Journal of Integrated
SCIENCE & TECHNOLOGY

Machine translation of low resource Indian language using deep learning approach

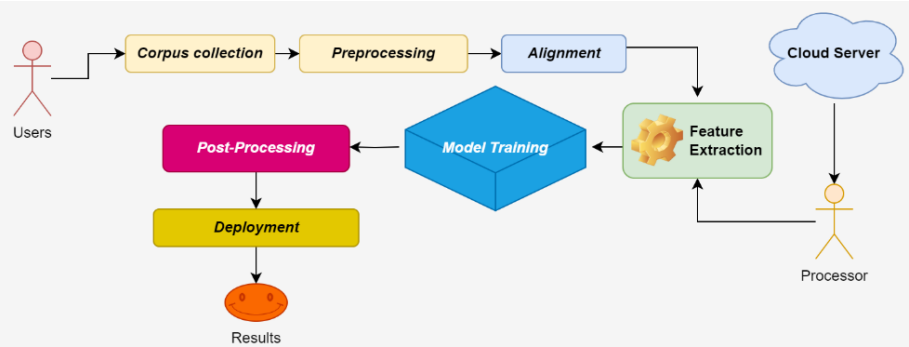
Madhuri Tayal,[1*] Aniket Tiwari,[2] Anuj Dharme,[3] Pratik K Agrawal,[4] Animesh Tayal,[5] Nilima V. Pardakhe[6]

[1]Department of Data Science, G.H. Raisoni College of Engineering and Management, Nagpur, Nagpur, India. [2] USI, AI and Data Department, Deloitte, Pune, India. [3]IBM, Pune, India. [4]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India. [5]Codemate IT Services Pvt LTD, Nagpur. India. [6]Department of Computer Science & Engineering, Prof. Ram Meghe Institute of technology & Research, Badnera, India.

Submitted on: 01-Dec-2024, Accepted and Published on: 09-Apr-2025

Article

ABSTRACT



The creation of an English-to-Bhojpuri machine translation (MT) system is presented in this paper, with an emphasis on the difficulties in translating words based on Devanagari script. Due to the sparse occurrence of distinct Devanagari words and the lack of training data, accurate translation is challenging in Bhojpuri, a language spoken in Bihar, India. A sequence-to-sequence model trained on a self-made dataset of 10,105 English-Bhojpuri sentence pairs is used to construct the system. Particular focus is placed on adding examples of these distinct words to the dataset. The model uses word embeddings and attention mechanisms to capture the semantic and contextual relationships required for precise translation. This machine translation (MT) system provides a customized solution for handling Indian languages in the Devanagari script by tackling the linguistic complexities of Devanagari words, improving the accuracy and fluency of English-to-Bhojpuri translation.

Keywords: Machine Translation, sparse occurrences, Neural network, Bahdanau attention mechanism

INTRODUCTION

In the ever-changing world of technology, where efficient communication is essential for business growth, overcoming language hurdles is imperative. The development of translation technologies is essential to enable smooth communication across linguistic boundaries. This project's main goal is to develop a system that uses natural language processing (NLP) to translate sentences from a variety of source languages to a target language. This effort focuses on Bhojpuri, a language that is mostly spoken in the Indian state of Bihar. The technical uses of machine translation from English to Bhojpuri include data analysis, education, localization, content translation, cross-lingual communication, and preservation of language. This project promotes efficient language exchange, improves accessibility, and enriches linguistic diversity and intercultural understanding.

Translation Services USA and Google Translate are two of the many translation companies that presently provide English to Bhojpuri translation services. Even while these services work effectively in most cases, there are problems when there are uncommon words in the text. In particular, uncommon English terms could be mistranslated or understood incorrectly semantically in the process of translation. Rare words in English are misinterpreted during translation and are either transliterated or semantically incorrect.

In Figure 1, we can see that Bhojpuri Translation consist Transliteration of words such as accessories, fashion and trend instead of their correct translation. This research paper focuses on the Bhojpuri Translation of Rare English words whose translation is not available easily. The reason behind this is the lack of training data for these words specifically and as result there are different types of error in the Translation performed by these such as Semantic, Lexical and Syntactic.

*Corresponding Author: Dr. Madhuri Tayal, Aniket Tiwari, Anuj Dharme, Pratik Agrawal
Tel: +91-8956113651, +91-8668292293, +91-8450967265
Email: madhuri.tayal@gmail.com, anitiw102@gmail.com, anujdharme15@gmail.com
pratik.agrawaal@gmail.com, nilimapardakhe@gmail.com

Cite as: *J. Integr. Sci. Technol., 2025, 13(6), 1127.*
URN:NBN:sciencein.jist.2025.v13.1127
DOI:10.62110/sciencein.jist.2025.v13.1127

Figure 1. Example of Translation by Google Translate

RELATED WORK

Several research works have been carried out in the domain of Machine Translation. Since our work is supposed to be low resource, A. Banerjee et.al.[1] explained the English-Marathi translation and methods for overcoming the difficulties of low-resource Neural Machine Translation (NMT). A variety of strategies are investigated, including enhancing parallel corpora, utilizing transfer learning, and utilizing Phrase Table Injection (PTI), back-translation, combining language corpora, pivoting, and multilingual embeddings. A comparison to a baseline trans- former model reveals a considerable increase in BLEU ratings. The paper gives first findings on Translation Edit Rate (TER) as a measure of human effort reduction in MT post-editing and includes detailed manual and automated evaluations.

