A Comparative Study of Transformer and Transfer Learning based MT models for English-Manipuri

Kshetrimayum Boynao Singh[1], **Ningthoujam Avichandra Singh**[1], **Loitongbam Sanayai Meetei**[1], **Ningthoujam Justwant Singh**[1], **Sivaji Bandyopadhyay**[2], and **Thoudam Doren Singh**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
[2]Dept. of CSE, Jadavpur University, India
{boynfrancis,avichandra0420,loisanayai,njustwant92,sivaji.cse.ju,thoudam.doren}@gmail.com

Abstract

In this work, we focus on the development of machine translation (MT) models of a low-resource language pair viz. English-Manipuri. Manipuri is one of the eight scheduled languages of the Indian constitution. Manipuri is currently written in two different scripts: one is its original script called Meitei Mayek and the other is the Bengali script. We evaluate the performance of English-Manipuri MT models based on transformer and transfer learning technique. Our MT models are trained using a dataset of 69,065 parallel sentences and validated on 500 sentences. Using 500 test sentences, the English to Manipuri MT models achieved a BLEU score of 19.13 and 29.05 with mT5 and OpenNMT respectively. The results demonstrate that the OpenNMT model significantly outperforms the mT5 model. Additionally, Manipuri to English MT system trained with OpenNMT model reported a BLEU score of 30.90. We also carried out a comparative analysis between the Bengali script and the transliterated Meitei Mayek script for English-Manipuri MT models. This analysis reveals that the transliterated version enhances the MT model performance resulting in a notable +2.35 improvement in the BLEU score.

1 Introduction

In an increasingly interconnected world, the role of machine translation cannot be overstated. It serves as a critical bridge for breaking down linguistic barriers and enabling effective communication across diverse cultures and languages. However, the efficacy of machine translation (MT) systems largely depends on the availability of adequate linguistic resources, particularly parallel corpora which are essential for training and fine-tuning neural machine translation models. While widely spoken languages benefit from abundant parallel data, many minority and indigenous languages with scant linguistic resources available for their inclusion in the machine translation landscape are left in the shadows of the digital age.

Manipuri language called Meiteilon is mainly spoken in the state of Manipur which lies in the northeastern part of India. Some speakers exist in the states of Assam, Tripura and Mizoram few speakers are also located in the country like Bangladesh and Myanmar. Manipuri language is also facing the challenge of linguistic resource scarcity. This situation impedes access to information and hampers communication for Manipuri speakers in an increasingly globalized world. Bridging this linguistic gap is not only a matter of preserving cultural heritage but also essential for promoting effective communication, education and access to vital information.

In this paper, we embark on a journey to analyse the challenge of English to Manipuri MT in a low-resource setting. Our approach hinges on the powerful techniques of transfer learning and pre-trained models which have shown remarkable success in natural language processing tasks including machine translation. Transfer learning allows us to harness the knowledge learned from high-resource languages and adapt it to the low-resource English-Manipuri translation task. Our goal is to investigate the potential of these techniques in enhancing the translation quality and fluency for Manipuri, despite the constraints of limited parallel data. By leveraging the wealth of linguistic information embedded in pre-trained models, we aim to bridge the language gap and contribute to the development of language technology for minority languages like Manipuri.

This paper is organized as follows: in Section 2, we provide an overview of related work in the fields of machine translation, transfer learning, and low-resource languages. Section 3 outlines the methodology employed, data preparation, and use of OpenNMT Klein et al. (2017) and mT5 Xue et al. (2020) pre-trained models for English to Manipuri MT. Section 4 provides the results and evaluation of the MT models using automatic metrics and qualitative analysis. In Section 5, we present insightful conclusions derived from the key findings.

2 Related Work

Over the past decade, numerous studies have been conducted on MT models for low-resource languages (Singh et al., 2021b) including unsupervised (Singh and Singh, 2020), transfer learning (Hujon et al., 2023) and multimodal (Gain et al., 2021; Meetei et al., 2023a,c) approaches among others. Singh and Bandyopadhyay (2010) carried out a study on supervised statistical methods in which the authors conducted a persuasive examination of the impact of morphosyntactic information and dependencies in the context of statistical machine translation employing Bengali script.

