Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair

Kuwali Talukdar, Farha Naznin, Shikhar Kumar Sarma, Kishore Kashyap, Mazida Akhtara Ahmed, Parvez Aziz Boruah
Department of Information Technology, Gauhati University, India
kuwalitalukdar@gmail.com, farha.gu@gmail.com, sks001@gmail.com,
kb.guwahati@gmail.com, 14mazida.ahmed@gmail.com, parvezaziz70@gmail.com

Abstract

Impressive results have been reported in various works related to low resource languages, using Neural Machine Translation (NMT), where size of parallel dataset is relatively low. This work presents the experiment of Machine Translation in the low resource Indian language pair Assamese-Bodo, with a relatively low amount of parallel data. Tokenization of raw data is done with IndicNLP tool. NMT model is trained with preprocessed dataset, and model performances have been observed with varying hyper parameters. Experiments have been completed with Vocab Size 8000 and 16000. Significant increase in BLEU score has been observed in doubling the Vocab size. Also data size increase has contributed to enhanced overall performances. BLEU scores have been recorded with training on a data set of 70000 parallel sentences, and the results are compared with another round of training with a data set enhanced with 11500 Wordnet parallel data. A gold standard test data set of 500 sentence size has been used for recording BLEU. First round reported an overall BLEU of 4.0, with vocab size of 8000. With same vocab size, and Wordnet enhanced dataset, BLEU score of 4.33 was recorded. Significant increase of BLEU score (6.94) has been observed with vocab size of 16000. Next round of experiment was done with additional 7000 new data, and filtering the entire dataset. New BLEU recorded was 9.68, with 16000 vocab size. Cross validation has also been designed and performed with an experiment with 8-fold data chunks prepared on 80K total dataset. Impressive BLEU scores of (Fold-1 through fold-8) 18.12, 16.28, 18.90, 19.25, 19.60, 18.43, 16.28, and 7.70 have been recorded. The 8th fold BLEU deviated from the trend, might be because of non-homogeneous last fold data.

1 Introduction

Assamese and Bodo both are official languages spoken in the North Eastern part of India. Approximate 16 million speakers use Assamese language, while around 3 million native speakers in the state of Assam, particularly in the Bodoland area uses Bodo language. Different natural language processing (NLP) works are reported for both the languages, including corpus creation, tagging, Wordnet development etc. But no attempt has been seen for Machine Translation in this language pair. A neural Machine Translation platform has been customized for Assamese-Bodo machine translation experiments. Preprocessing pipeline includes subword tokenization using Byte Pair Encoding (BPE). IndicNLP tool has been reported to work better for Indian language preprocessing, and hence for tokenization, we have adopted IndicNLP. A bilingual parallel corpus of 70000 parallel sentences (Assamese-Bodo) has been sourced from NLTM[1]. 11000 parallel sentences have been extracted from the Assamese and Bodo Wordnet data. (

Sahinur Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray and Sivaji Bandyopadhyay. 2020. *Multimodal Neural Machine Translation for English to Hindi*, Shared Task (Working Notes) – 7th Workshop on Asian Translation (WAT 2020), Hosted by the AACL-IJCNLP 2020

Sahinur Rahman Laskar and Partha Pakray. 2021. *Neural Machine Translation: Assamese–Bengali*. Modeling, Simulation and Optimization (pp.571-579)

Saiful Islam, Bipul Syam Purkayastha, 2018. *English to Bodo Machine Transliteration System for Statistical Machine Translation*, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, pp. 7989-7997 Number 10, 2018

Shikhar Kr Sarma et al., 2010) Also 5000 parallel sentences have been created in-house with the help of expert linguists.

