015), allows the decoder to focus on specific parts of the source sentence during each decoding step. This improves alignment and context preservation in translation. The attention mechanism involves three key components:

Alignment Score

$$e_{t,i} = v^\top \tanh(W_1 h_i + W_2 s_{t-1})$$

where h_i is the hidden state of the encoder at position i, and s_{t-1} is the decoder hidden state from the previous time step.

Attention Weights

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}$$

These are computed using a softmax function and represent how much attention the decoder pays to each encoder hidden state.

Context Vector

$$c_t = \sum_i \alpha_{t,i} h_i$$

This weighted sum of encoder states is passed to the decoder at each step, enhancing its ability to generate accurate and fluent translations.

Luong et al. later proposed global and local attention models, and Vaswani et al. introduced the Transformer architecture, which employs self-attention and eliminates recurrence for improved parallelization.

2.4 Neural Architectures for Low-Resource Translation

In low-resource settings, where large parallel corpora are unavailable, simpler models often outperform more complex ones. Recurrent architectures such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks remain popular due to their efficiency and lower data requirements.

Gated Recurrent Units (GRUs)

GRUs are a variant of RNNs that use gating mechanisms to control the flow of information. The update equations for GRUs are:

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1}) \\ \tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned}$$

where z_t is the update gate, r_t is the reset gate, and $\odot$ denotes element-wise multiplication. GRUs offer a good trade-off between performance and computational efficiency.

2.5 Related Works on English–Bodo Translation

Various approaches to neural machine translation have been explored in literature, with significant advancements in model architecture and performance. The above table summarizes a few prominent works relevant to this study.

Table 2.1: Comparison of Related Works on Neural Machine Translation

Author(s)	**Year**	**Model/Method**	**BLEU Score**
Sutskever et al.	2014	Seq2Seq with LSTM	34.8 (En–Fr)
Luong et al.	2015	Global/Local Attention (LSTM)	25.9 (En–De)
Vaswani et al.	2017	Transformer (Self-Attention)	41.8 (En–Fr)
Vaswani et al.	2017	Transformer (Self-Attention)	28.4 (En–De)
Parvez et al.	2023	Transformer (English–Bodo)	11.01 (En–Bodo)
...	...	...	...

2.6 Challenges in Low-Resource NMT

Multiple studies emphasize the unique obstacles faced when working with low-resource language pairs:

1. **Limited Parallel Corpora:** Most available datasets for Bodo are domain-specific and contain fewer than 50,000 sentence pairs.
2. **Morphological Richness:** Bodo is highly inflected, increasing vocabulary size and making it harder for models to learn representations for all variants.
3. **High Out-of-Vocabulary (OOV) Rates:** OOV problems are frequent due to limited data and morphological complexity.
4. **Script and Preprocessing:** The use of Devanagari script necessitates custom tokenization and normalization tools.
5. **Tooling and Linguistic Resources:** The absence of foundational NLP tools for Bodo hampers both data preparation and linguistic analysis.

These factors make English–Bodo translation particularly challenging and justify the need for domain-general NMT systems using flexible architectures.

2.7 Research Gap and Motivation

While previous works demonstrate the feasibility of English–Bodo NMT, they are either domain-limited or poorly evaluated across general contexts. Moreover, BLEU scores remain low across the board, regardless of the model type. There is a clear lack of:

- General-purpose translation models for Bodo,
- Evaluation across diverse sentence structures,
- Scalable, reproducible baselines for future comparison.

Our work addresses these gaps by designing a GRU-based encoder-decoder architecture with Bahdanau attention, trained and evaluated on publicly available English–Bodo parallel corpora. It contributes a strong foundational system for further research and practical deployment in multilingual applications.

Chapter 3

Proposed System

3.1 System Overview

The proposed system is an end-to-end Neural Machine Translation (NMT) pipeline designed to translate English sentences into Bodo. It is built on a **Sequence-to-Sequence (Seq2Seq)** architecture augmented with an **attention mechanism**, enabling the model to effectively handle variable-length input sequences and focus on relevant contextual information during translation.

The system comprises two main components:

- A **GRU-based encoder** that processes the input English sentence and generates a sequence of hidden states representing its contextual information.
- A **GRU-based decoder** that generates the translated Bodo sentence one token at a time, guided by an attention mechanism that dynamically weighs the encoder's outputs.

This encoder-decoder framework with attention allows the model to learn alignments between source and target tokens, leading to more accurate and context-aware translations. The system is fully trainable and operates in an end-to-end fashion, requiring no manual feature engineering or rule-based design.

3.2 Architecture

The proposed Neural Machine Translation model follows an encoder-decoder architecture with an integrated attention mechanism. Each component of the system plays a critical role in processing the input sequence and generating an accurate translation.