The field of MT has experienced substantial progress primarily propelled by the introduction of neural machine translation (NMT) models as evidenced by Vaswani et al. (2017). These innovative deep learning techniques have gradually supplanted traditional phrase-based and statistical methods resulting in remarkable enhancements in translation quality. Nonetheless, the effectiveness of these systems critically hinges on the availability of parallel corpora which consist of matching sentences in both the source and target languages (Singh and Singh, 2022). High-resource languages such as English, Spanish and Chinese benefit from extensive parallel dataset, leading to precise and fluent translations. Conversely, low-resource languages (Meetei et al., 2021, 2023b) often spoken by marginalized communities grapple with significant challenges in procuring adequate training data for machine translation.

Transfer learning has the potential to help low-resource languages overcome challenges by leveraging insights from rich languages. This method (Singh et al., 2021a) has gained prominence in natural language processing domains, including machine translation. Techniques like cross-lingual embedding and multilingual pre-trained models have been explored, enabling models to adapt and excel in resource-scarce environments. Cross-lingual embedding involves translating words or phrases from diverse languages into a shared vector space while multilingual pre-trained models capture cross-lingual representations during initial training allowing for fine-tuning on language-specific tasks. These strategies have shown promising results in low-resource machine translation scenarios, providing hope for underrepresented languages, such as minority or indigenous languages which face a critical threat of extinction due to a lack of support for documentation, educational materials and communication tools. This paper explores the potential of transfer learning and pre-trained models to improve English to Manipuri translation in low-resource contexts.

3 Methodology

In this section, we describe the methodology used to develop English-Manipuri MT systems in a low-resource setting (Figure 1). Our approach centers around supervised transformer based MT model and transfer learning based MT model to adapt the specific characteristics of the English-Manipuri language pair.

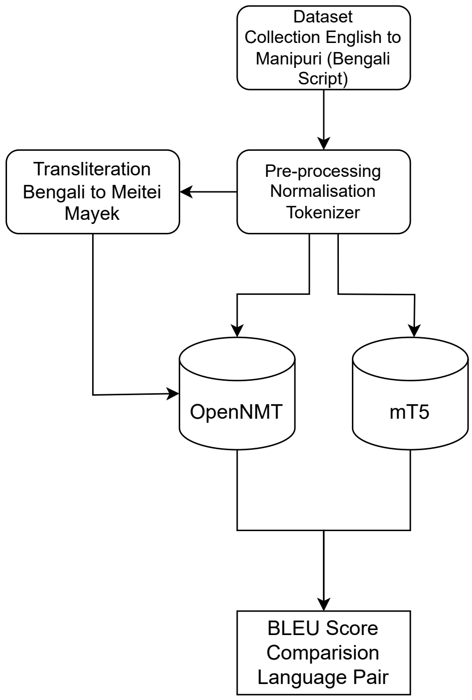

Figure 1: Workflow diagram

3.1 Data preparation

In preparing our parallel corpus for model training, we employed distinct tools for English and Manipuri languages in the Bengali script. For English, we utilized the Moses [1] toolkit, while for Manipuri, we used the IndicNLP library. Our preprocessing

[1] https://pypi.org/project/mosestokenizer/

journey began with language normalization, followed by tokenization. The collected dataset was subjected to a series of standard pre-processing procedures, encompassing tokenization, sentence segmentation, and rigorous cleaning to eliminate any extraneous noise or inconsistencies. Training the dataset is pre-processing with subword tokenization. For subword-based tokenization, we use a source and target BPE of 15000 subword tokens or vocabularies using sentence pieces over the parallel training dataset and apply them to the remaining testing and validation dataset. The subword tokenization (Sennrich et al., 2016) is carried out using the subword-nmt [2] tool.

Language	Sentence	Word	Average
Eng Train	69065	1494709	21
MniB Train	69065	1252459	18
Eng Valid	500	8335	16
MniB Valid	500	7145	14
Eng Test	500	8570	17
MniB Test	500	7324	14

Table 1: Figures from the experimental dataset for English to Manipuri (MniB) with Bengali script

The Manipuri text is written in Bengali script. Statistics of the training dataset are shown in Table 1. We collect parallel data comprising English-Manipuri sentence pairs from WMT23[3] shared task (Singh et al., 2023) and BPCC[4]. Table 2 presents the statistics of the dataset after transliterating the Manipuri text from Bengali to Meitei Mayek script using a rule-based transliteration approach.

