Neural Machine Translation for English - Manipuri and English - Assamese

Goutam Agrawal and **Rituraj Das** and **Anupam Biswas** and **Dalton Meiti Thounaojam**
National Institute of Technology, Silchar
{goutam_pg_22, rituraj_pg_22, anupam, dalton}@cse.nits.ac.in

Abstract

The internet is a vast repository of valuable information available in English, but for many people who are more comfortable with their regional languages, accessing this knowledge can be a challenge. Manually translating this kind of text, is a laborious, expensive, and time-consuming operation. This makes machine translation an effective method for translating texts without the need for human intervention. One of the newest and most efficient translation methods among the current machine translation systems is neural machine translation (NMT). In this WMT23 shared task: low resource indic language translation challenge, our team named ATULYA-NITS used the NMT transformer model for the English to/from Assamese and English to/from Manipuri language translation. Our systems achieved the BLEU score of 15.02 for English to Manipuri, 18.7 for Manipuri to English, 5.47 for English to Assamese, and 8.5 for Assamese to English.

1 Introduction

In countries like India, linguistic diversity is a significant aspect, with a multitude of languages varying across different regions. India officially recognizes 23 languages (Das et al., 2020) (e.g., Hindi, Sanskrit, Assamese, Odia, etc.), and alongside these, there are several hundred unofficial local languages spoken by communities. Despite India's vast population of approximately 1.4 billion, only about 11% of the population is proficient in English (Azam et al., 2013).

This language barrier becomes crucial when considering the abundance of valuable resources available on the internet, mostly in English, as a significant proportion of people in India cannot fully comprehend this content. Consequently, there arises a pressing need to translate such valuable information into local languages to facilitate knowledge sharing among the population. Such knowledge dissemination is crucial not just for business purposes but also for enabling the exchange of feelings, opinions, and actions, thereby fostering better communication and understanding among people from diverse linguistic backgrounds.

Manual translation of such copious amounts of content would be extremely laborious and time-consuming, making automatic machine translation an indispensable solution. However, machine translation for Indian languages presents its own set of challenges (Singh et al., 2021). One key challenge is the scarcity of parallel corpora, as there are fewer resources available for Indian languages compared to more widely spoken foreign languages. Moreover, the structural differences between Indian languages and English, particularly in terms of morphological richness and word order, pose significant obstacles to accurate translation. For instance, English follows a Subject-Verb-Object (SVO) word order, whereas Indian languages like Assamese and Manipuri, follow a Subject-Object-Verb (SOV) word order (Bora, 2015). Furthermore, English is a fusional language, while Assamese and Manipuri are agglutinative languages (Singh and Singh, 2022; , leading to distinct syntactic and morphological complexities that further complicate the translation process.

We participated in the Low-Resource Indic Language Translation task on translating two language pairs i.e. English to/from Assamese, and English to/from Mizo. We did the preprocessing of the given dataset and applied a neural machine translation technique i.e. transformer model. The performance was evaluated using the widely used evaluation metric BLEU. The rest of the paper is organized as follows: Section 2 discusses the existing machine translation systems and techniques tailored to Indian languages. In Section 3, we present details about the dataset, preprocessing of the dataset, and transformer model. In section 4, we discussed about the result. Finally, in Section 5, we

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 931–934

conclude with a discussion of the future prospects.

2 Literature Survey

Over the past few decades, machine translation (MT) has been the subject of extensive research. Researchers have explored various approaches in this field, including rule-based MT (Das and Baruah, 2014; Forcada et al., 2011), corpus-based MT, also known as data-driven MT (Laskar et al., 2022; Laitonjam and Singh, 2021; Singh and Bandyopadhyay, 2010), and hybrid-based MT (Laitonjam and Singh, 2022). Each of these approaches has its own advantages and disadvantages.

In rule-based MT, systems analyze the source text to create an intermediate representation, and depending on this representation, it can be further categorized into transfer-based (TBA) and interlingua-based (IBA) approaches. The corpus-based approach, on the other hand, relies on large parallel corpora consisting of text and their translations to acquire translation knowledge and is sub-divided into two sub-types, i.e. statistical machine translation (SMT) and example-based machine translation (EBMT). SMT generates translations using statistical models that combine language models and translation models with decoding algorithms. In contrast, EBMT uses existing translation examples to generate new translations. Hybrid-based machine translation combines aspects of both rule-based and corpus-based approaches to address their respective limitations.

The machine translation performance for Indian language pairs (e.g., Hindi, Bengali, Tamil, Punjabi, Gujarati, and Urdu) into English achieves an average accuracy of only 10%, (Khan et al., 2017) highlighting the need for improved machine translation systems for these languages. Neural Machine Translation (NMT) has emerged as a novel and promising technique for various languages, exhibiting remarkable results (Devi and Purkayastha, 2023; Laskar et al., 2022, 2021). In this paper, we have applied the transformer model to the English-Assamese and English-Manipuri language pair (Laskar et al., 2021; Singh and Singh, 2022)

3 Methodology and Evaluation

3.1 Dataset Details

The English-Assamese parallel corpus (Pal et al., 2023) comprised a grand total of 53,000 sentence pairs, while the Assamese monolingual corpus contained nearly 2.6 million sentences. Moving over to the English–Manipuri parallel corpus (Pal et al., 2023), it included a substantial 24,300 aligned sentence pairs. As for the Manipuri monolingual dataset, it contained roughly 2.1 million sentences.

