English to Bodo : Neural Machine Translation

Presented By -

Subhash Kumar Wary (202002022110)
Akher Uddin Ahmed (202102022100)
Birhang Borgoyary (202102023126)
Mohanji Prasad Sah (202102022111)

Project Supervisor -

Dr. Apurbalal Senapati

Department of Computer Science & Engineering

Central Institute of Technology Kokrajhar

Outline

1. Introduction
2. Problem Statement & Objectives
3. Background: Machine Translation Techniques
4. Comparision: RBMT, SMT and NMT - Advantages and Disadvantages
5. Related Work
6. Challenges
7. Model Architecture
8. Data Preparation
9. Model Building
10. Training Strategy
11. Evaluation Metrics
12. Results and Analysis
13. Conclusion and Future Work
14. References

Gantt Chart - Timeline of our Project Work

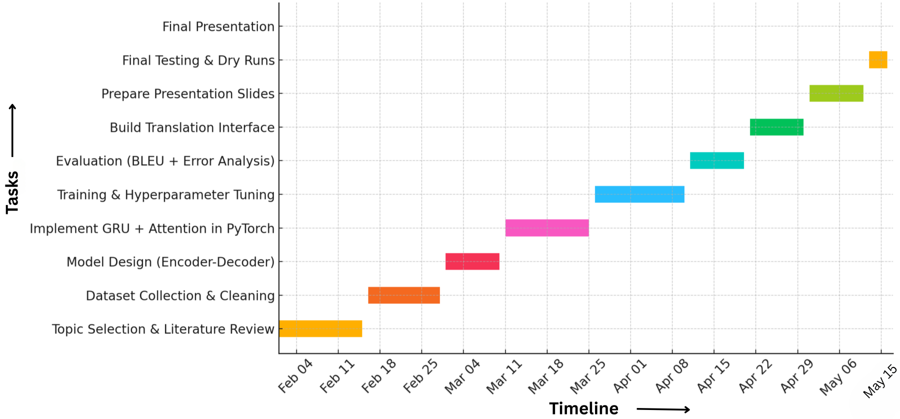


Figure 1: Gantt Chart - English to Bodo Nural Machine Translation Work

Introduction

- Bodo is spoken by over 1.5 million people but lacks MT tools.
- Rule-based and SMT approaches are insufficient.
- Neural Machine Translation (NMT) can leverage limited parallel data to improve translation quality.
- Our work implements an attention-based NMT system for English to Bodo.

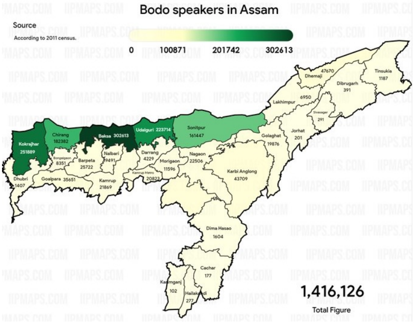


Figure 2: Bodo Language Speaking Region

Problem Statement & Objectives

Problem Statement:

- Scarcity of parallel corpora for English–Bodo.
- Existing tools fail to capture linguistic complexity.
- Need for context-aware, scalable translation models.

Objectives:

- Build an NMT model for English–Bodo using GRU and attention.
- Train on real parallel corpus from tourism and health domains.
- Evaluate performance using BLEU scores.

Background: Machine Translation Techniques

- **Rule-Based MT (RBMT):** Uses linguistic rules manually written; lacks flexibility.
- **Statistical MT (SMT):** Learns from aligned text data; suffers with limited data.
- **Neural MT (NMT):** Uses deep learning; captures context better through encoder-decoder architecture.