Neural Machine Translation (NMT) has been widely used in contemporary Machine Translation works. NMT requires good quality and reasonable size dataset of parallel sentences to train the model. Here comes the challenge for low resource language like Bodo. Bodo is relatively new for NLP works, and not much works have been done. Although in recent years, few NLP research works are reported, but till now in Assamese-Bodo pair, Neural Machine Translation has never been attempted. The current work, hence, is a novel attempt in order to give a first even experimental modelling for Assamese-Bodo Neural Machine Translation, and is the baseline work.

https://nplt.in/

As NMT demands significant amount of parallel corpus at sentences level, and as there is a scarcity of such resources, in the current work, we have tried to integrate resources obtaining from different sources, including available open sources data, and generating in-house using linguistic experts. Details of series of performances under different training parameters are reported here.

2 Related Works

NMT has nowadays become and widely used and acceptable Machine Translation platform, as it outperformed the earlier rule based and SMT approaches. Open NMT contributes to the efficiency, modularity and extensibility (Guillaume Klein et al., 2017). Faster training and efficient test strategies, modular structure, and scopes for incorporating new and innovative research directions are important features of OpenNMT. This also gives an exhaustive library for training and deployment of neural machine translation. Other NMT platform includes JoeyNMT which is a PyTorch based toolkit particularly designed for novices (Julia Kreutzer et al., 2020). In a simple code base, Joey NMT provides many important NMT features, and it supports RNN and Transformer models. It is reported that NMT shows very impressive translation results in short sentences, but the performance decreases as the number of tokens in a sentence increases. (Dzmitry Bahdanau et al., 2014). We also observed this performance bottleneck in our experiments, but have not included, within our scope in this paper, analysing the results based on length of sentences.

NMT systems are also reported to offer better performance in many language pairs of broad domains, however for narrow domain results still SMT are reported to perform better (Lucia Benková and Ľubomír Benko, 2020). Despite various limitations, and varying performances over range of parameters and characteristics of resources, NMT is the contemporary dominant approach for machine translation tasks for both research and applications.

For low resource language pairs, NMT becomes challenging. There are many NMT approaches proposed and experimented for low resource language pairs. A novel approach with multihead self attention has been experimented for English-Tamil and English Malayam (Himanshu et al.,2020). Here they have collected corpus from multiple sources, addressed the issues related with the resources, and then integrated and refined. Comparatively better BLEU scores were reported as 24.34 and 9.78 for the pairs.

Another work reported NMT experimentations for English-Tamil, English-Hindi, and English- Punjabi machine translations, where they have evaluated the NMT performances using BLEU as well as human evaluators for assessing quality of translation in terms of fluency adequacy (

Sahinur et al. 2020)

Other NMT works for Indian language pairs include Hindi-Gujrati NMT system (Parth Shah, Vishvajit Bakrola, 2020) where they have presented comparative evaluation with multiple evaluation matrix- BLEU, perplexity, and TER. Here it is claimed that the system outperforms the Google translation performance with a margin of 6 BLEU score.

Although Assamese-Bodo pair has not been experimented by anyone, and the current work is the first ever attempt, but Assamese-Bengali, and Assamese-English pairs have been presented with NMT experimentations in few reported works. The NMT system reported at (

Sahinur Rahman Laskar and Partha, 2021) claimed a BLEU score of of 10.10 for Bengali-Assamese direction, and 7.20 for Assamese-Bengali direction.

As part of similar attempts involving Bodo machine translation, English-Bodo parallel text

corpus has been developed for training phrased base SMT system on English-Bodo pair (Saiful Islam, Bipul 2018). English-Bodo NMT has been attempted for limited resources of tourism domain parallel corpus. This BiLSTM based work reported using of attention mechanism, and observed tourism domain BLEU of 17.5.

3 Data and Preprocessing

3.1 Data:

For our NMT model, the base dataset is the 70000 Assamese-Bodo parallel sentences. These sentences are of varying lengths in terms of number of tokens. We analysed the dataset based on the number of tokens in each sentence, and the statistical presentation is given in Table 1.