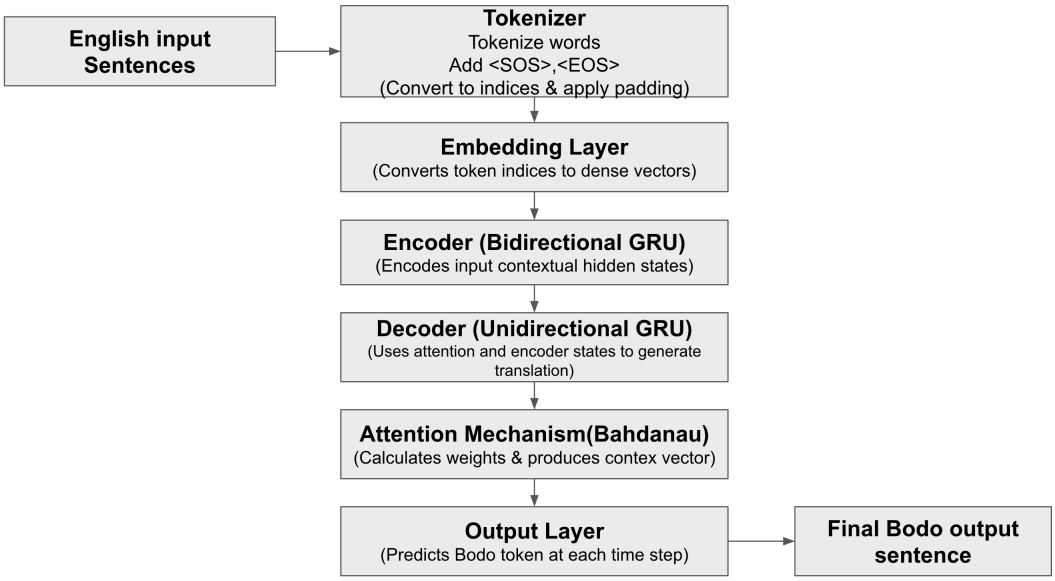

Figure 3.1: System Architecture of English-to-Bodo Neural Machine Translation

1. **Encoder:** The encoder is built using a two-layer Bidirectional GRU (Gated Recurrent Unit). It reads the input English sentence in both forward and backward directions, allowing the model to capture complete contextual information for each word. The outputs from both directions are concatenated to form a rich representation of the input.

2. **Decoder:** The decoder consists of a two-layer unidirectional GRU, which generates the translated Bodo sentence one token at a time. At each step, it takes the previous output token (or ground truth during training) and a context vector provided by the attention mechanism to predict the next token.

3. **Attention:** A Bahdanau-style additive attention mechanism is used to dynamically focus on relevant parts of the encoder's output at each decoding step. The attention module computes alignment scores between the current decoder state and each encoder hidden state, creating a weighted context vector that guides the decoder in generating the most appropriate translation.

3.3 Methodology

3.3.1 Data Preparation

This section describes the methodology adopted for building and training the English-to-Bodo Neural Machine Translation model, beginning with the data preparation process.

a) Dataset:

The dataset used for this project is a parallel corpus of English–Bodo sentence pairs. Each English sentence is aligned with its Bodo translation across corresponding lines in the provided files. The dataset is divided as follows:

- **Training Set:** 32,149 sentence pairs (`dev.eng` / `dev.brx`)
- **Validation Set:** 665 sentence pairs (`val.eng` / `val.brx`)
- **Test Set:** 444 sentence pairs (`tst.eng` / `tst.brx`)

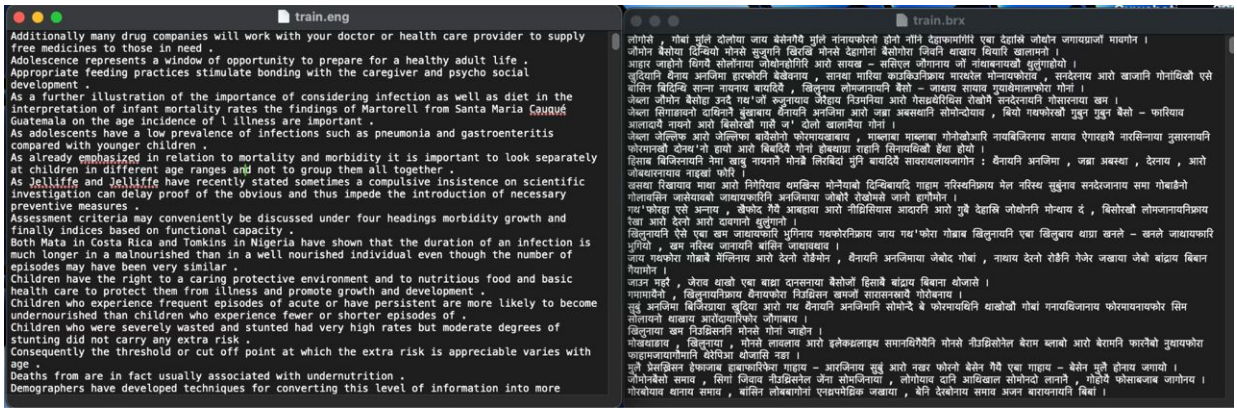

Figure 3.2: Datasets - English and Bodo

b) Preprocessing:

