Neural Machine Translation for a Low Resource Language Pair: English-Bodo

Parvez Aziz Boruah , Kuwali Talukdar , Mazida Akhtara Ahmed and **Kishore Kashyap**
Department of Information Technology
Gauhati University
Guwahati, Assam
parvezaziz70@gmail.com , kuwalitalukdar@gmail.com ,
14mazida.ahmed@gmail.com , kb.guwahati@gmail.com

Abstract

This paper represent a work done on Neural Machine Translation for English and Bodo language pair. English is a language spoken around the world whereas, Bodo is a language mostly spoken in North Eastern area of India. This work of machine translation is done on a relatively small size of parallel data as there is less parallel corpus available for english bodo pair. Corpus is generally taken from available source National Platform of Language Technology(NPLT), Data Management Unit(DMU), Mission Bhashini, Ministry of Electronics and Information Technology and also generated internally in-house. Tokenization of raw text is done using IndicNLP library and mosesdecoder for Bodo and English respectively. Subword tokenization is performed by using BPE(Byte Pair Encoder) , Sentencepiece and Wordpiece subword. Experiments have been done on two different vocab size of 8000 and 16000 on a total of around 92410 parallel sentences. Two standard transformer encoder and decoder models with varying number of layers and hidden size are build for training the data using OpenNMT-py framework. The result are evaluated based on the BLEU score on an additional testset for evaluating the performance. The highest BLEU score of 11.01 and 14.62 are achieved on the testset for English to Bodo and Bodo to English translation respectively.

1 Introduction

English is a widely spoken language around the glove and generally forms an official language for most of the people. Bodo is a local language of Bodo people spoken in Northeast of India (mostly in the state of Assam). Bodo language used the devanagari script for writing. Recently, most of the work are going on for Bodo language(Brahma et al., 2012). Since Bodo is a low resource language, there are less digital resources for Bodo language in the form of text corpus. This become quite challenging for Natural Language Processing(NLP) task for Bodo language like machine translation. Here, we performed a machine translation for English to Bodo and vice versa with limited amount of parallel corpus. This corpus is composed of parallel data from National Platform of Language Technology(NPLT), Data Management Unit(DMU), Mission Bhashini, Ministry of Electronics and Information Technology and our own in-house created parallel corpus.

Preprocessing work that includes normalization,tokenization, subword tokenization are performed sequentially on both the languages and then trained on two transformer encoder-decoder models. These two encoder-decoder models varies among each other basically in terms of number of layers and hidden size. OpenNMT-py framework(a framework containing libraries and tool for performing machine translation) is used for build and training the model(Klein et al., 2017). Normalization, i.e, conversion of all upper-case alphabets to lower-case is done on the English Text.

For tokenization of words in the text corpus of English and Bodo language mosesdecoder and IndicNLP library is been used respectively. Subword tokenization which is widely used for processing of the tokenized text into sub-word level is performed. For subword tokenization three techniques have been used i.e, BPE(Byte Pair Encoder) , Sentencepiece and Wordpiece subword. Two vocabulary size of 8000 and 16000 subwords are created from the data generated from each of the subword tokenizations. The two transformer models are trained individually using data generated by each of the subword tokenization(BPE, Sentencepiece and Wordpiece subword) technique. Training is done on both the directions(i.e, English to Bodo and Bodo to English) and using both the vocabulary sizes. The result are evaluated based on BLEU using sacredbleu library.

2 Related Works

Vaswani et al. (2017) have proposed the transformer model, an encoder decoder model which is solely based on attention mechanism. In their expirement this architecture is found to be efficient than recurrent model for NLP task. This transformer architecture is used by us for our translation model. Verma and Bhattacharyya (2017) in their paper have given a detail survey of NMT and the model and architecture used for NMT like LSTM, encoder-decoder model, attention mechanism. They have addressed two problems in NMT task. One is poor performanes for long sentences which can be solved by attention mechanism and OOV(out of vocabulary) which can be solved by Subword BPE tokenization. And (Tan et al., 2020) have given a detail specification of NMT models, architecture, tools, frameworks available for performing NMT tasks and other preprocessing libraries. Klein et al. (2017) in their paper have described an open-source toolkit for neural machine translation (NMT), OpenNMT. It is a framework for building Neural Machine Translation model and it prioritizes efficiency and modularity. In our experiment we have used OpenNMT for building our NMT model.