The study by V. Mujadia et.al.[2] presented Hindi-Marathi translation utilizing attention-based recurrent neural network architectures. In order to improve machine translation, the authors looked at language elements including morph and POS as well as reverse translation. They performed morph + BPE segmentation for preprocessing and used parallel and monolingual corpora. The promise of linguistic feature-driven NMT for low-resource comparable language translation was highlighted by the positive results obtained while training their models with a sequence-to-sequence NMT framework.

The study by V. Goyal et.al.[3] is similar to what we were aiming for as their study included one of the approaches using rarer sentence datasets too. In this work, neural machine translation was used to test phrase pairings in English and Hindi. They showed that simpler translation approaches, such as word embedding and byte pair encoding, produced superior outcomes for Indian languages when used in neural machine translation. The suggested model produced results that were adequate and exceeded earlier attempts at neural machine translation from English to Hindi.

In order to achieve a 15-point increase in the BLEU score without the addition of new data and a 20-point increase when incorporating a language pair with abundant resources, the work by P. Madaan et.al.[4] presented a straightforward data augmentation technique coupled with a multilingual transformer. This approach performs on par with or better than the best models that were submitted for the Shared Task, and it demonstrates a notable improvement over a standard multilingual transformer. These results show that the amount of data used affects multilingual transformers' sensitivity and show that the suggested augmentation method can significantly raise BLEU scores.

The study by A. Das et.al.[5] examines the Bengali-Hindi language pair's performance using an attention-based neural machine translation system. For comparison with the NMT model, MOSES, a phrase-based SMT model, served as the baseline system. When it comes to BLEU scores, the attention-based NMT model performs noticeably better than MOSES.

In order to facilitate neural machine translation, the study by M.T. Luong et.al.[6] suggests two straightforward and efficient attentional mechanisms: the global approach, which continuously attends to all source positions, and the local approach, which concentrate on a subset of source positions at a time. The WMT translation tasks in both directions between English and German are used to evaluate the efficacy of these models.[7] For both WMT'14 and WMT'15, the ensemble model sets new state-of-the-art results in the English-to-German translation direction.

R. Sennrich et.al.[8] reports the proof that neural machine translation systems can handle translation from an open vocabulary by representing uncommon and unknown words as a series of sub word units. This method works better and simpler than a back-off translation model. For word segmentation, a new byte pair encoding variant is presented that can encode open vocabularies with a compact symbol vocabulary of variable-length subword units. The analysis shows that the baseline NMT system performs poorly when translating not only out-of-vocabulary words but also uncommon in-vocabulary words. Performance can be improved by decreasing the vocabulary size of subword models.[9]

The study by P. Zaremoodi et.al.[10] addresses the problem of neural machine translation for language pairings with limited parallel training text by proposing a multi-task learning strategy. To enhance translation quality, monolingual linguistic resources are used on the source side. In particular, auxiliary tasks like named-entity identification, syntactic parsing, and semantic parsing are used to scaffold the machine translation process. By successfully incorporating syntactic and semantic knowledge into the translation model, this technique lessens the need for a lot of parallel material. English-to-French, English-to-Farsi, and English-to-Vietnamese translation assignments are used to empirically assess the efficacy of this method.

The research by Q. Wang et.al.[11] investigates the efficacy of a genuinely deep Transformer model in machine translation. Wider networks (Transformer-Big) and deeper language representations are the two primary research avenues that have been sought to improve Transformer-based models; nevertheless, the latter strategy has difficulties when it comes to training deep networks. Through 1) appropriate layer normalization and 2) a new technique of transmitting the combination of previous layers to the next, this work shows that a deep Transformer model can perform better than its Transformer-Big equivalent. It is demonstrated empirically that the deep system (30/25-layer encoder) outperforms the shallow Transformer-Big/Base baseline (6-layer encoder) by 0.4~2.4 BLEU points on the WMT'16 English-German, NIST OpenMT'12 Chinese-English, and WMT'18 Chinese-English tasks. In comparison to the Transformer-Big, the deep model is three times faster to train and 1.6 times smaller.

