IACS-LRILT: Machine Translation for Low-Resource Indic Languages

Dhairya Suman[1⋆] and **Atanu Mandal**[2†] and **Santanu Pal**[3‡] and **Sudip Kumar Naskar**[4†]
⋆Indian Association for the Cultivation of Science, Kolkata, India
†Jadavpur University, Kolkata, India
‡Wipro AI Lab, London, UK
{[1]dhairyasuman, [2]atanumandal0491, [3]santanu.pal.ju, [4]sudip.naskar}@gmail.com

Abstract

Even though, machine translation has seen huge improvements in the the last decade, translation quality for Indic languages is still underwhelming, which is attributed to the small amount of parallel data available. In this paper, we present our approach to mitigate the issue of the low amount of parallel training data availability for Indic languages, especially for the language pair English-Manipuri and Assamese-English. Our primary submission for the Manipuri-to-English translation task provided the best scoring system for this language direction. We describe about the systems we built in detail and our findings in the process.

1 Introduction

The ability to overcome linguistic barriers has emerged as the most critical issue in a society that is becoming increasingly interconnected. These linguistic barriers can be eliminated enabling effective communication among various linguistic communities, machine translation (MT) systems are not only capable of translating common languages but also less widely spoken or even endangered languages, ensuring that even marginalized communities can participate in the global conversation. The use of machine translation for regional Indian languages is both an intriguing and challenging application. India has an intricate mix of languages and dialects spoken all over its broad territory, making it a linguistically diverse nation (Mandal et al., 2021). Despite being culturally stimulating, this diversity poses substantial obstacles to effective communication. By automating the translation process and opening up content to speakers of different regional Indian languages, machine translation presents a viable remedy.

Due to deep learning, neural networks, and natural language processing developments, machine translation technology has made significant strides in recent years (Slocum, 1985). However, there are particular difficulties that must be overcome in order to adapt these technologies to the intricate linguistic features of Indian languages (Pal et al., 2013a). These difficulties include, among other things e.g., multi-word expressions (Pal et al., 2013b), the complexity of morphology, syntactic changes, and the scarcity of parallel training data (Pal, 2018). The challenge of producing accurate and relevant translations is further complicated by the requirement to preserve cultural nuances and context-specific meanings (Appicharla et al., 2023).

However, the translation problem for Indian regional languages is compounded by:

- **Morphological complexity:** Indian languages often exhibit rich morphology, leading to variations in word forms and sentence structures.
- **Low-resource languages:** Limited parallel training data is available for many Indian language pairs, leading to challenges in training accurate translation models.
- **Cultural and context preservation:** Accurate translation must account for context-specific meanings, idiomatic expressions, and cultural nuances.

So working on Indic languages has the challenge of designing translation models and techniques that address these complexities and constraints while achieving high-quality translations between Indian regional languages, contributing to effective cross-lingual communication and content accessibility in India's diverse linguistic landscape.

2 Related Work

Parul and Garg (2022) provides a survey of different approaches to Machine Translation (MT) for Indian languages, including Rule-based Machine Translation (RBMT), Corpus-based Machine

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 972–977
December 6–7, 2023.

Translation (CBMT), and Neural Machine Translation (NMT). Researcher (Parul and Garg, 2022) highlights the initial slow progress in MT research and the subsequent popularity of NMT. The paper emphasizes that while there has been significant research on MT for top-level languages, there is a scarcity of research for low-level languages spoken by fewer people. It discusses the use of different MT models, such as Anusaarka for direct MT, AnglaHindi for Interlingual translation, and CBMT for translation using stored data corpus.

Jha et al. (2023) presents the development and evaluation of a multilingual neural machine translation system for Indian languages using the mT5 transformer. The system was trained on the modified Asian Language Treebank multilingual dataset to translate text between English, Hindi, and Bengali. The system achieved acceptable Bilingual Evaluation Understudy (BLEU) scores, with the English-to-Bengali system achieving a maximum BLEU score of 49.87 and the Bengali-to-English system achieving an average BLEU score of 42.43. Jha et al. (2023) claims that the field of Natural Language Processing (NLP) research in low-resource languages has been expanding rapidly, with transformers being the latest state-of-the-art systems.

Jayanthi et al. (2020) states that India is a multicultural and multilingual country, with a large number of regional languages. English is provided as the second extra official language in India, but its usage is limited, leading to a communication gap. Machine translation can help minimize this gap by translating languages. Jayanthi et al. (2020) focuses on translating Indic languages, specifically Telugu, using a sequence-to-sequence framework with an encoder-decoder attention mechanism of neural machine translation. The proposed framework aims to convert the Telugu language into English and vice versa. Their approach framework was trained using a Telugu parallel corpus and achieved good accuracy. It overcomes the limitation of reduced accuracy when faced with unknown words by using an attention mechanism. As per the author, the sequence-to-sequence model used in this paper allows for the conversion of the native language into the desired language, and the attention mechanism helps handle rare words.