Choudhary et al. (2018) have performed an experiment to translate English to Tamil language using neural machine translation model with word embedding and byte pair encoding tokenization. Their model had been able to achieve a highest blue score of 8.33.

Indian Languages like Bodo, Assamese and most of the Northeastern Indian languages are comparatively new to NLP research. In recent years various works including standardization, developement of tools and technology, corpus development, wordnets, annotations etc. have been carried out for Bodo and Assamese languages(Sarma et al. (2010), Bhuyan and Sarma (2018), Sarma et al. (2012), Talukdar and Sarma (2023)). Works related to Machine Translation and Neural Machine Translation have also started in Assamese languages(Baruah et al. (2014), Hannan et al. (2019), Talkukdar et al. (2023)). Islam and Purkayastha (2018) have build a machine translation system for Bodo to English through the process of Bodo to English Machine Transliteration system. They have used the phase based statistical machine translation method for developing their model. Kalita et al. (2023) and Ahmed et al. (2023a) in their respective papers have shown some techniques for preprocessing and modeling of resources that are required for performing a neural machine translation training for English-Bodo and English-Assamese language pairs respectively.As the above research dicussed briefly about Preprocessing of the language pairs and later in Boruah et al. (2023) and Ahmed et al. (2023b) highlights multiple NMT models with varing hyperparameters for English-Bodo and English-Assamese respectively.

3 Methodology

3.1 Data Required

Data required for machine translation task is parallel corpus. Parallel corpus contains a collection of original texts in language L1 and their translations into an another languages L2(Stahlberg, 2019). We have performed experiments in both the direction(i.e, L1 to L2 and vice versa). The dataset is composed of parallel data of 67999 parallel sentences provided by National Platform for Language Technology(NPLT), 10000 taken from Data Management Unit(DMU), Mission Bhashini, Meity and 14411 parallel sentences created in-house so it becomes a total of 92410. A total of 600 sentences are been randomly derived from the dataset as validation set which is been used for validation and rest 91810 for training.

An additional corpus of 500 parallel English-Bodo sentences is created in-house for performing an evaluation of our NMT model. The result are evaluated based on BLEU score of this testset. This testset contents sentences from a diverse domain of administration, law, agriculture, education, health, technical and tourism. The results are given in table 3 and 4.

3.2 Preprocessing

Pre-Processing is an important part in most of the NLP tasks. Preprocessing generally is the arrangement or formatting of the raw text after which the training model can accept. Here we have performed normalization in the english text set. These normalization is done to have a uniformity among all the alphabets by having only lower-case alphabets in the text corpus. Normalization is not required for Bodo text as Bodo language used Devanagiri script for writing text and Devanagiri script do not have the concept of lower-case/upper-case alphabet.

After normalization, tokenization is required and is done by using IndicNLP library for Bodo lan-

Table 1: Subword Tokenization examples

Original text	wayanad is about 280 kms away from bengalooru .
Text after BPE subword tokenization	w@@ ay@@ an@@ ad is about 2@@ 80 kms away from beng@@ alo@@ or@@ u .
Text after Sentencepiece subword tokenization	_way ana d _is _about _2 80 _km s _away _from _bengal oor u _.
Text after Wordpiece subword tokenization	way ##ana ##d is about 28 ##0 kms away from bengal ##oor ##u .

guage and mosesdecoder library for English language. Performing tokenization, each words and punctuation are split with space.And each words and punctuations are treated as an individual tokens. For eg, the sentence "He is a good, honest and kind boy." after tokenization becomes "He is a good , honest and kind boy ." and each word and punctuation is considered as one token,i.e, ['He' ; 'is' ; 'a' ; 'good' ; ',' ; 'honest' ; 'and' ; 'kind' ; 'boy' ; '.']