Comparision: RBMT, SMT and NMT - Advantages and Disadvantages

Approach	Advantages	Disadvantages	Uses
Rule-Based MT (RBMT)	Transparent, fully controllable, no training data needed	Labor-intensive, inflexible, hard to scale across languages	Early MT systems (e.g., Systran), legal/medical texts, government use
Statistical MT (SMT)	Learns from bilingual corpora, adaptable to domains, customizable	Needs large data, poor context handling, may be disfluent	Pre-2016 Google Translate, domain-specific translation, moderate data setups
Neural MT (NMT)	Captures long-range context, fluent output, supports fine-tuning	Data and compute intensive, black-box nature, hard to debug	Current systems (e.g., Google, DeepL), low-resource MT, real-time apps

Related Work

- **Sutskever et al. (2014):** Introduced Seq2Seq with LSTM; BLEU: 34.8 (En–Fr)
- **Luong et al. (2015):** Integrated attention in NMT; BLEU: 25.9 (En–De)
- **Vaswani et al. (2017):** Transformer model; BLEU: 41.8 (En–Fr)
- **Parvez et al. (2023):** Transformer for English–Bodo; BLEU: 11.01
- **...**

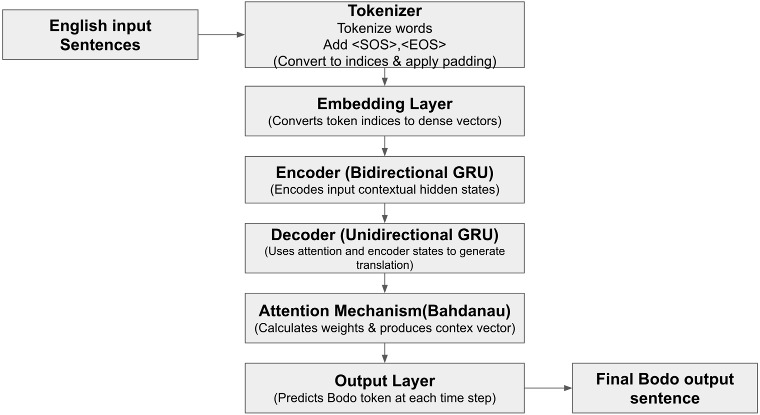
Challenges in Low-Resource MT

- **Data Scarcity:** Limited parallel English–Bodo corpora.
- **Domain Bias:** Available data is mainly from tourism and health.
- **Morphological Complexity:** Bodo has rich morphology with agglutinative structure.
- **OOV and Script Issues:** Handling unknown tokens and Devanagari script variations.

Model Architecture

- **Architecture:** Sequence-to-Sequence model with Attention.
- **Main Components are :**
- **Encoder:** Bidirectional GRU - Processes input sentence from both directions.
- **Decoder:** Unidirectional GRU - Generates translation word by word using hidden states.
- **Attention Mechanism:** Helps decoder focus on relevant source tokens.

Model Architecture



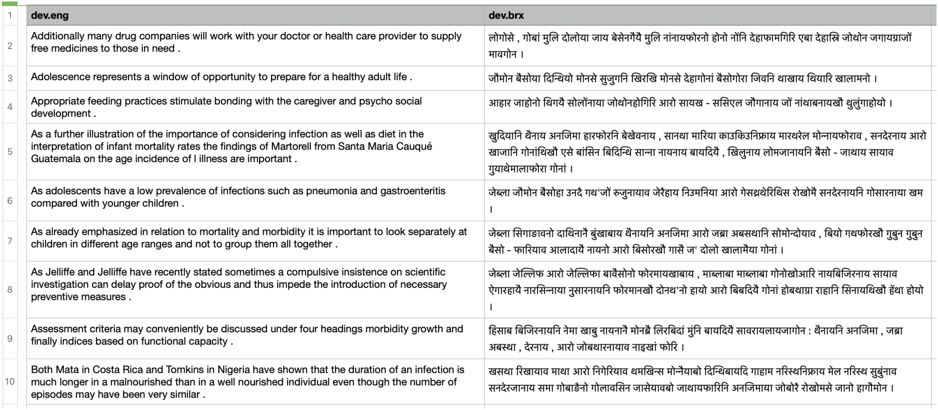
English input Sentences
Tokenizer
Tokenize words
Add <SOS>,<EOS>
(Convert to indices & apply padding)
Embedding Layer
(Converts token indices to dense vectors)
Encoder (Bidirectional GRU)
(Encodes input contextual hidden states)
Decoder (Unidirectional GRU)
(Uses attention and encoder states to generate translation)
Attention Mechanism(Bahdanau)
(Calculates weights & produces contex vector)
Output Layer
(Predicts Bodo token at each time step)
Final Bodo output sentence

Figure 3: System Architecture of English-to-Bodo Neural Machine Translation

Data Details and Preparation

- **Dataset:** 33,258 (Volume) sentence pairs from ALAYARAN corpus. (Source-https://get.alayaran.com/parallel-data/)
- **Splits:** Training (32,149), Validation (665), Test (444).
- **Steps Performed**: Tokenization, Indexing, Padding
- **Tokenization**: Split sentences into tokens
- **Indexing**: Convert tokens to numerical indices.
- **Padding**: Making all sentences the same length for batch processing.
- **Special Tokens**:
  `<SOS>`– start of sentence.
  `<EOS>`– end of sentence.
  `<PAD>`– padding.
  `<UNK>` – unknown/rare words.