The study by M.T. Luang et.al.[12] looks at how deep neural language models (NLMs) affect machine translation (MT), emphasizing how well they adapt to lengthy contexts. MT systems usually use NLMs with only one or two hidden layers, and the advantages of adding more layers are still unknown, despite the fact that deep neural networks have demonstrated success in voice and vision. This study shows that NLMs with three or four layers perform better than shallower models in terms of translation quality and confusion. To effectively train deep NLMs that jointly condition on source and target contexts, a variety of approaches are employed. The deep models outperform shallow NLMs by an average of 0.5 TER / 0.5 BLEU points when reranking n-best listings of a robust web-forum baseline. Additionally, a comparable gain of 1.0 TER / 0.5 BLEU points is noted when modified for an SMS-chat domain.

The study by Escribe, M. et.al.[13] assesses neural machine translation (NMT) tools' performance, especially when translating technical content. The publication of L'Apprentissage Profond, which was initially written in English (Deep Learning) and completely machine-translated into French before being expertly post-edited, marked a key milestone in this discipline. A framework for classifying errors was created in order to evaluate the quality of the translation, and a comparison between the unaltered and post-edited versions was carried out in order to spot reoccurring error patterns. The results show that although there were a few grammatical mistakes, the output's overall linguistic quality was adequate. Specialized language, untranslated passages, and stylistic changes were the most frequent mistakes. Some words were of publishable quality and stayed the same in the final edition, even though most of the machine-generated material needed several modifications per segment.

J. Zhou et.al.[14] reported a novel neural machine translation (NMT) approach that addresses the performance gap between single NMT models and traditional machine translation (MT) systems, as well as the constraints of shallow models. Based on deep Long Short-Term Memory (LSTM) networks, a novel kind of linear connection known as "fast-forward connections" is presented, coupled with an interleaved bi-directional architecture for stacking LSTM layers. Fast-forward connections are essential for gradient propagation and allow for the creation of a 16-layer deep network. A single attention-based model outperforms its shallow version by 6.2 BLEU points, achieving a BLEU score of 37.7 on the WMT'14 English-to-French test. This is the first time that a single NMT model has outperformed the best conventional model by 0.7 BLEU points and reached state-of-the-art performance. A BLEU score of 36.3 is achieved by the model even in the absence of an attention mechanism. The best-reported score on this assignment is 40.4, which is attained with extra strategies including managing unknown words and model ensembling. The more difficult WMT'14 English-to-German test is used to further validate the efficacy of the suggested models.

In the stydy by Niehues, J.[15], a paradigm for evaluating neural machine translation (NMT) systems' ability to continually learn new phrases is given. Even though machine translation has significantly improved as a result of recent developments in deep learning, NMT frequently finds it difficult to continuously adapt to new contexts. Although bilingual dictionaries are a promising tool for incorporating new information, their efficient use necessitates that the system be able to accurately model the morphology of both the source and destination languages and perform one-shot learning. The study combines different word representations with one-shot learning techniques and shows that overcoming these obstacles is crucial to making the most of multilingual dictionaries. This improves the capacity to translate new, uncommon words and phrases from 30% to 70%, with over 90% of the time the proper lemma is created.

G. Sahu et.al.[16] proposes an approach to produce more efficient augmented data for text classification by utilizing large language models' (LLMs') ability to follow instructions and carry out few-shot classifications. A common solution to problems resulting from a lack of training data is data augmentation, with more recent methods using LLMs like GPT-3 to produce more examples. The two main components of the suggested PromptMix approach are (1) creating difficult text augmentations close to class boundaries, which improves model learning but increases the possibility of false positives, and (2) relabeling the augmented data with an LLM classifier that relies on prompting to increase label accuracy. This method's performance is assessed on four text classification datasets - Banking77, TREC6, Subjectivity (SUBJ), and Twitter Complaints - in both 2-shot and zero-shot scenarios. Based on experimental results, it is easier to transfer knowledge from large models like GPT-3.5-turbo to smaller, more effective classifiers like DistilBERTbase and BERTbase when borderline examples are generated and relabeled.[17]

In order to improve sequence learning, a Attention Transformer model for translating similar languages by combining Recurrent Attention with Transformer architectures have been reported.[18] Recurrent Attention enhances order preservation during decoding, while Transformers allow for effective parallel processing. The model is well-suited for translating closely related languages because it successfully strikes a balance between the two benefits. At WMT-20, the method is assessed on translation tasks between Hindi and Marathi and Marathi and Hindi.

In order to improve translation, F. Meng et.al.[19] presented Interactive Attention, a novel attention mechanism for Neural Machine Translation (NMT) that allows the decoder and the source representation to interact both in reading and writing. Interactive Attention improves translation accuracy by tracking interaction history, in contrast to conventional attention, which passively reads a fixed source representation. Significant performance improvements are observed in experiments on the NIST Chinese-English task, surpassing the coverage and baseline models of attention-based NMT. On several test sets, the suggested method outperforms the phrase-based Moses system by 3.94 BLEU points and the open-source Groundhog system by 4.22 BLEU points.[20]

By explicitly modeling the relationship between past and future attention levels, the attention mechanism in Neural Machine Translation (NMT) can be improved.[21] In contrast to standard attention, the suggested method tracks attention history by using a recurrent network for every input word, capturing characteristics such as relative distortion and fertility. This parameterized attention

model enhances translation quality, as shown by experimental results.