S. and Bhattacharyya (2020) claims the use of Indowordnet helped handle ambiguity during translation and improved the performance of the machine translation systems. The author presents a comparative study of 440 phrase-based statistically trained models for 110 language pairs across 11 Indian languages and also discusses the principles followed in constructing the synsets, such as the minimality principle, coverage principle, and replaceability principle.

Research involving Indian languages is not very common due to the scarcity of parallel corpora. Baruah et al. (2014) using Statistical Machine Translation (SMT) with a small corpus (2,500 sentences), the Assamese-English bidirectional MT system for Assamese to English and English to Assamese obtained BLEU scores of 9.72 and 5.02, respectively. Das and Baruah (2014) investigated and reported a BLEU score of 11.32 for Assamese to English using SMT using 8,000 Tourism domain parallel sentences.

3 Method

3.1 Problem Definition

Given a source sentence in an Indian regional language, represented as $S = \{s_1, s_2, \ldots, s_n\}$, and a target sentence in a different Indian regional language or English, represented as $T = \{t_1, t_2, \ldots, t_m\}$, the objective of machine translation for Indian regional languages is to find the optimal translation function f that maximizes the translation quality while considering linguistic nuances, morphological complexities, and contextual information:

$$f^\star = \text{argmax}\, f(P(T \mid S)) \tag{1}$$

In equation 1, $f^\star$ represents the optimal translation function that produces the highest probability of the target sentence given the source sentence. $P(T \mid S)$ is the conditional probability of the target sentence T given the source sentence S, which is modelled using statistical or neural machine translation approaches. $S = \{s_1, s_2, \ldots, s_n\}$ denotes the sequence of words in the source sentence. $T = \{t_1, t_2, \ldots, t_m\}$ denotes the sequence of words in the target sentence. n is the length of the source sentence, and m is the length of the target sentence.

3.2 Dataset Description

Table 1 represents the Datasets for the language pair of Assamese-English and Manipur-English language pair in the WMT 2023 IndicMT[1] shared

[1] http://www2.statmt.org/wmt23/indic-mt-task.html

Language Pair	Train	Validation	Test
Assamese-English	50,000	2,000	2,000
English-Manipuri	21,686	1,000	1,000

Table 1: Dataset statistics for Workshop on Machine Translation (WMT) 23

task. As per the organizers' guidelines, no additional parallel data was allowed for training with only constrained submissions.

3.3 Experimental Setup

IndicBART (Dabre et al., 2022) and mbart-large-50 (Tang et al., 2020) have been adjusted for the bidirectional Assamese-English and English-Manipuri language pairs in our suggested study. We fixed the source and target lengths in both scenarios to "128". With batch sizes of "16" and "8", respectively, and learning rates of "2×10^{-5}" for both scenarios, we improved our suggested IndicBART and mbart-large-50 models, We applied weight decay of "0.01" for both scenarios.

3.4 Corpus Pre-processing

We used IndicBART (Dabre et al., 2022) developed by AI4Bharat[2] for some of the models. Using IndicBART for Indic languages other than Hindi or Marathi requires the language to be transliterated into the Devanagari script. Hence, we had to transliterate the data given into the Devanagari script to use those models.

3.5 Experiments

3.5.1 Bidirectional Assamese-English Language Pair

We first experimented by using IndicTrans (Ramesh et al., 2022) from AI4Bharat to get the responses on the Validation Set provided, but the BLEU scores on the same were unsatisfactory. We experimented by finetuning IndicBART from AI4Bharat on the Training Set and evaluating the responses on the given Validation Set. This gave us better results so we decided that these responses would be our Primary Submissions. IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on Indic Languages and English. Currently, it supports 11 Indian languages, Assamese, Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Kannada, Malayalam, Tamil, and Telugu based on mBART (Liu et al., 2020) architecture.

[2] https://ai4bharat.iitm.ac.in/

We used the transliteration module from the IndicNLP library (Kunchukuttan, 2020) for transliterations from Assamese to Devanagari, an example is shown in Figure 1.