Dataset Details



1	dev.eng	dev.brx
2	Additionally many drug companies will work with your doctor or health care provider to supply free medicines to those in need .	लोगोसे , गोबां मुलि दोलोया जाय बेसेनगैयै मुलि नांनायफोरनो होनो नोंनि देहाफामगिरि एबा देहास्रि जोथोन जगायग्राजों मावगोन ।
3	Adolescence represents a window of opportunity to prepare for a healthy adult life .	जौमोन बैसोया दिन्थियो मोनसे सुजुगनि खिरखि मोनसे देहागोनां बैसोगोरा जिवनि थाखाय थियारि खालामनो ।
4	Appropriate feeding practices stimulate bonding with the caregiver and psycho social development .	आहार जाहोनो थिगयै सोलोंनाया जोथोनहोगिरि आरो सायख - ससिएल जौगानाय जों नांथाबनायखौ थुलुंगाहोयो ।
5	As a further illustration of the importance of considering infection as well as diet in the interpretation of infant mortality rates the findings of Martorell from Santa Maria Cauqué Guatemala on the age incidence of I illness are important .	खुदियानि थैनाय अनजिमा हारफोरनि बेखेवनाय , सानथा मारिया काउकिउनिफ्राय मारथरेल मोन्नायफोराव , सनदेरनाय आरो खाजानि गोनांथिखौ एसे बांसिन बिदिन्थि सान्ना नायनाय बायदियै , खिलुनाय लोमजानायनि बैसो - जाथाय सायाव गुयाथेमालाफोरा गोनां ।
6	As adolescents have a low prevalence of infections such as pneumonia and gastroenteritis compared with younger children .	जेब्ला जौमोन बैसोहा उनदै गथ'जों रुजुनायाव जेरैहाय निउमनिया आरो गेसथ्रथेरिथिस रोखोमै सनदेरनायनि गोसारनाया खम ।
7	As already emphasized in relation to mortality and morbidity it is important to look separately at children in different age ranges and not to group them all together .	जेब्ला सिगाङावनो दाथिनानै बुंखाबाय थैनायनि अनजिमा आरो जब्रा अबसथानि सोमोन्दोयाव , बियो गथफोरखौ गुबुन गुबुन बैसो - फारियाव आलादायै नायनो आरो बिसोरखौ गासै ज' दोलो खालामैया गोनां ।
8	As Jelliffe and Jelliffe have recently stated sometimes a compulsive insistence on scientific investigation can delay proof of the obvious and thus impede the introduction of necessary preventive measures .	जेब्ला जेल्लिफ आरो जेल्लिफा बावैसोनो फोरमायखाबाय , माब्लाबा माब्लाबा गोनोखोआरि नायबिजिरनाय सायाव ऐगारहायै नारसिन्नाया नुसारनायनि फोरमानखौ दोनथ'नो हायो आरो बिबदियै गोनां होबथाग्रा राहानि सिनायथिखौ हेंथा होयो ।
9	Assessment criteria may conveniently be discussed under four headings morbidity growth and finally indices based on functional capacity .	हिसाब बिजिरनायनि नेमा खाबु नायनानै मोनब्रै लिरबिदां मुंनि बायदियै सावरायलायजागोन : थैनायनि अनजिमा , जब्रा अबस्था , देरनाय , आरो जोबथारनायाव नाइखां फोरि ।
10	Both Mata in Costa Rica and Tomkins in Nigeria have shown that the duration of an infection is much longer in a malnourished than in a well nourished individual even though the number of episodes may have been very similar .	खसथा रिखायाव माथा आरो निगेरियाव थमखिन्स मोन्नैयाबो दिन्थिबायदि गाहाम नरिस्थनिफ्राय मेल नरिस्थ सुबुंनाव सनदेरजानाय समा गोबाङैनो गोलावसिन जासेयावबो जाथायफारिनि अनजिमाया जोबोरै रोखोमसे जानो हागौमोन ।