Preprocessing was performed as follows:

- **Tokenization:** Sentences were split into tokens using whitespace-based tokenization for both languages.
- **Vocabulary Construction:** Separate vocabularies were created for English and Bodo. Each vocabulary includes special tokens such as `<PAD>`, `<SOS>`, `<EOS>`, and `<UNK>`.
- **Indexing:** Tokens were mapped to integer indices using the corresponding vocabulary.

- **Padding:** Sentences were padded to a fixed maximum length to ensure consistent input sizes for batch training.
- **Framing Sequences:** `<SOS>` and `<EOS>` tokens were appended to the beginning and end of each sentence to guide the decoder during training and inference.

c) Data Split:
The dataset is divided into three non-overlapping sets to support effective model training, tuning, and evaluation:

- **Training Set (32,149 pairs):** Used to train the model by optimizing parameters.
- **Validation Set (665 pairs):** Used to tune hyperparameters and monitor model performance during training.
- **Test Set (444 pairs):** Used for final evaluation of translation quality using BLEU scores and translation examples.

3.3.2 Model Building

The translation model is built using a Sequence-to-Sequence (Seq2Seq) architecture composed of Gated Recurrent Unit (GRU) layers and a Bahdanau-style attention mechanism to enhance alignment and context awareness during decoding.

Encoder: The encoder is implemented using a two-layer **bidirectional GRU**. It processes the input English sentence from both forward and backward directions, generating a sequence of hidden states that represent contextual information for each token. These hidden states are concatenated and passed to both the attention mechanism and the decoder.

Decoder: The decoder is a two-layer **unidirectional GRU** that generates the translated Bodo sentence token by token. At each time step, the decoder receives the previous output token (or ground truth during training) and a context vector produced by the attention mechanism. This design allows the decoder to generate context-aware translations by selectively attending to relevant encoder states.

Attention Mechanism: A **Bahdanau-style additive attention** mechanism is used to compute a context vector at each decoding step. It calculates alignment scores between the decoder's current hidden state and each encoder hidden state. These scores are passed through a softmax function to obtain attention weights, which are used to compute a weighted sum of the encoder outputs. This context vector enables the decoder to dynamically focus on relevant portions of the input sentence.

GRU Cell:
The GRU is a simplified variant of the LSTM, using fewer gates and parameters while

retaining the ability to model long-term dependencies. The GRU cell operates according to the following equations:

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1}) \\ \tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \end{aligned} \tag{3.1}$$

Where:

- x_t is the input at time step t
- h_{t-1} is the previous hidden state
- z_t and r_t are the update and reset gates
- $\tilde{h}_t$ is the candidate hidden state
- $\odot$ denotes element-wise multiplication
- σ is the sigmoid activation function

This architecture effectively captures sequential dependencies and allows the model to perform well in low-resource translation tasks such as English-to-Bodo.

The attention mechanism computes alignment scores and context as:

$$e_{t,i} = v_a^\top \tanh(W_a s_{t-1} + U_a h_i), \tag{3.2}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}, \tag{3.3}$$

$$c_t = \sum_i \alpha_{t,i} h_i, \tag{3.4}$$

where s_{t-1} is the decoder state at time $t-1$, h_i are encoder states, and $\alpha_{t,i}$ are the attention weights. These equations follow the formulation of Bahdanau et al. and Luong et al. in attentional NMT.

3.3.3 Training Strategy

The training strategy was designed to effectively optimize the Seq2Seq model with attention for English-to-Bodo neural machine translation. The following techniques and configurations were employed:

1. **Training Setup:** The model was trained on preprocessed English-Bodo sentence pairs using a batch size of 64. The training was conducted on GPU when available, ensuring accelerated computation.

2. **Loss Function and Optimization:** Cross-entropy loss was used as the objective function, with padding tokens (`<PAD>`) ignored using `ignore_index` to prevent them from affecting gradient updates. The optimizer chosen was **Adam** with a learning rate of 0.001. A learning rate scheduler (`ReduceLROnPlateau`) was used to reduce the learning rate when the validation loss plateaued, helping improve convergence and avoid overfitting.

3. **Gradient Clipping:** To mitigate the issue of exploding gradients, the gradient norms were clipped at a maximum value of 1.0.