Length	No of Sentences	Percentage of Total Dataset
Upto 10	28560	40.80%
>10<=20	27467	39.24%
>20<=30	12340	17.63%
>30	1633	02.33%
	70000	100.00%

Table 1: Sentence distribution based on number of tokens (70000 base dataset)

The new chunk of data is the 11500 Assamese-Bodo parallel sentences extracted from Wordnet database. North East Indian Language Wordnet is part of the Indo Wordnet, and consists of Wordnets of Assamese, Bodo, Manipuri, and Nepali languages. We extracted the 11500 parallel sentences for Assamese-Bodo pair, which are considered as gold standard dataset, because these were created by expert linguists. As part of dataset collection and creation efforts, 7000 gold standard Assamese-Bodo sentences were created by expert linguists. Standard test data of size 500 sentences was also created inhouse for testing the model, and to record performance of the model through BLEU scores.

Source	No of Parallel sentences	Training/ Testing
NLTK	70000	Training Dataset
Wordnet	11500	
Inhouse Creation	7000	
Inhouse Creation	500	Testing Dataset

Table 2: Summary of Datasets

Domain	No of Sentences	Percent distribution
Administration	79	15.80%
Agriculture	53	10.60%
Education	92	18.40%
Healthcare	54	10.80%
Law	124	24.80%
Science and Climate	16	3.20%
Tourism	82	16.40%
	500	100%

Table 3: Domain wise distribution of Test Dataset sentences

3.2 Data Preprocessing

Indian language specific NLP tools have been designed in order to boost Indian language NLP research works. One important work is the IndicNLPSuite, (Divyanshu et al.,2020) which is used in our current work for tokenization and detokenization phases. They have reported that this tool is suitable for handling morphological complexity of Indian languages, and hence we have adopted IndicNLP for preprocessing stages. Tokenization is also performed by tools like Moses, another open source toolkit for machine translation, that includes a collection of tools for training, tuning machine translation system. Then the dataset is subjected to subword tokenization in order to give adequate scope for reducing vocabulary size, and to handle out of vocabulary. The dataset has also been passed through filtering facilitating cleaning dataset removing blank lines and duplicities. For Data filtering we have developed a set of codes in Python. The preprocessed cleaned dataset is not fed to the next layer in the NMT model for training.

4 NMT Model Experiments

We trained the model initially with a dataset of 30000 parallel sentences. This was not to observe and record results and system performances, rather we used this phase for having a working test of the customized installed pipeline and model. Then a series of experiments has been designed and performed. We have adopted evolutionary model, so that as we travel through dynamics of the data and parameters, system continuously evolves and performs better and better. In fact, from our experimental results, it has been evident, the details of which is elaborated in the results and analysis section. We used OpenNMT for our NMT

experiments. Our model building is on the pytorch version of OpenNMT and based on encoder decoder architecture. Both the encoder and decoder parts consists of 3 layers of transformer blocks.

The first set of experiment is with 70000 parallel sentences dataset. The model is trained with 8000 vocab size, and an overall BLEU of 4.0 was recorded. Domain specific BLEU scores have been recorded, and presented in the next section. With same Vocab size, we now enhanced the dataset with the Wordnet data of 11500. Experiment has been performed in the same pipeline and with the same model, and a slight increase of over BULE score of 4.33 has been recorded.

Then the next level of experiment was performed with doubling the Vocab size to 16000. Experiment in the same pipeline and with the same model has been performed on the Wordnet enhanced dataset, to record significant increase in BLEU score to 6.94. Then we moved to the next round of experiment. This time the dataset is enhanced with gold standard additional dataset of 7000, created inhouse with expert linguists. Data filtering using Python based code set was operated on the integrated dataset, and model was training with 16000 vocab size. New BLEU recorded was 9.68.

As part of the experiments, we also designed cross validation of the system. Cross validation has been designed and performed with an experiment with 8-fold data chunks prepared on 80K total dataset. Impressive average BLEU scores of around 17 has been observed, which signifies model performance as an acceptable machine translation model for practice.

5 Results and analysis

Testing was done with the gold standard 500 sentences test dataset. Testing has been performed domain wise, as well as for overall dataset. BLEU scores have been recorded under varying vocab size as well as evolving dataset.