The study by T. Brants et.al.[22] investigates the advantages of statistical language modeling on a large scale for machine translation. To create language models with up to 300 billion n-grams, a distributed infrastructure is presented that can train on up to 2 trillion tokens. Fast, single-pass decoding with smoothed probabilities is made possible by the system. Furthermore, a brand-new smoothing technique called Stupid Backoff is put forth, which provides an effective substitute for Kneser-Ney Smoothing while preserving a level of quality that is comparable as training data grows.

A framework for quickly adapting machine translation (MT) systems to produce language varieties that deviate from the standard target language without the need for parallel source-variety data is presented.[23] Numerous such variants, such as sociolects and regional dialects, are low-resource and frequently disregarded in NLP solutions. The suggested method also makes it possible to adapt to typologically related languages with limited resources. Significant gains over competitive baselines are demonstrated by experiments that adapt English-Russian MT to Ukrainian and Belarusian, English-Norwegian Bokmål to Nynorsk, and English-Arabic to four Arabic dialects.

The Y. Moslem et.al.[24] investigated how real-time adaptive machine translation (MT) can be improved by incorporating in-context learning into large-scale language models (LLMs). Real-time adaptation is still difficult, even though domain adaptation in MT has improved. LLMs can mimic domain and style features without fine-tuning by being given translation pairs at inference time. Numerous experiments show that, especially for high-resource languages, few-shot in-context learning can outperform robust encoder-decoder MT models. Furthermore, combining MT with fuzzy matches enhances translation quality even more, particularly for languages with limited resources. Five language pairs - English-Arabic, English-Chinese, English-French, English-Kinyarwanda, and English-Spanish—are used to assess the method. The study by L. Wang et.al.[25] assesses the use of large language models (LLMs) for document-level machine translation (MT) discourse modeling. The study looks at three main areas: (1) how context-aware prompts affect discourse phenomena and translation quality; (2) how ChatGPT compares to commercial MT systems and sophisticated document-level MT models.

PROPOSED METHODOLOGY

The proposed Machine Translation model is illustrated in Figure 2.

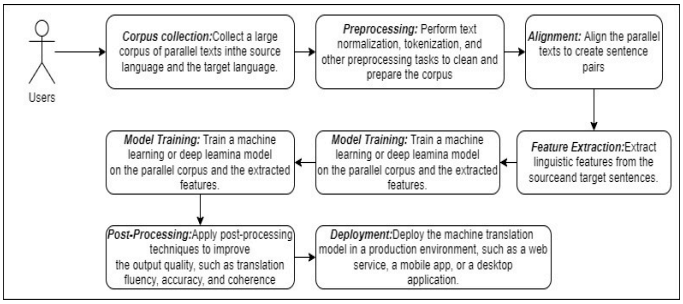


Figure 2. Steps of proposed method

A. Dataset

For this project, we have manually created a large parallel corpus that consists of 10105 sentence pairs in both the source language English and the target Bhojpuri language. The Bho- jpuri language's uncommon words, idioms, and expressions are given particular consideration in this dataset. We have created variants of pre-existing sentence pairs in order to perform data augmentation. This includes changing sentence structures, adding synonyms, or paraphrasing. By adding more data, the model becomes more resilient and able to handle a wider range of linguistic patterns. We have made sure that the dataset accurately represents a range of linguistic phenomena. To prevent biases in the model, this involves varying sentence lengths, grammatical constructions, and linguistic styles. While preparing the dataset we mostly focused on such sentences that contain complex words (like Machine Learning, Artificial Intelligence, and many more) which are not correctly translated by currently available translators like Google Translate.

158	Depending on how the match goes, customers could join and rejoin around key moments	खेल कईसे चलेला एकरा आधार प ग्राहक प्रमुख पल के आसपास शामिल हो सकतारे अवुरी फेर से जुड़ सकतारे
159	Despite feeling tired and overwhelmed, she pushed herself to finish the project on time, staying up late into the night to ensure it was completed to the best of her abilities.	थकान आ अभिभूत महसूस कइला का बावजूद ऊ अपना के समय पर एह परियोजना के पूरा करे खातिर धकेल दिहली, देर रात ले जागल रहली जेहसे कि ई सुनिश्चित हो सके कि ऊ अपना क्षमता के मुताबिक पूरा होखे.
160	Despite her initial reservations, she quickly warmed up to the idea and was soon fully on board.	शुरुआती आरक्षण के बावजूद उ जल्दी से ए विचार से गरम हो गईली अवुरी जल्दिए पूरा तरीका से ए पे आ गईली।
161	Despite his initial reluctance, he eventually came to see the value in taking risks and trying new things.	शुरुआती अनिच्छा के बावजूद अंत में उनुका जोखिम उठावे अवुरी नया चीज़ आजमावे के कीमत देखाई देलस।

Figure 3. Dataset- English to Bhojpuri

B. Preprocessing

The preprocessing stage involves using NLP techniques to clean up the sentences. Removing all the punctuation characters from the text, and then converting the text into lowercase to ensure uniformity. Additionally, removed all the digits(numbers) from the text, replaced one or more whites- pace characters with a single space and removed leading and trailing whitespaces from the text. In simpler terms, preprocessing involves removing unnecessary elements and reducing the complexity of the data while retaining important information.