4. **Training Loop:** Each epoch included both training and evaluation phases. During training, teacher forcing was applied, allowing the decoder to receive the ground-truth token at each step. Evaluation was conducted without teacher forcing to simulate real inference conditions. Losses for both phases were logged for visualization and early stopping purposes.

5. **Early Stopping:** Early stopping was implemented using a patience parameter. If the validation loss did not improve for a defined number of consecutive epochs, training was halted to prevent overfitting. The best-performing model (lowest validation loss) was saved during training.

6. **Visualization:** Training and validation loss trends were plotted and saved for later analysis, providing insight into model convergence and overfitting behavior.

This training strategy ensured that the model was robustly optimized, monitored for generalization performance, and prevented from overtraining on the dataset.

3.3.4 Evaluation Metrics

The performance of the neural machine translation model is evaluated using the BLEU (Bilingual Evaluation Understudy) score, a standard metric for assessing the quality of machine-generated translations by comparing them to human references.

BLEU measures the overlap of n-grams between the candidate and reference translations, penalizing overly short outputs through a brevity penalty. It computes modified precision scores for n-grams up to a defined order (commonly 4).

The BLEU score is formally defined as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.5}$$

Where:

- p_n is the modified n-gram precision for n-grams of size n,

- w_n is the weight for each n-gram precision (usually $w_n = \frac{1}{N}$),
- BP is the brevity penalty, defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{3.6}$$

  where c is the length of the candidate translation and r is the length of the reference translation.

BLEU scores range from 0 to 100, with higher values indicating closer alignment between the generated and reference translations. In this work, BLEU-1 to BLEU-4 were computed to capture unigram to 4-gram matching quality.

3.4 Implementation

1. **Hardware Requirements :**

   The model was implemented and tested on a system with the following hardware specifications:

   - **Processor:** Intel Core i5 / i7 (8th Gen or above) or AMD Ryzen 5 equivalent
   - **RAM:** Minimum 8 GB (16 GB recommended for faster training and data handling)
   - **GPU:** NVIDIA GPU with CUDA support (e.g., GTX 1650 / RTX 2060 or higher)
   - **Storage:** At least 10 GB free disk space (for dataset, models, and outputs)
   - **Operating System:** MacOS, Windows 10/11

2. **Software Requirements :**

   The following software tools and libraries were used to implement and run the neural machine translation system:

   - **Programming Language:** Python 3.8 or above
   - **Development Environment:** Jupyter Notebook (via Anaconda or standalone)
   - **Libraries & Frameworks:**
     - **PyTorch** – for building and training neural networks
     - **NumPy** – for numerical operations

  - **TorchText** – for text preprocessing and dataset handling
  - **Matplotlib** – for visualizing training loss and BLEU scores
  - **NLTK** – for BLEU score computation and tokenization
  - **pickle** – for saving and loading tokenizers and model files
- **Package Manager:** `pip` or `conda`
- **Operating System:** Compatible with Windows, macOS, and Linux (Linux preferred for GPU training)

These software tools enabled efficient prototyping, training, evaluation, and visualization of the translation model in a modular and scalable manner.

3. **Dataset Used :**

- **Name of Dataset :** parallel eng brx tourism health
- **Source :** https://get.alayaran.com/parallel-data/

Chapter 4

Results and Analysis

4.1 Training and Validation Loss

Training a neural machine translation model involves minimizing the difference between predicted and actual target sequences. During the training phase, both training and validation loss were recorded over 20 epochs to evaluate how well the model learned and generalized.

Initially, the model exhibited relatively high training and validation loss values due to random initialization and lack of context understanding. However, by epoch 5, the training loss began to decrease steadily, and the validation loss followed a similar trend. This indicated that the model was learning meaningful patterns in the data.

Key observations:

- **Teacher Forcing:** Implemented during training, it enabled faster convergence by guiding the decoder with ground-truth outputs during sequence generation.
- **Padding Mask:** Helped exclude `<PAD>` tokens from loss computation, improving learning efficiency and accuracy.
- **Loss Graph:** A plot using `matplotlib` demonstrated a consistent downward slope for both loss curves, with minimal divergence between them.

The convergence of training and validation loss without overfitting is a strong indicator of good model generalization, especially given the limited size and scope of the dataset used.