Figure 4: Datasets - English and Bodo

Model Building

- Implemented using **PyTorch** Framework.
- **Embedding Layer:** Converts token indices to dense vectors.
- **Encoder:** Bidirectional GRU processes input sequence.
- **Decoder:** Unidirectional GRU generates Bodo sentence.
- **Bahdanau Attention:** Learns soft alignment between source and target tokens.
- **Output Layer:** Maps decoder outputs to vocabulary probabilities.

Gated Recurrent Unit (GRU)

GRU Overview:

- Efficient alternative to LSTM with fewer parameters.
- Handles vanishing gradients better than standard RNNs.

GRU Equations:

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

- z_t: Update gate – how much of past state to retain.
- r_t: Reset gate – how much of the past state to forget.
- $\tilde{h}_t$: Candidate Hidden State - New memory created based on current input and reset gate
- h_t: Final Hidden State – Final hidden state combining old and new memory.

Training Strategy

- **Batch Size:** 64
- **Optimizer:** Adam with learning rate = 0.001
- **Loss Function:** Cross-entropy loss (ignoring `<PAD>` tokens)
- **Hardware:** Trained on GPU for efficiency
- **Regularization Techniques:**
  - **Teacher Forcing** : Decoder uses correct previous word as input.
  - **Early Stopping**: Stops training if validation loss doesn't improve.
  - **Learning Rate Scheduler**: Reduces learning rate when validation loss stops improving.
  - **Validation Monitoring**: Tracks both training and validation loss.

Evaluation Metrics

- **BLEU Score (Bilingual Evaluation Understudy):**
  - Measures n-gram overlap between candidate and reference translations.
  - Ranges from 0 to 100 (higher is better).
  - We used BLEU-1 to BLEU-4 for evaluation.
- **BLEU Formula:**

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

- **BP (Brevity Penalty):**

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

where c: candidate length, r: reference length

BLEU Score Results

- **BLEU-1:** 14.64
- **BLEU-2:** 6.23
- **BLEU-3:** 2.89
- **BLEU-4:** 1.33

Observation: Higher n-grams show lower scores due to limited data and phrase errors. BLEU-1 indicates basic word alignment is learned.

Training and Validation Loss

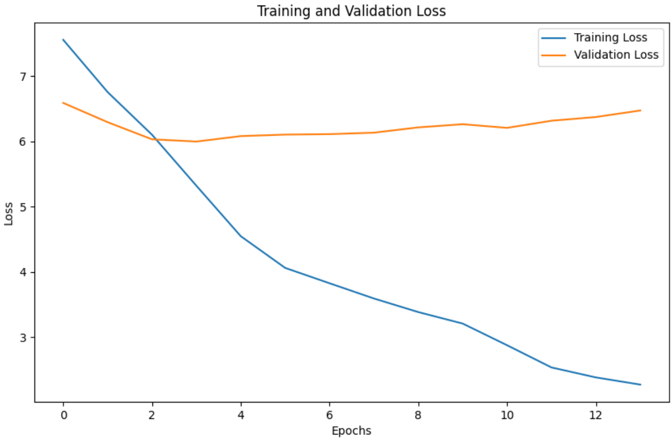


Figure 5: Training and Validation Loss Curve

Translation Sample

- **Simple Sentences:** Mostly Translated accurately.
- **Complex Sentences:** Partial or missing translations.
- **Common Errors:**
  - Repetition of words
  - Omission of key phrases
  - Handling unknown tokens `<UNK>`