C. Preparing the data

To know the model when to start and stop, START and END tokens are employed, which prepend and append to the target language. Figure 4 shows the added tokens in the preprocessed data.

```
([['how are you',
   'all is well',
   'long time to see',
   'what is your name',
   'where are you from',
   'nice to meet you'],
  ['<START> का हाल बा <END>',
   '<START> सब बढ़िया बा <END>',
   '<START> बड़ी दिन से भेंट ना भईल ह <END>',
   '<START> तोहार नाव् का ह <END>',
   '<START> तु कहाँ से हव <END>',
   '<START> तोसे मिल कर अच्छा लगल <END>'])
```

Figure 4. Preparation of data by appending tokens

The data comes in the form of a character, which is converted to a numeric representation using TensorFlow's inbuilt tokenizer. In the context of natural language processing tasks, tokenization is a crucial step where words or subwords are converted into numerical representations (sequences of integers) that can be fed into a neural network. This step is essential for preparing the data for training machine learning models, especially neural networks. To restrict the vocabulary, oov(Out-of-Vocabulary) token is used, which replaces any unpopular term with the parameter value and is deemed as not being in the vocabulary.

D. Model Architecture

Using the GRU and Bahadanau attention mechanism, the GRU encoder is in charge of taking input sequences (i.e. English character's sequences), processing them, and creating an output sequence and final hidden state. It processes sequential data during the forward pass, initializing its hidden state to zeros. A GRU is the recurrent neural network layer that is used, and because of its effectiveness in capturing temporal dependencies, it is a good choice for processing sequential data.

By giving the model the ability to concentrate on distinct segments of the input sequence while producing each element of the output sequence, the Bahdanau attention mechanism significantly improves the model's performance. Through the selective attention of relevant sequence segments, this mechanism enhances the model's ability to capture dependencies between various segments of the input and output, thereby improving translation accuracy.

This attention mechanism is built into the decoder so that the model can generate each output element while concentrating on particular segments of the input sequence. Bahdanau attention is used to implement the attention mechanism. By taking into account the context provided by the encoder and focusing on distinct input regions, this approach enables the decoder to make more informed predictions. The model dimensions and parameters are shown in Table 1.

Table 1. Model Dimension

Parameter	Dimension
Batch size	32
Embedding dimension	256
GRU hidden state	1024
Encoder output shape	(32, 18, 1024)
Encoder Hidden state shape	(32, 1024)
Decoder output shape	(32, 2669)
Attention result shape	(32, 1024)
Attention weight shape	(32, 18, 1)

E. Training

The sequence-to-sequence model used in machine translation has a structured training process that effectively maximizes model performance. Every training cycle comprises a forward pass through the decoder and encoder, in which the model is guided during prediction by means of teacher forcing. As the model processes the input sequence, these values are tracked and the loss and accuracy are calculated at each time step. Time steps are used to average the total loss, and batch accuracy is tracked continuously. Gradients are calculated and used by the Adamax optimizer to update the model's parameters, ensuring improvements in performance with each iteration.

The model is trained over a predefined number of epochs, initializing the encoder's hidden state and tracking variables for accuracy and total loss at the beginning of each epoch. As the training loop goes through batches of data, the computed gradients are used to update the model's parameters. To give an overview of the model's development and to provide insights into its progress, logging takes place at regular intervals, usually every 100 batches. The model's performance is summarized at the end of each epoch by computing the average loss and accuracy. The model saves snapshots of its state every two epochs as a checkpoint to protect the training progress. Because of this systematic approach, training is more effective and the model can gradually improve its translation capabilities over several epochs.[26]

RESULTS AND EVALUATION

The above model gets trained for around 30 epochs by maintaining 95% of dataset for training and 5% for testing, using GRU (Gated Recurrent Unit) model. The dataset for this project contains 10105 sentence pairs of English and Bhojpuri sentences and we got an accuracy of 51.25%. Table 2 shows the evaluation report for the last six epochs.