After tokenization, subword tokenization is done on the tokenized text. The main concept of subwords is that frequent words are to be included as a word in the vocabulary, whereas rare words should be split into frequent sub-words. These subwords can merged with another subwords to form new words. Subword tokenization helps in reducing the vocabulary size and handling of unknown tokens (out of vocabulary). Three methods of subword tokenization have been explored.i.e, BPE(Byte Pair Encoder) , Sentencepiece and Wordpiece subword. The vocabulary of the model is created in this phase where each words and subwords are listed in the vocabulary file. The size of the vocabulary can be given while performing the subword tokenization.We have considered two vocabulary sizes of 8000 and 16000 on each of the three subword tokenization process and then training is performed individually by taking these two different vocabulary(Kudo and Richardson, 2018)(Song et al., 2020). Table 1 shows a raw text in different subword tokenized format.

3.3 Machine Translation Model

After performing the preprocessing step our data is ready as an input to the Neural Machine Translation(NMT) model. We built two NMT model using OpenNMT framework. OpenNMT is an open source framework for neural machine translation and neural sequence learning tasks. It is developed in two version: one using the tensorflow library and other using the pytorch library. For our experiments we have taken the pytorch version of OpenNMT(OpenNMT-py). We built two NMT model based on transformer encoder-decoder architecture. In Model 1 the encoder and decoder both consist of 3 layers each and in Model 2 the encoder and decoder both consist of 6 layers each(Klein et al., 2017). The hyperparameters use for our model are given in Table 2.

Both the models are trained with the subword tokenized text file. Also the validation file is given to the models to performed a validation after each 10000 steps. Other files that need to be given to the models are the source and target vocabulary files i.e, the english and bodo vocabulary files. We have trained the model for a total of 100000 steps. Each step consist of a fix batch size of 256 tokens. The training and validation batch size is taken as 256 and 512 for Model 1 and Model 2 respectively with a batch type of tokens i.e, each batch consist of 256/512 subword tokens. A validation is performed with the validation data after an interval of 10000 training steps. Also at each 10000 steps interval a checkpoint is saved and the checkpoint with highest accuracy value is considered for testing the test data. Both Model 1 and Model 2 are trained individually using data generated by each of the subword tokenization(i.e, BPE, Sentencepiece

Table 2: Parameters/hyper-parameters value

Parameters	**Values for Model 1**	**Values for Model 2**
Number of Encoder Layers	3	6
Number of Decoder Layers	3	6
No. of attention heads	4	8
Encoder hidden layer size	256	512
Decoder hidden layer size	256	512
Transformer feedforward size	1024	2048
Word Vector size	256	512

and Wordpiece subword technique) on both the language direction. Both vocabulary sizes of 8000 and 16000 are considered and trained individually on all the models.

After training of the models, the testset is translated using the checkpoint with highest validation accuracy. Before translating the testset, all the preprocessing is required for the testset, ie, normalization, tokenization, subword tokenization. The subword tokenized file of the source language is translated using the train model which gives a machine translated subword file for the target language. This machine translated file is converted back to detokenize form and is compared with the target file to give the BLEU score. The BLEU score is computed using sacrebleu library.