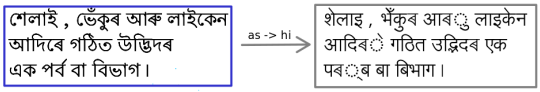

Figure 1: Transliteration from Assamese to Hindi

These experiments are discussed below:

- **Primary Submission**
  We took the training data and fine-tuned it on IndicBART for the translation settings from Assamese to English. This model gave good BLEU scores on the Validation set hence, this model was selected as the Primary System.

- **Contrastive - 1**
  Here, it was considered that since Assamese and Bengali share linguistic similarities, it may be that IndicBART fine-tuned on the training dats but this time for translations from English-Bengali, did give results, surprisingly similar to the Primary System

- **Contrastive - 2**
  Here we used IndicTrans from AI4Bharat, the translator was built and the responses on the Test Set were calculated. Note, that for this system no Transliteration was required.

For the models that used IndicBART, we had to transliterate the data from Assamese to Hindi using the IndicNLP transliterator. Moreover, the responses generated by these models, when the target language was Assamese also had to be back-transliterated from Hindi to Assamese for the evaluation of the Validation Set.

3.5.2 Bidirectional English-Manipuri Language Pair

Since resources available for the Manipuri language are very scarce, we decided to use existing models available for Bengali and Assamese. This was because Manipuri shares its script with Assamese and Bengali, so even with morphological differences the models gave good scores for Manipuri. We used mbart-large-50 (Tang et al., 2020) from Facebook and IndicBART by AI4Bharat.

For the language pair English-Manipuri there were no existing transliteration tools that we found,

Framework	BLEU	ChrF	RIBES	TER	COMET
English-to-Assamese					
Primary	34.82	56.58	0.87	55.10	0.77
Contrastive-1	34.71	56.59	0.87	54.75	0.78
Benchmark	8.57	25.24	0.44	86.14	0.59
Contrastive-2	6.57	39.71	0.45	86.26	0.79
Assamese-to-English					
Primary	66.36	75.88	0.93	37.44	0.84
Contrastive-1	66.33	75.88	0.93	37.38	0.84
Contrastive-2	23.19	48.42	0.61	71.79	0.75
Benchmark	11.28	28.70	0.53	83.10	0.56
English-to-Manipuri					
Primary	25.78	49.94	0.84	60.43	0.71
Contrastive-1	25.82	49.93	0.84	60.57	0.71
Benchmark	21.58	45.97	0.61	69.76	0.69
Contrastive-2	9.69	40.45	0.54	81.18	0.67
Manipuri-to-English					
Primary	69.75	78.16	0.94	32.08	0.84
Contrastive-1	69.75	78.16	0.94	32.10	0.84
Benchmark	24.86	46.37	0.64	70.26	0.63
Contrastive-2	22.10	48.03	0.63	72.19	0.70

Table 2: Results of Primary, Contrastive-1, and Contrastive-2 submissions evaluated on Benchmark results for the language pair Assamese-English and English-Manipuri.

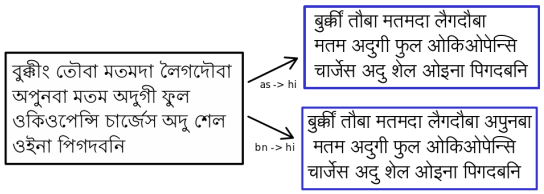

Figure 2: Transliteration from Manipuri to Hindi

but it was thought that, since Manipuri has script similarities with Bengali and Assamese we can experiment with transliteration tools from Bengali and Assamese to Hindi with the expectation for good results and it turns out it does give good results. For this task too we used the transliteration tools from the IndicNLP library, an example is shown in Figure 2.

We discussed in detail about these experiments as follows:

- **Primary Submission**
  The data was first transliterated into Hindi using the transliteration from Bengali to Hindi, then we finetuned IndicBART on the Training Data and evaluated the responses given for the Validation Set. This model gave the highest score on the Validation Set and hence, was picked as the Primary model.

- **Contrastive - 1 Submission**
  This model was similar to the Primary model, but instead of the transliteration and Translation settings, Bengali was the Indic language instead of Assamese.

- **Contrastive - 2 Submission**
  For this model, we fine-tuned mbart-large-50 with the Bengali-English configuration. This model gave a lesser score on the validation set than the models discussed before, even though this was a larger model.

Similar, to the Primary and Contrastive - 1 system for Task 1, responses from the models that used IndicBART had to be back-transliterated from Hindi to the Indic language, when the Indic language was the target language.

3.6 Post-processing

Along with the back-transliteration that was required for the models using IndicBART when the target language was the Indic language. We also

had to do some post-processing of the responses received, we saw that often the responses had random Chinese characters and emoticons in the responses. The emoticons were chalked up to encoding errors while saving the responses to a text file, on the other hand, the Chinese characters were something that we think were errors because of the model itself. These noisy characters were manually removed to ensure that they don't affect the accuracy.

