Iterative Back Translation Revisited: An Experimental Investigation for Low-resource English Assamese Neural Machine Translation

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

Abstract

Back Translation has been an effective strategy to leverage monolingual data both on the source and target sides. Research have opened up several ways to improvise the procedure, one among them is iterative back translation where the monolingual data is repeatedly translated and used for re-training for the model enhancement. Despite its success, iterative back translation remains relatively unexplored in low-resource scenarios, particularly for rich Indic languages. This paper presents a comprehensive investigation into the application of iterative back translation to the low-resource English-Assamese language pair. A simplified version of iterative back translation is presented. This study explores various critical aspects associated with back translation, including the balance between original and synthetic data and the refinement of the target (backward) model through cleaner data retraining. The experimental results demonstrate significant improvements in translation quality. Specifically, the simplistic approach to iterative back translation yields a noteworthy +6.38 BLEU score improvement for the English-Assamese translation direction and a +4.38 BLEU score improvement for the Assamese-English translation direction. Further enhancements are further noticed when incorporating higher-quality, cleaner data for model retraining highlighting the potential of iterative back translation as a valuable tool for enhancing low-resource neural machine translation (NMT).

1 Introduction

Low Resource Neural Machine Translation (NMT) is an open problem where even the most successful state-of-the-art methods or models have a limited impact and adapting NMT systems under these restrictions is still a challenging task. The basic requirement for any popular neural model is an enormous amount of data which becomes an inherent limitation for low-resource languages where only tens of thousands of parallel data is available. Although there is a dearth of parallel data, monolingual data is abundantly available for most of the cases. A well-known, simple and effective method is Back-Translation (BT) where a target to source NMT model is trained on the parallel data and is used to translate the monolingual target data to generate a relatively large pseudo-parallel dataset. The pseudo-parallel dataset can thus be utilised to appease the data hungry NMT models which appears to be a suitable method for low resource scenarios. (Sennrich et al., 2015a; Jain et al., 2021) have demonstrated successful implementation of the BT approach. Simple enough as it seems to be, several factors such as the data quality (Hoang et al., 2018), ratio of original to synthetic data size (Poncelas et al., 2018; Hoang et al., 2018), model tuning on clean data, impact of merging the original parallel data to the synthetic parallel data (Sennrich et al., 2016), noise induction (Wu et al., 2019) etc influence a BT model. We investigate back translation for low-resource language pair: English-Assamese. As pointed out by (Poncelas et al., 2018), model performance is negatively affected by the back translated data which is error-prone in comparison to the *'perfect'* human translations. To mitigate this effect of BT and in order to enhance the translation quality we adopt BT in an iterative manner, inspired by (Hoang et al., 2018), hypothyzing to produce better translations with every iteration.

Contributions of this paper:

1. Experimental analysis on Byte Pair Encoding (BPE) and SentencePiece (SP) in order to find the suitable vocabulary size for the data at hand.

2. Simplified version of the standard Iterative Back Translation approach to leverage the use of existing target monolingual data for better

translation quality and model performance for both directions.

3. Experimental analysis on balancing original and synthetic data as well as model tuning on semantically cleaner data to enhance backward model.

The rest of the paper is organized as follows: Section 2 describes some available reported works on BT, Section 3 presents the methodology followed, experiments done and results obtained and Section 4 concludes the paper.

2 Related Work

The whole idea of back translation in low resource scenario is to augment the limited data. Therefore, extensive research has been conducted in this direction to achieve improved results by investigating in the directions described in the following subsection.

2.1 Back Translation

1. *Data Augmentation:* Apart from using only target monolingual data, source side data is also exploited to inflate the data size (Tonja et al., 2023)(Abdulmumin et al., 2021)(Jain et al., 2021). In a similar work, (Nguyen et al., 2020) trained k forward and backward models in the rth round. The forward models are used to produce k translation sets of the source side data while the backward models generates k translation sets of the target side data. All the translation thus produced are merged to obtain a much larger sized data. This method has the advantage to exponentially increase the data size with no requirement of extra monolingual data but at the same time, the augmented data are more or less duplicated at least at source or target side which might hinder in attaining a generalised model.

2. *Noise induction:* (Wu et al., 2019) hits two targets at the same time namely:

   (a) *Data size expansion:* Exploits both source and target monolingual data producing forward and back translations to add up in the original parallel bitext.

   (b) *Model Generalization:* Noise is induced in the source sentences by a systematic analytical approach to produce a robust translation model. Also, tuning on clean data is observed to produce improved results [Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data].