Domain	8000 Vocab size	
	BLEU against training with 70000 dataset	BLEU with Wordnet enhanced dataset (70000+11500)
Administration	7.15	4.86
Agriculture	5.6	6.83
Education	6.31	9.82
Healthcare	2.8	4.11
Law	1.13	2.84
Science and Climate	2.6	8.17
Tourism	2.9	7.11
Overall	4.0	4.33

Table 4: BLEU scores with 8000 Vocab size

Although overall BLEU scores were not impressive, however reasonable performance quality was observed in domain specific translation, particularly in domains like administration, education and tourism.

Domain	16000 Vocab size	
	BLEU with Wordnet enhanced dataset (70000+11500)	BLEU with enhanced dataset (70000+Wordnet+7000 inhouse created) subjected to data filtering
Admin.	7.0	11.60
Agriculture	6.39	10.36
Education	9.58	14.13
Healthcare	7.25	9.06
Law	2.09	3.79
Science Climate	6.14	5.04
Tourism	8.27	13.47
Overall	6.94	9.68

Table 5: BLEU scores with 16000 Vocab size and enhanced, filtered dataset

We also performed cross validation as part of the whole experiment. Cross validation designed with 80000 dataset, creating 8 fold data chunks has resulted impressive BLEU scores.

Data Fold	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Fold-6	Fold-7	Fold-8
BLEU	18.12	16.28	18.90	19.25	19.60	18.43	16.28	7.70

Table 6: Cross validation with 8 fold data chunking on 80000 dataset

We consider that the final set of BLEU score as presented in Table 5 signifies our baseline performance for the Assamese-Bodo NMT. It is evident from the series of experiments that data amount, as well as data preprocessing for quality parallel resources are crucial for training a

machine translation model. The domain specific BLEU scores presented above reflects the performance of the system adaptive wo a wide variety of testing data, although scopes could be clearly observed for further enhancement of model performance with adequate quality dataset, particularly domain specific dataset, thereby facilitating effective learning of range of vocabularies alongwith syntactic characteristic mapping from source to target language.

Limitations

Machine learning is always data critical, and the major limitation here is the parallel resources for the low resource language pair Assamese-Bodo. Both the languages are relative new for NLP, and very few published dataset are available for NLP tasks. Although monolingual data are being available resulting out of discrete research works in Assamese an Bodo mostly at academic level, parallel dataset, that too, of good quality, and reasonable in size, is very rare.

References

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

Biswajit Brahma, Anup Kr. 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, Mumbai, India. The COLING 2012 Organizing Committee.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. 2016. *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv:1409.0473

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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

Himanshu Choudhary, Shivansh Rao, and Rajesh Rohilla. 2020. Neural Machine Translation for Low-Resourced Indian Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3610–3615, Marseille, France. European Language Resources Association.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Julia Kreutzer, Jasmijn Bastings, Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices. arXiv:1907.12484v3

Lucia Benková and Ľubomír Benko. 2020. *Neural Machine Translation as a Novel Approach to Machine Translation.* DIVAI 2020 The 13th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, Slovakia

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Parth Shah, Vishvajit Bakrola. 2020. *Neural Machine Translation System of Indic Languages -- An Attention based Approach.* arXiv:2002.02758v1

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Sahinur Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray and Sivaji Bandyopadhyay. 2020. *Multimodal Neural Machine Translation for English to Hindi***,** Shared Task (Working Notes) – 7th Workshop on Asian Translation (WAT 2020), Hosted by the AACL-IJCNLP 2020

Sahinur Rahman Laskar and Partha Pakray. 2021. *Neural Machine Translation: Assamese–Bengali*. Modeling, Simulation and Optimization (pp.571-579)

Saiful Islam, Bipul Syam Purkayastha, 2018. *English to Bodo Machine Transliteration System for Statistical Machine Translation*, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, pp. 7989-7997 Number 10, 2018

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