4 Results and Analysis

Table 2 lists the findings of our experiments. We list our observations here:

- As we discussed in section 3.6 we believe that there might be noise in the responses saved that we missed or couldn't manually find, which can contribute to a lesser score even though the translations are accurate.
- We also believe that there might be some issues in translation because of transliteration problems while back-transliterating we often came across responses that still had some words in Hindi. Due to this we also believe that there might have been errors in transliteration from the Assamese/Manipuri to Hindi.
- For task 4, we also consider that the transliteration and translation models used were configured to Assamese and Bengali, so even though the models were fine-tuned on the data but still we assume that because of the morphological differences, there might be gaps in the understanding and generating of language by the model.
- An interesting observation that can be made is that there exists a large gap in the scores for when English is the target language and when the target language is the Indic language. This error can be attributed to the model understanding the target languages morphologically well, but not being able to generate the language that well.

5 Conclusion and Future Work

In this paper, We discussed the models and procedures our team used for the language pairs Assamese-English and English-Manipuri. According to our experiments, we claimed that Using language models like IndicBART and mbart-large-50 results in improvement for the low-resourced individual languages results. We hope that this will enable us to develop more precise and superior translation models for languages and domains with limited resources specially for Indian Languages where there is a presence of large language diversity. We also believe that, as seen with Manipuri, a language with very few resources for processing we can use languages close and similar to it to aid in its processing and create a better way of processing those low-resource languages. In future, we will include our models in online post-editing platforms (Pal et al., 2016; Nayak et al., 2015; Vela et al., 2019).

Acknowledgements

This research was supported by the TPU Research Cloud (TRC) program, a Google Research initiative and funded by the 'VIDYAAPATI: Bidirectional Machine Translation Involving Bengali, Konkani, Maithili, Marathi, and Hindi' under the Project titled 'National Language Translation Mission (NLTM): BHASHINI'.

References

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2023. A case study on context encoding in multi-encoder based document-level neural machine translation. *arXiv preprint arXiv:2308.06063*.

Kalyanee Baruah, Pranjal Das, Abdul Hannan, and Shikhar Sarma. 2014. Assamese-english bilingual machine translation. *International Journal on Natural Language Computing*, 3.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for natural language generation of indic languages. In *Findings of the Association for Computational Linguistics*.

Pranjal Das and Kalyanee Kanchan Baruah. 2014. Assamese to english statistical machine translation integrated with a transliteration module. *International Journal of Computer Applications*, 100:20–24.

N Jayanthi, Aluri Lakshmi, Ch Suresh Kumar Raju, and B Swathi. 2020. Dual translation of international and indian regional language using recent machine translation. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 682–686.

Abhinav Jha, Hemprasad Yashwant Patil, Sumit Kumar Jindal, and Sardar M N Islam. 2023. Multilingual indian language neural machine translation system

using mt5 transformer. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Atanu Mandal, Santanu Pal, Indranil Dutta, Mahidas Bhattacharya, and Sudip Kumar Naskar. 2021. Is attention always needed? a case study on language identification from speech. *ArXiv*, abs/2110.03427.

Tapas Nayak, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 36–42.

Santanu Pal. 2018. A hybrid machine translation framework for an improved translation workflow.

Santanu Pal, Mahammed Hasanuzzaman, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013a. Impact of linguistically motivated shallow phrases in pb-smt. In *ICON 2013*. https://www.researchgate.net/publication ....

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013b. Mwe alignment in phrase based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, pages 61–68.

Santanu Pal, Sudip Kumar Naskar, Marcos Zampieri, Tapas Nayak, and Josef van Genabith. 2016. CATaLog online: A web-based CAT tool for distributed translation with data capture for APE and translation process research. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 98–102, Osaka, Japan. The COLING 2016 Organizing Committee.

Parul and Kamal Deep Garg. 2022. Machine translation system for indian language: Survey. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 468–473.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Sreelekha S. and Pushpak Bhattacharyya. 2020. Indowordnet's help in indian language machine translation. *AI Soc.*, 35(3):689–698.

Jonathan Slocum. 1985. A survey of machine translation: Its history, current status and future prospects. *Computational Linguistics*, 11(1):1–17.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Mihaela Vela, Santanu Pal, Marcos Zampieri, Sudip Naskar, and Josef van Genabith. 2019. Improving CAT tools in the translation workflow: New approaches and evaluation. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 8–15, Dublin, Ireland. European Association for Machine Translation.