3.2 Data Preprocessing

The dataset may contain repetition of sentences with the same source and the same target translation, sentences with the same source but different translations, sentences with different source text but the same translation. To address these issues, a solution was implemented by selecting unique sentence pairs from all available sentences and removing the duplicates. Sentences repeated more than once were completely removed to avoid ambiguity in determining the correct translation for a given source and vice versa. This preprocessing step aimed to ensure that the training and test sets did not contain the same sentences, which could result in better predictions for the test set but incorrect predictions for new sentences. Some additional preprocessing steps were carried out, including removing sentences with a length greater than 50, removing noisy translations and unwanted punctuations, filtering out sentences in other languages by applying language identification, and filtering out sentences containing HTML tags, illegal characters, and invisible characters. Finally, the dataset was split into training, testing, and validation sets, following shuffling. The English-Assamese parallel corpus was segregated into 49,500 for training, 2,000 for validation, and 1,000 for testing. Similarly, the English-Manipuri parallel corpus was divided into 21,000 for training, 2,000 for validation, and 1000 for testing.

3.3 Transformer Model

The Transformer model(Vaswani et al., 2017) is a powerful architecture used in tasks like machine translation. It excels in natural language processing, employing a technique called "self-attention" to process sequential data effectively. Unlike traditional models, it considers the context of the entire sequence, using multiple self-attention mechanisms known as "attention heads" to capture different relationships between words. Positional encoding is added to understand the word order. In machine translation, it consists of an encoder and a decoder communicating through attention mechanisms.

For the task of the English-Assamese language pair, along with the provided parallel corpus (Pal et al., 2023), we also used the monolingual corpus to create the vocabulary for the English and Assamese languages. The vocabulary extracted from the monolingual corpus generated a total of 107483 unique tokens in the Assamese language. The vocabulary size of the English language was 35487.

We used the transformer model to train the data. For the whole process, we used Google Colab and trained the model using a T4 GPU provided by Colab. We trained the model for 2000 training steps and 250 validation steps. We set the word vector size to 512 and used 6 layers of 512 hidden nodes. We set the transformer feed-forward size to 2048 and used 8 attention heads. We set the learning rate to 1 while using Adam optimization(Kingma and Ba, 2014). We used a batch size of 2048 with a dropout probability of 0.1 and used a label smoothing regularization technique to prevent overconfidence. The whole training process took around 4 hours when we used the batch size of 2048.

The vocabulary extracted from the monolingual corpus generated a total of 84072 unique tokens in the Manipuri language. For the English-Manipuri language pair, we trained the model for 1500 training steps and 150 validation steps, and all the remaining were similar to English-Assamese. The whole training process took around 3 hours.

4 Evaluation

4.1 Evaluation Metric

The Bilingual Evaluation Understudy (BLEU) score is a useful tool for determining the differences between translations produced by machines and those created by human translators (Papineni et al., 2002). This assessment method compares and aligns the number of n-grams in the translated output with the number of n-grams in the source text. In this context, a bigram comparison entails analyzing every word pair, while a unigram comparison relates to each individual token. It's significant to notice that this evaluation ignores the comparison's precise wording. This methodology is an improved version of a simple precision-based evaluation strategy.

4.2 Result

BLEU, chrf2, RIBES, and TER evaluation metrics on both language pairs are shown in Table 1.

Language Pair	BLEU	Chrf2	RIBES	TER
English-Assamese	5.47	21.66	0.21	0.5
Assamese-English	8.5	24.26	0.25	0.47
English-Manipuri	15.02	35.96	0.28	0.43
Manipuri-English	18.7	38.49	0.32	0.41

Table 1: The experimental result of language pairs on different evaluation metrics

5 Conclusion

In this paper, we applied NMT to the two most difficult language pairs (English-Assamese and English-Manipuri). We showed that the transformer model performs better for Indian languages. We achieved a fairly good BLEU score for the English-Manipuri language pair. So, this model can be used for domains such as tourism and education. Moreover, this transformer model is useful for various English-Indian language pair translations.

References

Mehtabul Azam, Aimee Chin, and Nishith Prakash. 2013. The returns to english-language skills in india. *Economic Development and Cultural Change*, 61(2):335–367.

Manas Jyoti Bora. 2015. Word order.

Aankit Das, Samarpan Guha, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar. 2020. A hybrid meta-heuristic feature selection method for identification of indian spoken languages from audio signals. *IEEE Access*, 8:181432–181449.

Pranjal Das and Kalyanee K Baruah. 2014. Assamese to english statistical machine translation integrated with a transliteration module. *International Journal of Computer Applications*, 100(5).

Maibam Indika Devi and Bipul Syam Purkayastha. 2023. An exploratory study of smt versus nmt for the resource constraint english to manipuri translation. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 329–338. Springer.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. 2017. Machine translation approaches and survey for indian languages. *arXiv preprint arXiv:1701.04290*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-english machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88.

Lenin Laitonjam and Sanasam Ranbir Singh. 2022. A hybrid machine transliteration model based on multi-source encoder–decoder framework: English to manipuri. *SN Computer Science*, 3:1–18.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent & Fuzzy Systems*, 42(5):4727–4738.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021. Neural machine translation for low resource assamese–english. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 35–44. Springer.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Muskaan Singh, Ravinder Kumar, and Inderveer Chana. 2021. Machine translation systems for indian languages: review of modelling techniques, challenges, open issues and future research directions. *Archives of Computational Methods in Engineering*, 28:2165–2193.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, 209:118187.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-english example based machine translation system. *Int. J. Comput. Linguistics Appl.*, 1(1-2):201–216.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.