Language	Sentence	Word	Average
MniM Train	69065	1478491	21
MniM Valid	500	7514	15
MniM Test	500	7324	14

Table 2: Figures from the experimental dataset for English to Manipuri with Meitei Mayek (MniM) script

3.2 OpenNMT

This MT model is a supervised transformer based model (Vaswani et al., 2017). The model is trained for 300000 steps and validated after every 5000 steps. We set the parameter of batch type to tokens and batch size to 2048. The models are trained using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2 and the dropout set to 0.1. The early stopping mechanism is employed where the training is stopped when the accuracy does not improve for 30 consecutive validations. In our transformer-based model, each source encoder-decoder also has 4 layers, with a word vector size of 512 and a shared encoder and decoder embedding.

3.3 mT5

This MT system involves fine-tuning the mT5 (Multilingual Translation) model (Xue et al., 2020) for English to Manipuri translation in a low-resource context. Transfer learning is employed to fine-tune models like mT5, a multilingual pre-trained model variant of Text-to-Text Transfer Transformer (T5), for low-resource scenarios. The mT5-base model is fine-tuned using the simple-transformers library and fine-tuned for 30000 training steps with the 5 epochs, Train batch size, and evaluation batch size of 10. We used the FLORES development set flores200 (NLLB Team, 2022) dataset with mT5-base model ”google/mt5-base”[5] is initialized with learned weights and adapted to the English-Manipuri translation task. Task-specific fine-tuning involves training the model on the curated English-Manipuri parallel dataset, using standard NMT training procedures and regularization techniques to prevent overfitting and enhance the model’s generalization ability and robustness.

The model’s base architecture is pre-trained on a vast multilingual corpus, capturing cross-lingual transferable knowledge. The fine-tuning procedure involved pre-trained mT5 model on the assembled English-Manipuri dataset and checkpoints were saved to ensure that the model could be restored for evaluation. We chose the mT5 model for our experiments because of its versatility and effectiveness in multilingual translation tasks. mT5 is a transformer-based model that has demonstrated excellent performance in a variety of language pairs.

4 Results and Discussion

4.1 BLEU Score

The BLEU (Bilingual Evaluation Understudy) score is a widely used metric for assessing the qual-

[2] https://github.com/rsennrich/subword-nmt
[3] https://www2.statmt.org/wmt23/indic-mt-task.html
[4] https://ai4bharat.iitm.ac.in/bpcc/
[5] https://huggingface.co/google/mt5-base

ity of machine translations. It measures the overlap between the generated translations and reference translations in terms of n-grams (typically up to 4-grams). Higher BLEU scores indicate better translation quality.

Model	BLEU Score
Eng-MniB(mT5)	19.13
Eng-MniB(OpenNMT)	29.05
MniB-Eng(OpenNMT)	30.90
Eng-MniM(OpenNMT)	33.25

Table 3: BlEU Score of the systems for Manipuri-Bengali-Script (MniB) and Manipuri-Meitei-Mayek-Script (MniM)

The fine-tuned mT5 model achieved a substantially lower BLEU score compared to the OpenNMT model as shown in Table 3. This result underscore the substantial performance advantage of the transformer based MT model over the transfer learning based MT model for the English-Manipuri language pair.

4.2 Qualitative analysis

In addition to automated metrics, we conducted human evaluations to assess the fluency, accuracy and naturalness of the translations.

In addition to automated metrics, we conducted human evaluations to assess the fluency, accuracy and naturalness of the translations. Manipuri is a language with significant linguistic variation and automated ratings for English to Manipuri translations are often lower than those for Manipuri-to-English translations, despite the availability of an acceptable translation output. As a result, we solicit the assistance of a bilingual native Manipuri speaker fluent in English to evaluate the translation. outputs for the English to Manipuri job. Tables 4 present three randomly chosen samples from the test set for each of the Manipuri to English and English to Manipuri translation models for subjective evaluation.