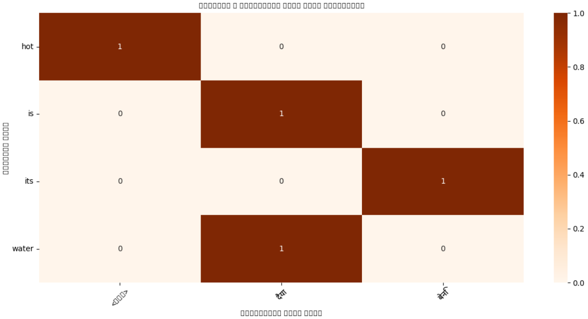
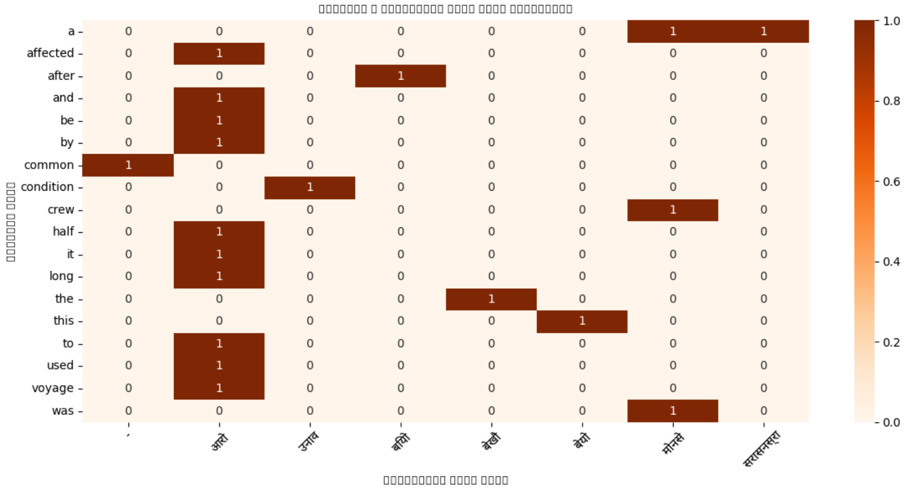


Figure 6: Word Level Evaluation of a Sentence

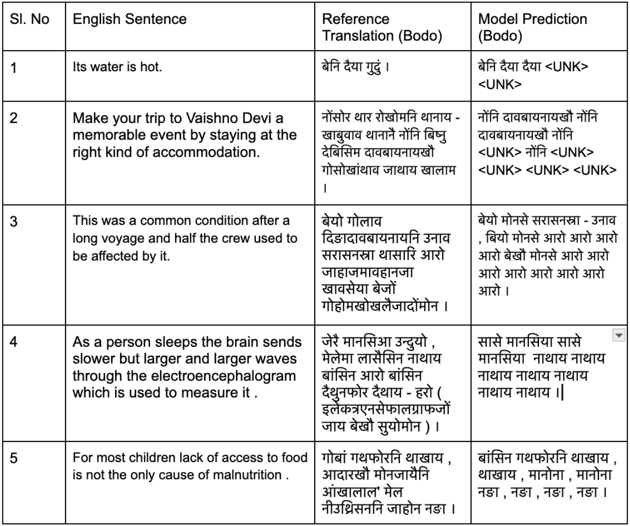
Translation Sample



	,	आरो	उनाव	बिथो	बेखौ	बेयो	मोनसे	सरासनस्रा
a	0	0	0	0	0	0	1	1
affected	0	1	0	0	0	0	0	0
after	0	0	0	1	0	0	0	0
and	0	1	0	0	0	0	0	0
be	0	1	0	0	0	0	0	0
by	0	1	0	0	0	0	0	0
common	1	0	0	0	0	0	0	0
condition	0	0	1	0	0	0	0	0
crew	0	0	0	0	0	0	1	0
half	0	1	0	0	0	0	0	0
it	0	1	0	0	0	0	0	0
long	0	1	0	0	0	0	0	0
the	0	0	0	0	1	0	0	0
this	0	0	0	0	0	1	0	0
to	0	1	0	0	0	0	0	0
used	0	1	0	0	0	0	0	0
voyage	0	1	0	0	0	0	0	0
was	0	0	0	0	0	0	1	0

