NITS-CNLP Low-Resource Neural Machine Translation Systems of English-Manipuri Language Pair

Kshetrimayum Boynao Singh[1], **Ningthoujam Avichandra Singh**[1], **Loitongbam Sanayai Meetei**[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,sivaji.cse.ju,thoudam.doren}@gmail.com

Abstract

This paper describes the transformer-based Neural Machine translation (NMT) system for the Low-Resource Indic Language Translation task for the English-Manipuri language pair submitted by the Centre for Natural Language Processing in National Institute of Technology Silchar, India (NITS-CNLP) in the WMT 2023 shared task. The model attained an overall BLEU score of 22.75 and 26.92 for the English to Manipuri and Manipuri to English translations respectively. Experimental results for English to Manipuri and Manipuri to English models for character level n-gram F-score (chrF) of 48.35 and 48.64, RIBES of 0.61 and 0.65, TER of 70.02 and 67.62, as well as COMET of 0.70 and 0.66 respectively are reported.

1 Introduction

Our team from Centre for Natural Language Processing at National Institute of Technology Silchar, India (NITS-CNLP) participated in Low-Resource Indic Language Translation task for English-Manipuri language pair in the WMT 2023 shared task (Pal et al., 2023). The shared task involves developing Machine Translation (MT) systems with relatively small parallel datasets. Neural Machine translation (NMT) has been a trending topic for the last few years for translating human languages. Manipuri's MT task is still in its infancy because of the limited resources. Singh and Bandyopadhyay (2011a) conducted a study on supervised statistical methods in which the authors present a convincing study of the impact of morphosyntactic information and dependencies in the context of statistical machine translation. In another work, Singh and Bandyopadhyay (2011b) showed that the expression grounded Statistical Machine translation (SMT) system improves by incorporating verbal features including named entities and reduplicated multiword expressions. Despite the advancement in MT tasks, its investigation in low-resource languages is limited. MT researchers have introduced several approaches to overcome this bottleneck such as data augmentation using back-translation (Sennrich et al., 2016a), multilingual approach (Singh and Singh, 2022a), semi-supervised approach (Cheng et al., 2016; Singh and Singh, 2022b) and exploiting cues from multiple modalities (Gain et al., 2021; Meetei et al., 2023). There are also reports of a comparative study of MT systems on the low resource machine translation focusing on Indian languages such as Assamese (Baruah et al., 2021) and Mizo (Devi et al., 2022; Thangkhanhau and Hussain, 2023).

Driven by the benefits of NMT over traditional MT systems and the encouraging outcomes achieved by NMT in recent times, a study to assess its efficacy in the domain of Indian languages is conducted. Specifically, we have developed and assessed NMT models for translating English to Manipuri and Manipuri to English. The predicted translations are evaluated using automatic evaluation metric and qualitative analysis.

1.1 About the language

Manipuri is the lingua franca of Manipur and has been in existence since 2000 years back till present times with records preserved in the classical cultural heritage of literature. Manipuri is a language of Tibeto-Burman sub-family of the Sino-Tobentan languages family which is locally called as Meeiteilon/Meiteilon (hereon Meiteilon). It is one of the 22 official languages of the India included in the 8th schedule of the Indian constitution[1]. Meiteilon had its original script named Meitei/Meetei Mayek (hereon Meitei Mayek) which was in use up to the 18th century and was replaced later with the Bengali script. However, the wave of revivalist movement

[1]https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 967–971

emerged later leading to the formation of Meitei Mayek Advisory committee in the year 1973. In 1982, the Government of Manipur announced its decision to include Manipuri in the school education and efforts to revive the Meitei Mayek are still on.

2 System Overview

2.1 Dataset

Language	Sentence	Word	Avg
Eng-Training	21687	390730	18
Man-Training	21687	330319	15
Eng-Validation	1000	16905	16
Man-Validation	1000	14469	14
Eng-Testing	1000	14886	14
Man-Testing	1000	12775	12

Table 1: Statistics of the experimental dataset. (Avg = Average Sentence length

The Manipuri text is written in Bengali script. Statistics of the training dataset are shown in Table 1.

2.2 Data preparation

Training the dataset is pre-processing with subword tokenization. For subword based tokenization we use a source and target BPE of 10000 subword tokens or vocabularies using sentences pieces over the parallel training dataset and applied to the remaining testing and validation dataset. The subword tokenization (Sennrich et al., 2016b) is carried out using the subword-nmt tool [2].