Table 4 shows the results of the MT system for translating English to Manipuri. The 1mT5 Output, 1OpenNMTOutput, and 1MniOutputM give a very close meaning to each other with reference sentences. In the Output 2 section, the words "হন্থবা, ৱাংপা, য়ামদবা" are all synonymous. But, in context of English translation "Lack of" and "decreased" are terms used to describe a reduction or insufficiency of something. However, they are used in slightly different contexts in the output. In the context of Meitei Mayek transliteration the word "ꯄꯨꯝꯕꯦ" is not included in the output.

In output3 translation OpenNMT translates similar meaning and expression with the references while mt5 translates similar meaning with different expression. To enhance the quality of translation from Manipuri to English, we strive to improve the accuracy and semantic richness of the machine-generated output. In Transliteration Meitei Mayek word "ꯁꯣꯏꯗꯅꯥ ꯐꯪꯒꯗꯕꯅꯤ" signifies an absolute requirement while "ꯆꯪꯒꯗꯕꯅꯤ" suggests a take care or preference. In the back translation 'Must receive' conveys an uncompromising obligation, while 'should have access' reflects a thoughtful recommendation or desirable option. The translation model demonstrates superior performance with shorter sentences, However, when it comes to longer sentences, it faces challenges.

5 Conclusion

In this paper, we report a comprehensive study by implementing MT models using transformer based model (OpenNMT) and transfer learning technique using pretrained model (mT5) for English-Manipuri in a low resource setting. The experimental results demonstrate that OpenNMT model outperform the mT5 model by a factor of 9 in terms of BLEU score. The performance of Manipuri to English using OpenNMT model report a BLEU score of 30.90 using Bengali script. OpenNMT model shows a promising results of BLEU score of 33.25 for English to Manipuri languages pair using transliterated Meitei Mayek script. Despite the fact that transfer learning technique emerge as an approach to build machine translation models for resource-constrained languages by leveraging pre-trained models, the study's findings indicate that this technique performs less effectively than the transformer model trained on the same parallel dataset. Future research may be needed to explore the nuances of transfer learning in this specific domain and to identify potential refinements or enhancements to improve its efficacy for machine translation in resource-constrained languages.

Based on subjective evaluation, the translation quality is generally deemed satisfactory taking into account of the small dataset and the use of a single test reference dataset. This research underscores the potential of supervised transformer based model and transfer learning technique using