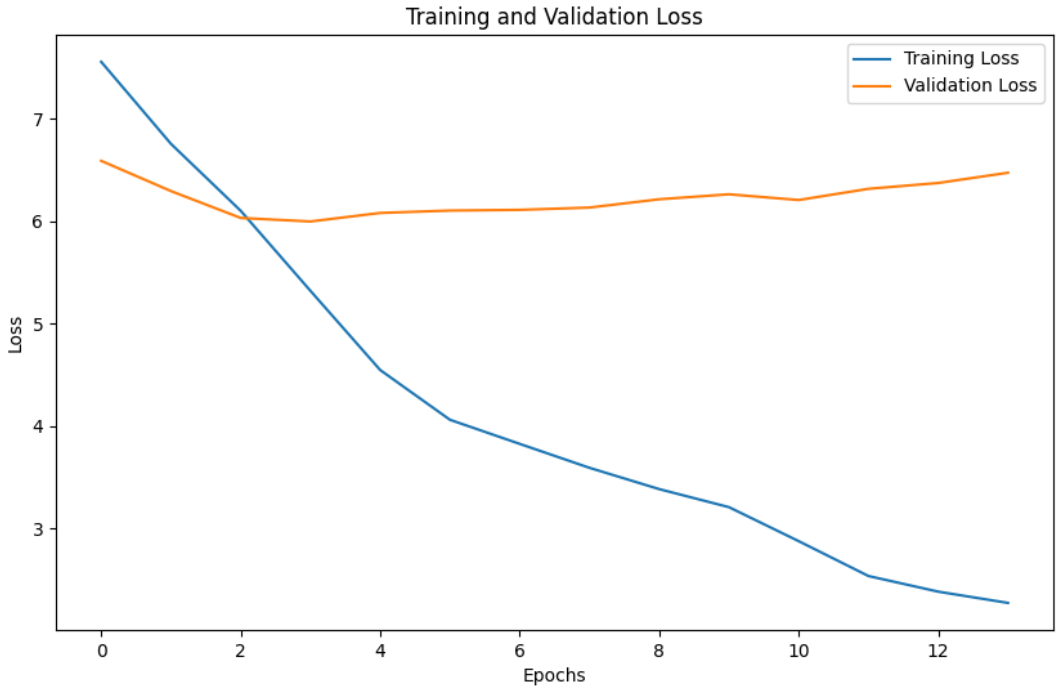

Figure 4.1: Training and Validation Loss Curve

4.2 BLEU Score Evaluation

The Bilingual Evaluation Understudy (BLEU) score is a widely used metric for evaluating the quality of machine translation systems by comparing machine-generated translations against one or more reference translations. BLEU evaluates the precision of n-grams (contiguous word sequences) in the predicted output, giving an insight into how similar the machine translation is to human reference translations.

In our work, BLEU-1 through BLEU-4 scores were computed to assess the performance of the trained English-to-Bodo neural machine translation model.

The results indicate a gradual drop in score with higher-order n-grams, which is expected due to the increased difficulty in matching longer sequences, especially in low-resource language settings. The BLEU-1 score of 14.64 suggests that the model captures a reasonable amount of unigram (single-word) matches between predictions and references. However, the low BLEU-4 score of 1.33 reveals that the model struggles with producing longer coherent phrases that match the reference translations exactly.

This performance is typical for initial models in low-resource language translation tasks where parallel corpus data is limited. The low BLEU-2 to BLEU-4 scores highlight the need for more training data, improved model architecture, or data augmentation techniques to enhance translation fluency and accuracy.

Despite these challenges, the BLEU scores provide a quantifiable starting point and demonstrate that the model has begun to learn useful word and phrase correspondences between English and Bodo.