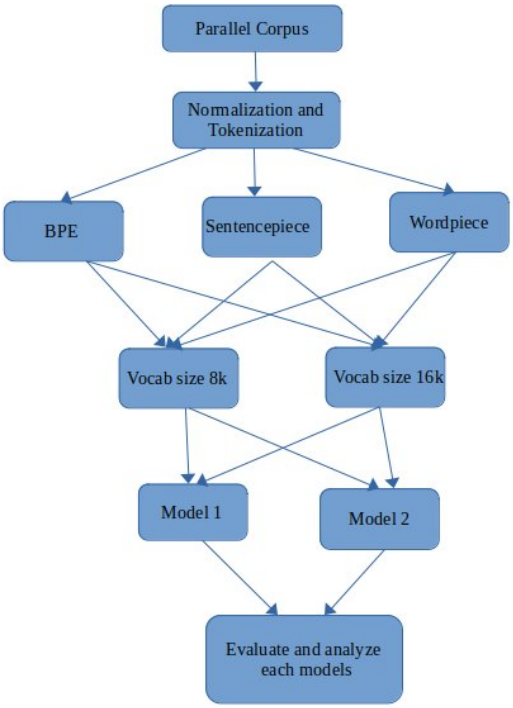

Figure 1: Flow steps of the whole process

4 Result and Discussion

The model is trained with our 92410 training data with 600 data extracted randomly from the whole dataset as validation data. The validation accuracy is monitored in each 10000 steps in all the training processes. The two models trained with the three different subword tokenization methods individually with both 8k vocabulary and 16k vocabulary sizes in both the directions gives us a total of 24 different BLEU score. The result of the training of English to Bodo is shown in Table 3 and the results of the training of Bodo to English is shown in Table 4.

Table 3: English to Bodo Translation

Subword Technique	**BLUE of Model 1**	**BLUE of Model 2**
BPE(8k vocab)	10.48	**11.01**
Wordpiece(8k vocab)	10.23	10.32
Sentencepiece(8k vocab)	10.99	10.70
BPE(16k vocab)	8.67	9.64
Wordpiece(16k vocab)	9.94	10.44
Sentencepiece(16k vocab)	9.39	10.33

Table 4: Bodo to English Translation

Subword Technique	**BLUE of Model 1**	**BLUE of Model 2**
BPE(8k vocab)	13.52	13.73
Wordpiece(8k vocab)	14.06	**14.62**
Sentencepiece(8k vocab)	12.79	13.58
BPE(16k vocab)	11.69	12.86
Wordpiece(16k vocab)	13.59	13.88
Sentencepiece(16k vocab)	13.37	14.01

From table 3 we get to known that the Model 2 trained with BPE subword data with a vocabulary size of 8000 vocabs has the highest BLEU score of 11.01 for the English to Bodo translation. And from table 4 we get to known that the Model 2 trained with Wordpiece subword data with a vocabulary size of 8000 vocabs has the highest BLEU score of 14.62 for the Bodo to English translation. From both table 3 and 4 it is observed that for all 23 cases Model 2 shows a higher BLEU score than Model 1 except for one. And comparing table 3 and 4 we observed that in each of the trained models the Bodo to English translation shows a higher BLEU score than English to Bodo translation.

5 Conclusion

In this paper we have performed experiments on Neural Machine Translation for English to Bodo and Bodo to English using transformer encoder decoder model. Here we have trained the model using

92410 parallel sentences which we have have collected from National Platform of Language Technology(NPLT), Data Management Unit(DMU), Mission Bhashini, Ministry of Electronics and Information Technology and our own in-house created parallel corpus We have performed preprocessing such normalization, tokenization and Subword tokenization. Three methods of subword tokenization have been explored(i.e, BPE, Sentencepiece and Wordpiece subword) trained individually with two different vocabulary sizes of 8000 and 16000 on both Model 1 and Model 2. For English to Bodo translation Model 2 trained with a vocabulary size of 8000 from BPE subword tokenized text data gives the highest Bleu score of 11.01 and for Bodo to English translation same Model 2 trained with a vocabulary size of 8000 but from Wordpiece subword tokenized text data gives the highest Bleu score of 14.62. Since from our expirements we have seen that BPE subword tokenization has given the best result for English-to-Bodo translation and Wordpiece subword tokenization has given the best result Bodo-to-English translation, this can be a potiential objective of further reasearch to identify the affect of different subword tokenization methods for different language pairs.We observed that Model 2 which has more number of layers, hidden sizes etc. than Model 1 gives a higher result among both the model. In most of our cases, training the models with 8000 vocabulary sizes gives higher BLEU score than training the same model with 16000 vocabulary sizes. Also in all the cases, the same model trained with same subword data with same vocabulary sizes in Bodo to English direction gives a higher BLEU score than English to Bodo direction.