2.3 MT model

Our MT models are trained using OpenNMT toolkit (Klein et al., 2017) and is based on the transformer model (Vaswani et al., 2017).

2.4 Model parameter

Our models are 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. 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 has 4 layers and decoder also has 4 layers, with a word vector size of 512 and a shared encoder and decoder embedding. We measure the performance of our models by using BLUE (Papineni et al., 2002) and chrF (Popović, 2015).

3 Results and Discussion

In this section, we discuss the experimental results and the performance of models. The reported BLEU score is calculated upon the de-tokenized text using sacrebleu (Post, 2018) while remaining score such as chrF, RIBES, TER and COMET are calculated using tool provided by the organizer[3]. The scores of the system are given in Table 2. The Manipuri to English translation model obtained a BLUE score of 26.92 while English to Manipuri obtained 22.75.

3.1 Qualitative analysis

Automated evaluation metrics such as BLEU measures the precise lexical matches between the translated output and the reference sentences. However, it is inevitable for a natural language to exhibit linguistic variations in terms of synonyms and transformations between active and passive mode of communication. As a result, despite preserving the intended meaning of the source sentence in the translation output, the automated score based on the n-gram match suffers. Manipuri is a language of considerable linguistic diversity and the automated scores for English to Manipuri translations are typically lower than those for Manipuri to English despite the provision of a translation output of acceptable quality. Therefore, we enlist the services of a bilingual native speaker of Manipuri with fluency in English to evaluate the translation outputs for the English to Manipuri task. Table 3 and Table 4 present four randomly selected source sentences from the test set for each of the Manipuri to English and English to Manipuri translation models along with their corresponding output sentences with reference sentences to carry out subjective evaluation.

In Table 3, the difference between the reference and MT results are reported. For Source1, Source2 and Source4, the Manipuri to English MT system outputs (OutputE1, OutputE2 and OutputE4) are close to the reference sentence. In OutputE2, the sentence formation is incorrect where

[2]https://github.com/rsennrich/subword-nmt

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

MT system	BLEU	chrF	RIBES	TER	COMET
en-mni	22.75	48.35	0.61	70.02	0.70
mni-en	26.92	48.64	0.65	67.62	0.66

Table 2: BLEU score and the character n-gram F-score (chrF), RIBES, TER and COMET values of the English→Manipuri (en-mni) and Manipuri→English (mni-en) translation model.

Result	Samples
Source1:	দোক্টর অমা খুদক্তা কৌবীয়ু , নত্ত্রগা অনাবদু হোস্পিটালদা পুনবা এম্বুলেন্স অমা থৌরাং তৌবীয়ু ।
References1:	*call a doctor immediately , or arrange for an ambulance to take the casualty to hospital*
OutputE1:	**send for a doctor immediately , arrange for an ambulance to take the causality to hospital.**
Source2:	ই-বোক্স অসি শিজিন্নবদা য়াম্না লাই অমসুং কমপ্লেন্ত অদুগী অরোন-অথুপ অদু ঙাক্কুনা থম্বদা মতেং পাংগনি ।
References2:	*e-box is very simple to operate and will help to maintain the confidentiality of the complaint.*
OutputE2:	**the e-box will help you protect and will help with regard to confidentiality of the complaint.**
Source3:	মসি বেঙ্গলোর মেথদগী ওন্ন-তৈনবনি ।
References3:	*it is the reverse in bangalore method.*
OutputE3:	**it is to be done from bangalore method.**
Source4:	ভারত্তা অহল ওইরবা মীওইশীংগী মীশীং অসি লেপ্তনা হেনগৎলক্লি ।
References4:	*there has been a steady rise in the population of older persons in india.*
OutputE4:	**the number of older persons has been increasing**

Table 3: Sample input and output of the Manipuri to English MT system.