Table 4.1: BLEU Scores for English–Bodo Translation

BLEU Score Type	**Score**
BLEU-1	14.64
BLEU-2	6.23
BLEU-3	2.89
BLEU-4	1.33

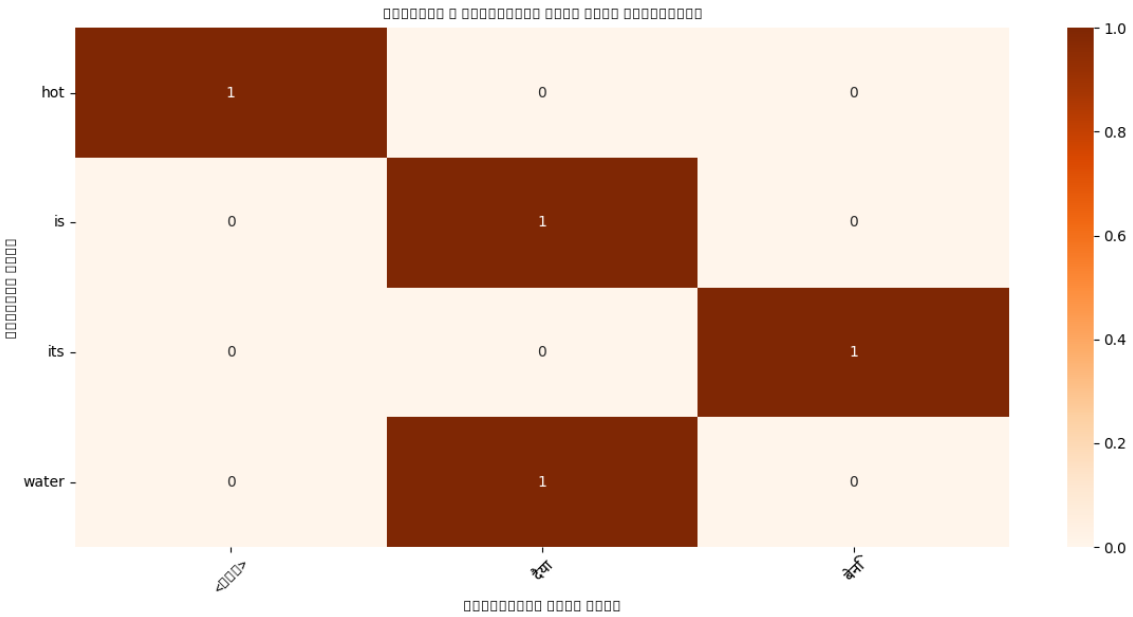

Figure 4.2: Word Level Evaluation of a Sentence

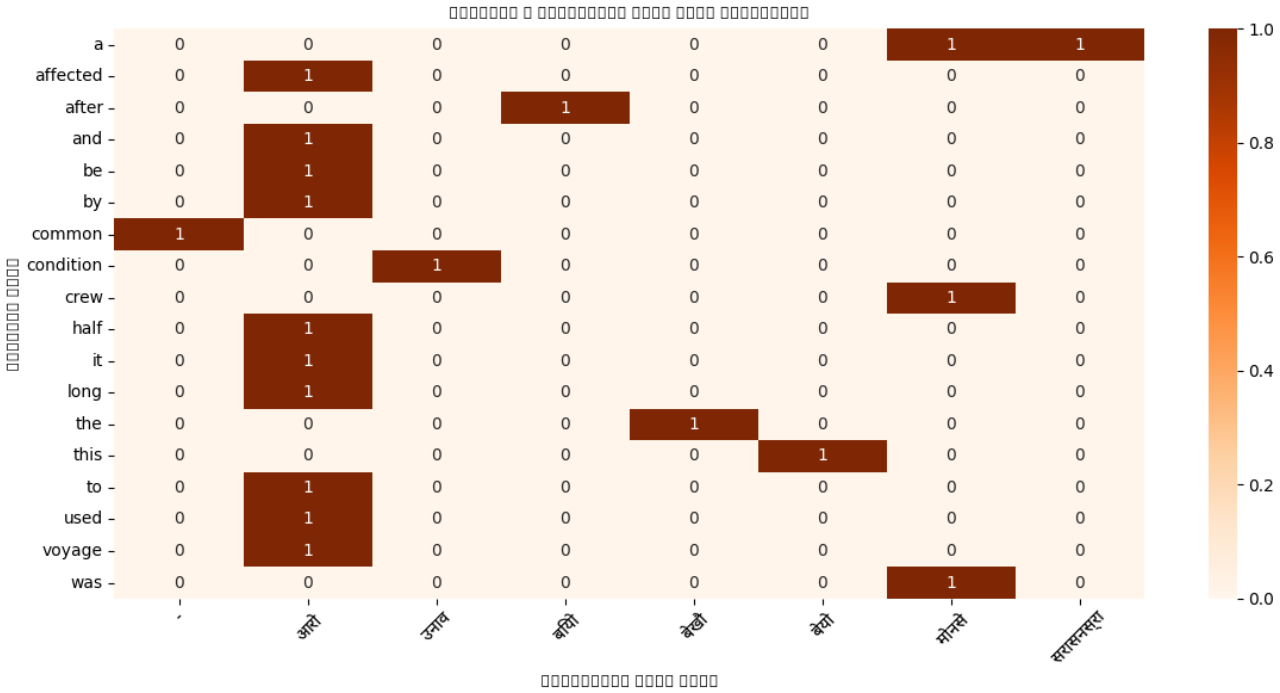

Figure 4.3: Word Level Evaluation of a Sentence

4.3 Translation Examples

The following examples illustrate the translation capabilities of the proposed model. Short and commonly used phrases are generally translated accurately, while longer or less frequent expressions may result in partial or incorrect translations.