Figure 7: Word Level Evaluation of a Sentence

Sample Translation Output



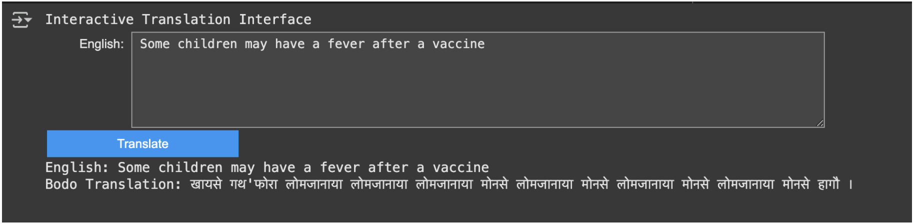
Sl. No	English Sentence	Reference Translation (Bodo)	Model Prediction (Bodo)
1	Its water is hot.	बेनि दैया गुदुं ।	बेनि दैया दैया <UNK> <UNK>
2	Make your trip to Vaishno Devi a memorable event by staying at the right kind of accommodation.	नोंसोर थार रोखोमनि थानाय - खाबुवाव थानानै नोंनि बिष्नु देबिसिम 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 8: Sample Translation Outputs of the English–Bodo NMT Model

Analysis

- Model learned basic English–Bodo word alignment.
- Sentence fluency is limited due to dataset size.
- Attention mechanism improves performance over simple encoder–decoder.
- Training was stable across 15 epochs.

Interface



English: Some children may have a fever after a vaccine

```
English: Some children may have a fever after a vaccine
Bodo Translation: खायसे गथ'फोरा लोमजानाया लोमजानाया लोमजानाया मोनसे लोमजानाया मोनसे लोमजानाया मोनसे लोमजानाया मोनसे हागौ ।
```

Figure 9: Interactive Translation Interface

Conclusion and Future Work

Conclusion:

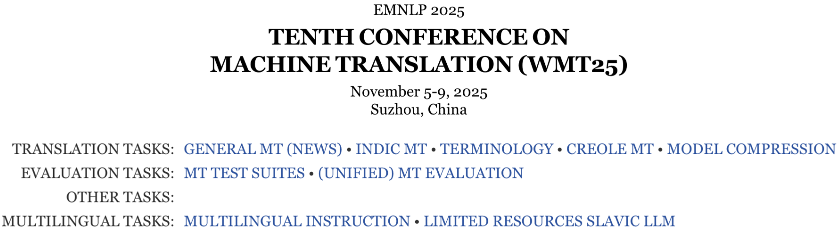
- Built a functional NMT model for English–Bodo.
- Demonstrated feasibility in low-resource conditions.
- BLEU scores provide a baseline for future improvement.

Future Work:

- Add back-translation and data augmentation.
- Integrate subword tokenization (BPE).
- Use pretrained multilingual models (e.g., mBART, mT5).
- Develop user-facing applications (web/mobile).
- Include human evaluation metrics.

WMT25 (Conference) Participating in the Shared Tasks

The shared tasks conference will feature scientific papers on topics related to Machine Translation models in a Low-Resource Indic Language Translation Systems.



EMNLP 2025

TENTH CONFERENCE ON
MACHINE TRANSLATION (WMT25)

November 5-9, 2025
Suzhou, China

TRANSLATION TASKS: GENERAL MT (NEWS) • INDIC MT • TERMINOLOGY • CREOLE MT • MODEL COMPRESSION
EVALUATION TASKS: MT TEST SUITES • (UNIFIED) MT EVALUATION
OTHER TASKS:
MULTILINGUAL TASKS: MULTILINGUAL INSTRUCTION • LIMITED RESOURCES SLAVIC LLM

WMT25 - Official Website (https://www2.statmt.org/wmt25/)

References

- Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V 2014 – Advances in neural information processing systems
- Bahdanau et al., 2015 – Neural Machine Translation by Jointly Learning to Align and Translate
- Vaswani et al., 2017 – Attention Is All You Need
- Narzary et al., 2019 – Attention Based English-Bodo Neural Machine Translation
- Som et al., 2025 – Part-of-Speech Tagger for Bodo Language
- Naveen, Palanichamy and Trojovsk, Pavel 2025 – Overview and challenges of machine translation for contextually appropriate translations
- Dataset Source - https://get.alayaran.com/parallel-data/ (Accessed on Feb 20 2025)

Team and Acknowledgment

Team Members:
Subhash Kumar Wary
Akher Uddin Ahmed
Birhang Borgoyary
Mohanji Prasad Sah

Supervisor: Dr. Apurbalal Senapati

Department of Computer Science and Engineering
Central Institute of Technology Kokrajhar
May 2025

Thank You!
Q&A