Result	Samples
Source1:	*encouraging the appropriate government to assume the fullest responsibility for the administration of occupational safety , health and environment at workplace*
References1:	সেফটি , হেলথ এন্ড এনভাইরনমেন্ট এট ৱার্কপ্লেসকী মতাংদা মতিক চাবা লেজিস্লেসন অমা শেমবা
OutputM1:	সেফটি , হেলথ এন্ড এনভাইরনমেন্ট এট ৱার্কপ্লেসকী মতাংদা মতিক চাবা লেজিস্লেসন অমা শেমগৎপা ।
Source2:	*the chairperson of the national authority shall preside over the meetings of the national authority .*
References2:	নেসনেল ওথোরিটিগী চিয়ারপর্সননা নেসনেল ওথোরিটিগী মিটিংশীংগী থৌরম মপু ওইগনি ।
OutputM2:	নেসনেল ওথোরিটিগী চিয়ারপর্সননা নেসনেল ওথোরিটিগী মীটিং অদুগী মীফম পাংথোক্কনি ।
Source3:	*this causes pain .*
References3:	মসিনা নাবা ফাউহল্লি ।
OutputM3:	মসিনা নাবা থোকহল্লিবা মরমশীং
Source4:	*compensation for accredited social health activist*
References4:	এক্রেদিতেদ সোসিএল হেলথ এক্টিভিষ্ট গী কম্পেন্সেসনট
OutputM4:	এক্রেদিতেদ সোসিএল হেলথ এক্টিভিষ্ট গীদমক কম্পেন্সেসন পীবা

Table 4: Sample input and output of the English to Manipuri MT system.

the words such as “maintain” and “operate” are incorrectly translated. In OutputE1, we observe a case where the word “call” is translated to its antonym “send” and “casualty” to “causality”. In OutputE4, “steady rise” is translated as “increasing” which could be considered as a synonym of the phase. Apart from the missing words “in india”, the output sentence preserve the meaning of the source sentence. In OutputE3, the MT output is not able to retain the intended meaning of the source sentence.

The Table 4 shows the results of the MT system for translating English to Manipuri. The OutputM1, OutputM3 and OutputM4 give a close meaning to the reference sentence. In OutputM1, the word “শেমবা” (meaning “build”) is translated to its infinitive form of the verb “শেমগৎপা” (meaning “to build”). The word “causes” has multiple translations in Manipuri such as “ফাউহল্লি”, “থোকহল্লিবা”and “মরমশীং” which are used in different context. In OutputM3, we observe that the translations of the word “causes” is repeated showing the challenges of the MT model in translating such words. In OutputM4, we observe a case where a word as multiple translation in Manipur but can be used in the same context. The word “for” can be translated as “গী” or “গীদমক” in Manipuri. Apart from the extra verb “পীবা” (meaning “give”), OutputM4 is grammatically correct despite not having a perfect n-gram match. In the case of OutputM2, the sentence is observe to have a poor adequacy with the incorrect translation for the word “chairperson” but the structure of sentences is well formed and grammatically correct.

4 Conclusion

Enabling MT for low-resource languages poses several challenges due to the lack of parallel resources available for training. In this work, we report the performance of the MT systems trained on low resource setting for English to Manipuri and Manipuri to English using a transformer-based encoder and decoder architecture. The automatic evaluation shows that English to Manipuri MT system achieved 22.75 BLEU and Manipuri to English MT system achieved 26.92 BLEU. The automated scoring mechanism is inadequate in capturing the linguistic nuances of the morphologically complex Manipuri language which requires the use of multiple references. Based on the subjective evaluation, we observed that the translation quality is deemed satisfactory and fluent in some cases, given the relatively small size of the dataset and the utilization of a single test reference.

Limitations

Translation model performs better for short sentences as compared to the longer sentences. There are several out of vocabulary words due to the fact that the model is built on a constraint environment.

Acknowledgements

This work is sponsored by MEITY Ref. No. 11(1)/2022-HCC(TDIL)-Part(4). We also acknowledge CNLP (Centre for Natural Language Processing) and Department of Computer Science and Engineering at National Institute of Technology Silchar for providing and giving access to the computing facilities.

References

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–32.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974.

Chanambam Sveta Devi, Bipul Syam Purkayastha, and Loitongbam Sanayai Meetei. 2022. An empirical study on english-mizo statistical machine translation with bible corpus. *International journal of electrical and computer engineering systems*, 13(9):759–765.

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.

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 Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Salam Michael Singh and Thoudam Doren Singh. 2022a. An empirical study of low-resource neural machine translation of manipuri in multilingual settings. *Neural Computing and Applications*, 34(17):14823–14844.

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

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011a. Bidirectional statistical machine translation of manipuri english language pair using morpho-syntactic and dependency relations. *International Journal of Translation*, 23(1):115–137.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011b. Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system. In *Proceedings of 5th international joint conference on natural language processing*, pages 1304–1312.

Haulai Thangkhanhau and Jamal Hussain. 2023. Construction of mizo–english parallel corpus for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

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.