Particulars of Input/Output and Models	Sample Input and Output
	Sample 1
1Eng (English)	long term measures
1MniB(Manipuri in Bengali Script)	অশাংবা মতমগী ওইবা থৌরাংশীং
1MniM(Manipuri in Meetei Mayek)	ꯑꯁꯥꯡꯕꯥ ꯃꯇꯝꯒꯤ ꯑꯣꯏꯕꯥ ꯊꯧꯔꯥꯡꯁꯤꯡ
Eng to MniB	
1mT5 Output	মতম শাংনা পাংথোকপা পান্বৈশিং
1OpenNMT Output	অশাংবা মতমগী ওইবা পান্বৈশীং
Eng to MniM	
1Mni OutputM(Meetei Mayek)	ꯑꯁꯥꯡꯕꯥ ꯃꯇꯝꯒꯤ ꯑꯣꯏꯕꯥ ꯄꯥꯅ꯭ꯕꯩꯁꯤꯡ
MniB to Eng	
1OpenNMT BT Output:	long term measures
	Sample 2
2Eng	lack of financial resources
2MniB	ফাইনান্সিএল রিসোর্সশীং হন্থবা
2MniM	ꯐꯥꯏꯅꯥꯅ꯭ꯁꯤꯑꯦꯜ ꯔꯤꯁꯣꯔ꯭ꯁꯁꯤꯡ ꯍꯅ꯭ꯊꯕꯥ
Eng to MniB	
2mT5 Output	ফাইনান্সিএল রিসোর্সশীং ৱাৎপা
2OpenNMT Output	ফাইনান্সিএল রিসোর্সশীং য়ামদবা
Eng to MniM	
2Mni OutputM	ꯐꯥꯏꯅꯥꯅ꯭ꯁꯤꯑꯦꯜ ꯔꯤꯁꯣꯔ꯭ꯁꯁꯤꯡ
MniB to Eng	
2OpenNMT BT Output	decreased financial resources,
	Sample 3
3Eng	every pregnancy is special and every pregnant woman must receive special care .
3MniB	মীরোনবা খুদিংমক মরুওই অমসুং মীরোনবী নুপী খুদিংমক্না অখন্নবা কেয়র শোয়দনা ফংগদবনি।
3MniM	ꯃꯤꯔꯣꯟꯕꯥ ꯈꯨꯗꯤꯡꯃꯛ ꯃꯔꯨꯑꯣꯏ ꯑꯃꯁꯨꯡ ꯃꯤꯔꯣꯟꯕꯤ ꯅꯨꯄꯤ ꯈꯨꯗꯤꯡꯃꯀ꯭ꯅꯥ ꯑꯈꯅ꯭ꯅꯕꯥ ꯀꯦꯌꯔ ꯁꯣꯌꯗꯅꯥ ꯐꯪꯒꯗꯕꯅꯤ ꯫
Eng to MniB	
3mT5 Output	মীরোনবা খুদিংমক অখন্নবা ওই অমসুং মীরোনবী নুপী পুম্নমক্না অখন্নবা কেয়র লৌগদবনি।
3OpenNMT Output	মীরোনবা খুদিংমক মরুওই অমসুং মীরোনবী নুপী খুদিংমক্না অখন্নবা কেয়র ফংহনবা তাই।
Eng to MniM	
3Mni OutputM	ꯃꯤꯔꯣꯟꯕꯥ ꯈꯨꯗꯤꯡꯃꯛ ꯃꯔꯨꯑꯣꯏ ꯑꯃꯁꯨꯡ ꯃꯤꯔꯣꯟꯕꯤ ꯅꯨꯄꯤ ꯈꯨꯗꯤꯡꯃꯀ꯭ꯅꯥ ꯑꯈꯅ꯭ꯅꯕꯥ ꯀꯦꯌꯔꯗꯕ ꯂꯧꯒꯗꯕꯅꯤ ꯫
MniB to Eng	
3OpenNMT BT Output	every pregnancy is special and every pregnant woman should have access to specialized facilities .

Table 4: Sample input and output of the English-Manipuri MT systems

pre-trained models in addressing challenges posed by low-resource languages like Manipuri. Thus, this work showcases the capabilities of transformer based model as an effective approach as compared to mT5 for English-Manipuri MT system.

Acknowledgements

This research is supported by MeitY (Govt. of India) Ref. No. 11(1)/2022-HCC(TDIL)-Part(4). We are grateful to the CNLP lab and Dept. of CSE at NIT Silchar for providing computing resources.

References

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44.

Aiusha V Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2023. Transfer learning based neural machine translation of english-khasi on low-resource settings. *Procedia Computer Science*, 218:1–8.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Loitongbam Sanayai Meetei, Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. Do cues in a video help in handling rare words in a machine translation system under a low-resource setting? *Natural Language Processing Journal*, 3:100016.

Loitongbam Sanayai Meetei, Salam Michael Singh, Alok Singh, Ringki Das, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023b. Hindi to english multimodal machine translation on news dataset in low resource setting. *Procedia Computer Science*, 218:2102–2109.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. Low resource multimodal neural machine translation of english-hindi in news domain. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 20–29.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023c. Exploiting multiple correlated modalities can enhance low-resource machine translation quality. *Multimedia Tools and Applications*, pages 1–21.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Salam Michael Singh, Loitongbam Sanayai Meetei, Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021a. On the transferability of massively multilingual pretrained models in the pretext of the indo-aryan and tibeto-burman languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 64–74.

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

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. *International Journal of Computational Linguistics and Applications*, pages 201–216.

Thoudam Doren Singh, Cristina España i Bonet, Sivaji Bandyopadhyay, and Josef van Genabith. 2021b. Proceedings of the first workshop on multimodal machine translation for low resource languages (mmtlrl 2021). In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.