Table 2. Evaluation table when model is trained

Loss	Accuracy
0.0232	0.4624
0.0195	0.4638
0.0162	0.4641
0.0137	0.4641
0.0131	0.4771
0.0129	0.5125

Figure 5 makes it very evident how training epochs affect the model's performance. The model experienced more thorough training as the number of epochs rose, which led to an increase in accuracy. At the same time, there was a drop in the loss, suggesting that the model got better at reducing errors and producing predictions that were more accurate. This trend is illustrated visually in Figure 5, where the plotted data shows how the model's performance changes with each new epoch. The idea that a larger

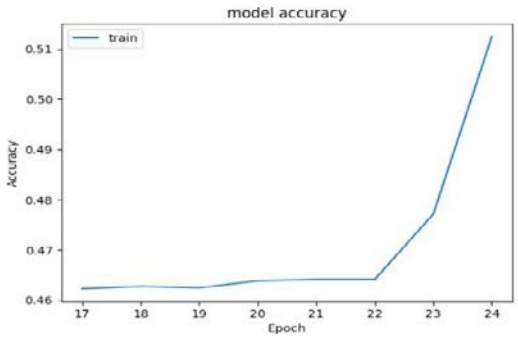


Figure 5. Model accuracy

number of training epochs positively affects the model's learning process and eventually improves its predictive abilities is supported by this empirical data.[27]

The Table 3 presents evaluation measures for English-Bhojpuri translation, like unigram and bigram precision, brevity penalty, and BLEU score. These metrics examine the machine translation system's accuracy, completeness, and overall quality.

Table 3. Evaluation Metrics

Parameters	Value(%)
Unigram Precision	70.00
Bigram Precision	52.38
Brevity Penalty	1.00
Bleu Score	0.62

The translated output of some English sentences is illustrated in Figure 6.

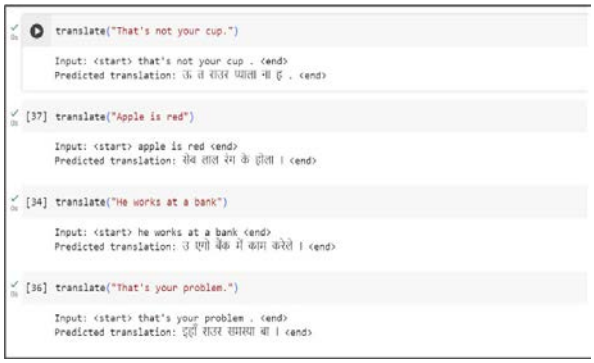


Figure 6. Translation Results

Our dataset currently holds 10105 sentence pairs of English to Bhojpuri translations and we are aiming to improve the size of the dataset further in the near future to about 50000 pairs. We have attempted two models on this dataset till now, LSTM and GRU, we will be aiming to improve the model's accuracy via increased dataset or by using another model such as BERT.

The results of this research show that machine translation from English to Bhojpuri has advanced significantly, especially when it comes to handling uncommon words and idioms. With a final accuracy of 51.25% and a BLEU score of 0.62, the model was trained on a meticulously selected parallel corpus of 10,105 sentence pairs. The evaluation metrics show that the model successfully captures contextual and lexical relationships between the source and target languages, including bigram precision (52.38) and unigram precision (70.00). The model's ability to effectively learn linguistic patterns is further supported by the declining loss values and increasing accuracy over epochs.

Existing machine translation services that translate from English to Bhojpuri include Google Translate and Translation Services USA; however, they frequently have trouble with uncommon words, colloquial expressions, and syntactic accuracy. Large-scale multilingual models are the mainstay of these traditional systems, but they lack adequate Bhojpuri training data, which results in inaccurate transliterations and semantic distortions. Our model addresses the translation of uncommon words and idioms more successfully than these generic systems because it concentrates on Bhojpuri-specific linguistic patterns.

Approaches like transformer-based models and statistical machine translation (SMT) have been investigated in earlier research on machine translation for low-resource languages. Although transformer models such as MarianMT and mBART have proven to be highly accurate in languages with ample resources, they necessitate a substantial amount of pre training data, which is not readily available for Bhojpuri. For Bhojpuri translation with limited data availability, our GRU-based encoder-decoder model with Bahdanau attention offers an alternate method that uses sequence-to-sequence learning to increase contextual accuracy.[28]

Our dataset currently holds 10105 sentence pairs of English to Bhojpuri translations and we are aiming to improve the size of the dataset further in the near future to about 50000 pairs. We have attempted two models on this dataset till now, LSTM and GRU, we will be aiming to improve the model's accuracy via increased dataset or by using another model such as BERT.

DISCUSSION

We will compare our proposed method with similar works from Neural Machine Translation (NMT) field that focus on low-resource language translation within this section. The key features of Table 1 appear in a structured format.

1. Dataset and Language Pair

The field of low-resource NMT research has studied Hindi-English, Hindi-Marathi and Bengali-Hindi language pairs according to previous investigations. The method stands apart because it dedicates attention to the translation of English into Bhojpuri which maintains low visibility in comparison to other language pairs. We built our own spacious parallel corpus which included 10,105 sentence pairs but focused especially on technical language.