Sl. No	English Sentence	Reference Translation (Bodo)	Model Prediction (Bodo)
1	Its water is hot.	बेनि दैया गुदुं ।	बेनि दैया दैया <UNK> <UNK>
2	Make your trip to Vaishno Devi a memorable event by staying at the right kind of accommodation.	नोंसोर थार रोखोमनि थानाय - खाबुवाव थानानै नोंनि बिष्नु देबिसिम दावबायनायखौ गोसोखांथाव जाथाय खालाम ।	नोंनि दावबायनायखौ नोंनि दावबायनायखौ नोंनि <UNK> नोंनि <UNK> <UNK> <UNK> <UNK>
3	This was a common condition after a long voyage and half the crew used to be affected by it.	बेयो गोलाव दिङादावबायनायनि उनाव सरासनस्रा थासारि आरो जाहाजमावहानजा खावसेया बेजों गोहोमखोखलैजादोंमोन ।	बेयो मोनसे सरासनस्रा - उनाव , बियो मोनसे आरो आरो आरो आरो बेखौ मोनसे आरो आरो आरो आरो आरो आरो आरो आरो ।
4	As a person sleeps the brain sends slower but larger and larger waves through the electroencephalogram which is used to measure it .	जेरै मानसिआ उन्दुयो , मेलेमा लासैसिन नाथाय बांसिन आरो बांसिन दैथुनफोर दैथाय - हरो (इलेकत्रएनसेफालग्राफजों जाय बेखौ सुयोमोन) ।	सासे मानसिया सासे मानसिया नाथाय नाथाय नाथाय नाथाय नाथाय नाथाय नाथाय ।
5	For most children lack of access to food is not the only cause of malnutrition .	गोबां गथफोरनि थाखाय , आदारखौ मोनजायैनि आंखालाल' मेल नीउथ्रिसननि जाहोन नङा ।	बांसिन गथफोरनि थाखाय , थाखाय , मानोना , मानोना नङा , नङा , नङा , नङा ।

Figure 4.4: Sample Translation Outputs of the English–Bodo NMT Model

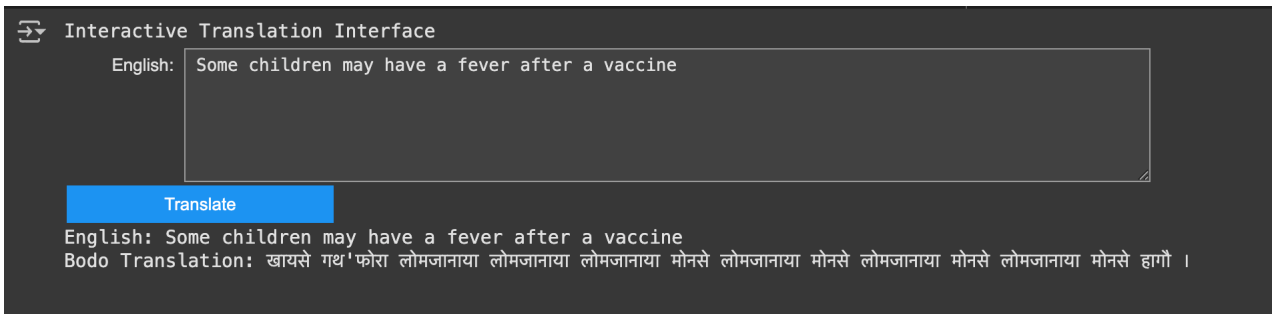

Figure 4.5: Interactive Translation Interface

4.4 Analysis

The results obtained from the English–Bodo Neural Machine Translation system reveal several insights into the model's performance, limitations, and potential for improvement.

1. Translation Quality: The model shows a basic level of translation capability, as evidenced by the BLEU-1 score of 14.64, indicating that the model can often predict individual words correctly. However, the rapidly declining scores for BLEU-2 to BLEU-4 suggest that the model struggles to generate fluent and syntactically correct multi-word sequences.

2. Error Patterns: Qualitative examination of translation outputs reveals common issues such as:

- Incomplete translations, where only a part of the source sentence is translated.
- Repetition or omission of key content words.
- Word-order errors, particularly with verb and object placements.
- Frequent use of generic or filler tokens due to vocabulary limitations or alignment confusion.

These patterns highlight the difficulty of translating between morphologically and syntactically divergent languages like English and Bodo, especially with a limited dataset.

3. Effect of Data Scarcity: The performance of the model is significantly impacted by the size and quality of the parallel corpus. With only a modest amount of training data available, the model lacks sufficient exposure to varied sentence structures, idioms, and grammatical constructs in both languages. This hinders generalization and translation fluency, especially for longer and more complex sentences.

4. Model Stability and Training Behavior: Throughout training, both training and validation loss decreased, indicating that the model was learning without overfitting significantly. The use of early stopping, learning rate scheduling, and gradient clipping helped maintain training stability. However, the limited improvement in

BLEU scores suggests that while the model learns to reconstruct frequent patterns, it still lacks linguistic depth.

5. Overall Evaluation: The implemented model provides a foundational baseline for English–Bodo neural machine translation. It demonstrates the feasibility of applying attention-based Seq2Seq models to low-resource translation tasks but also emphasizes the importance of additional resources.