Acknowledgement

The work is done as part of Project ISHAAN: Machine Translation Project (English-Assamese-Bodo) at Gauhati University, sponsored by Ministry of Electronics and Information TechnologyMeity, Govt. of India.

References

Mazida Akhtara Ahmed, Kishore Kashyap, and Shikhar Kumar Sarma. 2023a. Pre-processing and resource modelling for english-assamese nmt system. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–6. IEEE.

Mazida Akhtara Ahmed, Shikhar Kumar Sarma, and Kishore Kashyap. 2023b. Tokenization effect on neural machine translation: An experimental investigation for english-assamese. In *The 14th International Conference On Computing, Communication And Networking Technologies(ICCCNT)*. IEEE.

Kalyanee Kanchan Baruah, Pranjal Das, Abdul Hannan, and Shikhar Kr Sarma. 2014. Assamese-english bilingual machine translation. *arXiv preprint arXiv:1407.2019*.

Manash Pratim Bhuyan and Shikhar Kumar Sarma. 2018. Automatic formation, termination & correction of assamese word using predictive & syntactic nlp. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 544–548. IEEE.

Parvez Aziz Boruah, Shikhar Kumar Sarma, Kishore Kashyap, and Simanta Kalita. 2023. Performance evaluation of english to bodo neural machine translation system with varying model architecture and vocabulary size. In *The 14th International Conference On Computing, Communication And Networking Technologies(ICCCNT)*. IEEE.

Biswajit Brahma, Anup Barman, Shikhar Kr Sarma, and Bhatima Boro. 2012. Corpus building of literary lesser rich language-bodo: Insights and challenges. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 29–34.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for english-tamil. In *Proceedings of the third conference on machine translation: shared task papers*, pages 770–775.

Abdul Hannan, Shikhar Kr Sarma, and Zakir Hussain. 2019. Marie: a statistical approach to build a machine translation system for english assamese language pair. *International Journal of Computer Sciences and Engineering, Available: https://doi.org/10.26438/ijcse/v7i3*, 774779.

Saiful Islam and Bipul Syam Purkayastha. 2018. English to bodo machine transliteration system for statistical machine translation. *International Journal of Applied Engineering Research 13*, pages 7989–7997.

Simanta Kalita, Parvez Aziz Boruah, Kishore Kashyap, and Shikhar Kumar Sarma. 2023. Nmt for a low resource language bodo: Preprocessing and resource modelling. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–5. IEEE.

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Shikhar Kr Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Deka, and Anup Barman. 2012. A structured approach for building assamese corpus: insights, applications and challenges. In *Proceedings of the 10th workshop on Asian language resources*, pages 21–28.

Shikhar Kr Sarma, B Brahma, M Gogoi, and Mane Bala Ramchiary. 2010. A wordnet for bodo language: Structure and development. In *Global Wordnet Conference (GWC10), Mumbai, India*.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Felix Stahlberg. 2019. Neural machine translation: A review and survey. *arXiv preprint arXiv:1912.02047*.

Kuwali Talkukdar, Shikhar Kumar Sarma, and Kishore Kashyap. 2023. Influence of data quality and quantity on assamese-bodo neural machine translation. In *The 14th International Conference On Computing, Communication And Networking Technologies(ICCCNT)*. IEEE.

Kuwali Talukdar and Shikhar Kumar Sarma. 2023. Parts of speech taggers for indo aryan languages: A critical review of approaches and performances. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–6.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, pages 5–21.

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

Ajay Anand Verma and Pushpak Bhattacharyya. 2017. Literature survey: Neural machine translation. *CFILT, Indian Institute of Technology Bombay, India*.