2. Data Augmentation and Preprocessing

The previous works used morph-based segmentation along with Byte Pair Encoding (BPE) as their preprocessing techniques. The proposed method builds upon previous methods to incorporate structured data augmentation techniques that use paraphrasing as well as synonym replacement and sentence restructuring. Our preprocessing steps include tokenization as well as case normalization and out-of-vocabulary token mechanisms that strengthen model resilience.

3. Model Architecture

This research implements a GRU-based encoder-decoder system with Bahdanau attention as its core component whereas earlier works relied mainly on transformers alongside recurrent models with attention. GRU remains our selection because it maintains efficient performance in sequence dependency processing even though it consumes fewer computational resources than LSTMs and transformers.

4. Training Strategy

The training process adopts teacher forcing in combination with Adamax optimization to refine translation performance systematically during thirty epochs of training. The model operation creates multiple checkpoint points regularly which

stabilizes the process and allows ongoing improvements across time.

5. Evaluation and Performance

Several researches documented BLEU score elevation between 15 and 20 points through the fusion of multilingual transformers and augmentation approaches. The training data comprised of English-Bhojpuri pairs which enabled our model to reach 51.25% accuracy and 0.62 BLEU score performance levels.

Table 4. Comparative Analysis

Feature	Related Work	Proposed Work
Language Pair	Hindi-English, Hindi-Marathi, Bengali-Hindi	English-Bhojpuri
Dataset Size	Pre-existing corpora	Manually created 10,105 sentence pairs
Data Augmentation	Morph-based segmentation, BPE	Paraphrasing, synonym replacement, sentence restructuring
Preprocessing	Morph-based, BPE segmentation	Tokenization, case normalization, out-of-vocabulary handling
Model Architecture	Transformer, RNN with attention	GRU with Bahdanau attention
Training Method	Parallel corpora, multilingual approach	Teacher forcing, Adamax optimization, checkpointing
Evaluation Metrics	BLEU improvements (15-20 points), Translation Edit Rate (TER)	BLEU Score: 0.62, Accuracy: 51.25%

The innovative method stands out as the analysis shows its ability to tackle obstacles in English-Bhojpuri machine translation effectively. The implementation of strong data augmentation and GRU-based design produces good outcomes yet requires additional data expansion and alternative architectural choices like BERT for further advancement.

AUTHOR CONTRIBUTION

The ideation, planning, and implementation of this study were all done by the authors working together. Each author contributed differently but cooperatively to the study's successful conclusion. In order to create the dataset, Aniket and Anuj oversaw the manual curation of 10,105 parallel sentence pairs in Bhojpuri and English. Idioms, uncommon words, and culturally specific expressions received particular attention. In order to guarantee high-quality input data, they also used data preprocessing techniques like text normalization, tokenization, and cleaning. They used data augmentation techniques to improve model performance, producing lexical and syntactic sentence variations. The model's implementation and architecture were the responsibility of Aniket and Anuj. In order to increase translation accuracy, this involved integrating the attention mechanism, optimizing hyperparameters, and designing and coding the GRU-based encoder-decoder model with Bahdanau attention. Additionally, Pratik helped with model debugging, guaranteeing computational effectiveness and resolving training stability concerns. Performance analysis, evaluation, and model training were managed by Dr. Madhuri. In order to improve convergence, this required fine-tuning the model over 30 epochs, employing teacher forcing, and optimizing the learning process with the Adamax optimizer. BLEU scores, precision metrics, and loss reduction trends were used in her quantitative assessment. Furthermore, Animesh carried out an error analysis, pointing out translation mistakes pertaining to syntax, semantics, and handling uncommon words. They also offered suggestions for improvement. The study was well-structured and fully expressed thanks to the collaborative efforts of all authors during the manuscript's drafting and review stages. In order to place the study within the larger field of low-resource machine translation, each author contributed to the literature review. In order to analyze the findings, make insightful deductions, and polish the concluding discussion section, the writers also had lengthy conversations.

CONCLUSION

Bhojpuri translation is a need as it helps to cater the business, professional and personal needs of people living in Bihar, Jharkhand and Uttar Pradesh. It is spoken by a large volume of the workforce in all India as well. Although some translators exist for this language, the translation available cannot be used extensively as it unable to translate many rare and complex words used in the English language. In this research, we concentrated on creating a machine translation system for translating English into Bhojpuri, a language with low resources in general. The challenge of translating rare words was addressed due to limited training data. Preparing the dataset, preprocessing the data, and using an encoder-decoder architecture with GRU model and Bahadanau attention mechanism were all parts of the methods we suggested. The machine translation model identifies areas for improvement even though it successfully translates many English sentences into accurate Bhojpuri. Because of the dynamic nature of language and its many contextual nuances, it will always be difficult to translate every sentence perfectly. We expect continued improvements in translation accuracy and fluency as machine translation technology develops and is refined, which will promote efficient cross-linguistic communication.

CONFLICT OF INTEREST STATEMENT

The author declared no conflict of interest for the publication of this work.