Chapter 5

Conclusion & Future Works

5.1 Conclusion

This project explored the development of a Neural Machine Translation (NMT) system for English to Bodo—a low-resource language pair—using a Sequence-to-Sequence model with attention mechanism. The model was trained and evaluated using a manually prepared parallel corpus, and its performance was quantitatively assessed using BLEU scores and qualitatively through example translations.

The results show that while the model is able to learn basic word-level mappings, it faces challenges in generating fluent and semantically accurate multi-word translations. The BLEU-1 score of 14.64 reflects some success in capturing individual word correspondences, but the significantly lower BLEU-2 to BLEU-4 scores reveal limitations in grammatical fluency and contextual understanding.

Despite these limitations, the project successfully demonstrates the viability of applying modern NMT architectures to low-resource languages, and it establishes a foundational framework for further research. The process of building the dataset, designing the model architecture, and evaluating the system has provided valuable insights into both the technical and linguistic challenges of machine translation in underrepresented languages.

5.2 Future Works

There are several directions in which this work can be extended to enhance both translation quality and model robustness:

- **Data Augmentation:** One of the most critical steps forward is expanding the parallel corpus. Techniques such as back-translation, crowd-sourced translation, or crawling bilingual data from public sources can significantly improve training data diversity.
- **Transfer Learning:** Pretrained multilingual models such as mBART, mT5, or MarianMT can be fine-tuned for English–Bodo translation to leverage knowledge from high-resource language pairs and multilingual data.
- **Subword Tokenization:** Implementing subword units (e.g., Byte-Pair Encoding or SentencePiece) can help in better handling rare or morphologically complex words, which are especially common in Bodo.
- **Linguistic Incorporation:** Integrating syntactic and morphological information from Bodo grammar may help the model understand word structure and ordering better.
- **Human Evaluation:** In addition to automated metrics, human evaluation of translation quality—based on fluency, adequacy, and grammaticality—can provide more comprehensive feedback for iterative model improvement.
- **Web or Mobile Integration:** The model can be deployed as a simple web or mobile application to help promote the use and preservation of the Bodo language through practical tools.

Our work marks an important step toward advancing natural language processing for Bodo and similar low-resource languages. With continued efforts in data curation and model development, more accurate and fluent translations can be achieved, supporting language preservation and technological inclusivity.

References

[1] Shahnawaz Ayoub, Yonis Gulzar, Faheem Ahmad Reegu, and Sherzod Turaev. Generating image captions using bahdanau attention mechanism and transfer learning. *Symmetry*, 14(12):2681, 2022.

[2] Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, and Bidyut Kr Patra. Statistical machine translation for indic languages. *Natural Language Processing*, 31(2):328–345, 2025.

[3] Mikel L Forcada. Making sense of neural machine translation. *Translation Spaces*, 6(2):291–309, 2017.

[4] Amita Gupta. Global and local discourses in india's policies for early childhood education: policy borrowing and local realities. *Comparative Education*, 58(3):364–382, 2022.

[5] Yasir Abdelgadir Mohamed, Akbar Khanan, Mohamed Bashir, Abdul Hakim HM Mohamed, Mousab AE Adiel, and Muawia A Elsadig. The impact of artificial intelligence on language translation: a review. *Ieee Access*, 12:25553–25579, 2024.

[6] Mwnthai Narzary, Gwmsrang Muchahary, Maharaj Brahma, Sanjib Narzary, Pranav Kumar Singh, and Apurbalal Senapati. Bodo resources for nlp-an overview of existing primary resources for bodo. *AIJR Proceedings*, pages 96–101, 2021.

[7] Sanjib Narzary et al. Attention based english-bodo neural machine translation system for tourism domain. In *2019 International Conference on Computing Methodologies and Communication (ICCMC)*, pages 335–343. IEEE, 2019.

[8] Palanichamy Naveen and Pavel Trojovskỳ. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10), 2024.

[9] Margaret Dumebi Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.

[10] Dhrubajyoti Pathak, Sanjib Narzary, et al. Part-of-speech tagger for bodo language using deep learning approach. *Natural Language Processing*, 31(2):215–229, 2025.

[11] Dhrubajyoti Pathak, Sanjib Narzary, Sukumar Nandi, and Bidisha Som. Part-of-speech tagger for bodo language using deep learning approach. *Natural Language Processing*, 31(2):215–229, 2025.

[12] Sonali Sharma et al. Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics*, 12(7):1716, 2023.

[13] Bidisha Som, Rekha Kalita, and Ramesh Kumar Mishra. Culture cues facilitate object naming in both native and second language: evidence from bodo–assamese bilinguals. *Journal of Cultural Cognitive Science*, 2(1):45–57, 2018.

[14] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.