REFERENCES AND NOTES

1. A. Banerjee, A. Jain, S. Mhaskar, et al. Neural Machine Translation in Low-Resource Setting. In *Proceedings of Machine Translation Summit XVIII: Research Track*; **2021**; pp 36–39.
2. V. Mujadia, D.M. Sharma. NMT-based Similar Language Translation for Hindi – Marathi. In *Proceedings of the Fifth Conference on Machine Translation*; **2020**; pp 414–416.
3. V. Goyal, P. Mishra, D.M. Sharma. Linguistically informed Hindi-english neural machine translation. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*; **2020**; pp 3698–3703.
4. P. Madaan, F. Sadat. Multilingual Neural Machine Translation involving Indian Languages. In *Proceedings of the WILDRE5{--}5th Workshop on Indian Language Data: Resources and Evaluation*; **2020**; pp 29–32.

5. A. Das, P. Yerra, K. Kumar, S. Sarkar. A study of attention-based neural machine translation model on Indian languages. *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*. 2016, pp 163–172.
6. M.T. Luong, H. Pham, C.D. Manning. Effective approaches to attention-based neural machine translation. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*; **2015**; pp 1412–1421.
7. S. Kumar, A. Anastasopoulos, S. Wintner, Y. Tsvetkov. Machine Translation into Low-resource Language Varieties. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*; **2021**; Vol. 2, pp 110–121.
8. R. Sennrich, B. Haddow, A. Birch. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*; **2016**; Vol. 3, pp 1715–1725.
9. M. Garg, S. Kumar, A.K.J. Saudagar. Natural Language Processing and Information Retrieval: Principles and Applications; Oxford University Press, **2023**.
10. P. Zaremoodi, G. Haffari. Neural machine translation for bilingually scarce scenarios: A deep multi-Task learning approach. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. 2018, pp 1356–1365.
11. Q. Wang, B. Li, T. Xiao, et al. Learning deep transformer models for machine translation. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020, pp 1810–1822.
12. M.T. Luong, M. Kayser, C.D. Manning. Deep neural language models for machine translation. *CoNLL 2015 - 19th Conference on Computational Natural Language Learning, Proceedings*. 2015, pp 305–309.
13. M. Escribe. Human Evaluation of Neural Machine Translation: The Case of Deep Learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop*; **2019**; pp 36–46.
14. J. Zhou, Y. Cao, X. Wang, P. Li, W. Xu. Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*. 2016, pp 371–383.
15. J. Niehues. Continuous learning in neural machine translation using bilingual dictionaries. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*; **2021**; pp 830–840.
16. G. Sahu, O. Vechtomova, D. Bahdanau, I.H. Laradji. PromptMix: A Class Boundary Augmentation Method for Large Language Model Distillation. In *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*; **2023**; pp 5316–5327.
17. G. Sahu, P. Rodriguez, I.H. Laradji, et al. Data Augmentation for Intent Classification with Off-the-shelf Large Language Models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2022, pp 47–57.
18. Farhan, M. Rafi. Attention Transformer Model for Translation of Similar Languages. *5th Conference on Machine Translation, WMT 2020 - Proceedings*. 2020, pp 387–392.
19. F. Meng, Z. Lu, H. Li, Q. Liu. Interactive Attention for Neural Machine Translation. *arXiv preprint*. 2016, pp 2175–2181.
20. A. Hosseini, S. Reddy, D. Bahdanau, et al. Understanding by Understanding Not: Modeling Negation in Language Models. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. 2021, pp 1301–1312.
21. Y. Zichao, Z. Hu, D. Yuntian, C. Dyer, A. Smola. Neural machine translation with recurrent attention modeling. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*; **2017**; Vol. 2, pp 383–387.
22. T. Brants, A.C. Popat, P. Xu, F.J. Och, J. Dean. Large language models in machine translation. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007, pp 858–867.
23. S. Kumar, A. Anastasopoulos, S. Wintner, Y. Tsvetkov. Machine Translation into Low-resource Language Varieties. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*; **2021**; Vol. 2, pp 110–121.
24. Y. Moslem, R. Haque, J.D. Kelleher, A. Way. Adaptive Machine Translation with Large Language Models. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023*. 2023, pp 227–237.
25. L. Wang, C. Lyu, T. Ji, et al. Document-Level Machine Translation with Large Language Models. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2023, pp 16646–16661.
26. P.R. Pardhi, J.K. Rout, P.K. Agrawal, N.K. Ray. Blockchain for decentralized malware detection on android devices. *J. Integr. Sci. Technol.* **2025**, 13 (4), 1084.
27. P.K. Agrawal. A Novel Mapper Machine Learning Algorithm for Semantic Domain Mapping for Domain Database Updation. *SN Comput. Sci.* **2023**, 4 (5), 536.
28. S. Kadu, B. Joshi, P. Agrawal. Optimization for Image Sentiment Analysis Using Novel Dual Moth Flame Algorithm. *SN Comput. Sci.* **2025**, 6 (2), 134.