3. *Data Quality:* This aspect has been investigated by multiple researchers such as (Hoang et al., 2018; Poncelas et al., 2018) where it has been shown that the quality of translations does produced influence the learning of a model as normally the machine generated data, especially the ones trained with limited data, have a substantial rate of error. And hence, enhancing the translation quality definitely boosts up the model performance and vice-versa (Akella et al., 2020).

4. *Ratio of Synthetic to original parallel text:* As (Sennrich et al., 2016) had experimentally shown that merging the original parallel text with the back translations produces better models, an important aspect in this data fusion is also the proportion of the errorneous back translations as compared to the clean parallel text. In this regard, (Poncelas et al., 2018) has presented a detailed analysis.

The techniques of back translation which holds good for a particular setting may not work the same way for another as a lot of factors come to play as described above. (Hoang et al., 2018) have reported that dual learning (He et al., 2016) which is a refined version of BT, effective in many cases, have failed for them. The language specialties differ from language to language demanding special care in each case, not to speak of the morphologically rich languages especially the Indic ones. On exploring the available literatures (Laskar et al., 2021; Kandimalla et al., 2022; Talukdar and Sarma, 2023; Kalita et al., 2023) for the underexplored North-Eastern Indic languages , we found limited works on back translation on Assamese, a North-Eastern Indic language, recognized as a low resource language with limited linguistic resource. BT, therefore, seems to be a suitable data augmentation technique. This paper is therefore, dedicated towards a systematic analysis of BT on the English-Assamese pair.

2.2 Iterative Back Translation

The core idea behind is that the insatiable neural models face a significant challenge in low-resource settings due to the scarcity of large-scale

data. Consequently, the resulting translations in such scenarios in either direction often fall short of expectations. Back translating target monolingual data to generate a synthetic parallel corpus is expected to boost up the target→source model which, in turn, can be used to reproduce better translations for the same monolingual data. The objective is to enhance the target→source model anticipating quality translations. The process can be repeated until saturation.

3 Methododology, Experiments and Results

All experiments are carried out using the Pytorch version of the open-source NMT toolkit known as OpenNMT-py[1] on a system equipped with a 256GB SSD and an Intel® Xeon(R) Gold 6230R processor, accompanied by an NVIDIA RTX A6000 with 48 GB of GPU RAM. Due to the limited availability of a single extensive data source, the genuine parallel training data comprises a combination of four distinct datasets: National Platform for Language Technology[2] (NPLT), Samanantar(Ramesh et al., 2022), PMIndia (PMI) (Haddow and Kirefu, 2020), and an in-house parallel data collection known as Machine Aided Translation (MAT). The individual contribution of each data source is detailed in table 1.

Table 1: Training Set Composition

DataSet	Size
NPLT	70,000
Samanantar	1,41,227
PMI	9,780
MAT	15,157
Total	**2,36,164**

Monolingual Source: We use the IndicCorp (Kakwani et al., 2020) Assamese monolingual data. It has 1.39M sentences sourced mainly from news articles.

To test the performance of the models developed, a comprehensive test set is designed to include complexity in all levels, be it domain-wise or of various sentence lengths. It is developed by linguistic experts comprising of seven (7) domains which includes Administration, Agriculture, Education, Judiciary, Health, Tourism and Technology and Climate (combined). Every domain also holds sentences with four different sentence length buckets namely, (1-10), (11-20), (21-30) and 30+.

3.1 Data Filtering, Pre-processing and Vocabulary Size Selection

As the training data is obtained from public sources, it is observed that it contains erroneous data which is probably scraped. To obtain a clean set of parallel training sentences, a basic filtering routine is applied which comprises of the following steps:

1. Removal of blank lines.
2. Removal of duplicate (source, target) pairs.
3. Removal of disproportionate (source, target) pairs where the source sentence is significantly (twice) greater than the target sentence and vice-versa.
4. Removal of (source, target) pair if either of them is too long or too short. Typically we keep sentences with the condition 5 < Sentence Length < 35.

Normalization: The filtered data is then normalized to maintain uniformity in the content. All English text is lowercased and for Assamese, the normalization procedure of IndicNLP (Kakwani et al., 2020) is used which does normalize ambiguous characters like '।' (dari, which is similar to pipe '|'), 'ঃ' (bisarga, having resemblance with colon ':') but fails to distinguish the oordhocoma (') from the punctuation markers (Ahmed et al., 2023). This character is normalized to punctuation which is tokenized where the words with oordhocoma are split during tokenization. A work-around as suggested in (Ahmed et al., 2023) is used to handle the fault.

Tokenization: After experimenting with several tokenizers like nltk2, Moses (Koehn et al., 2007), OpenNMT tokenizer and the tokenization module of IndicNLP, it is found that the best NMT results are obtained with Moses for English and IndicNLP for Assamese text.

Selection of appropriate subwording method: To find the most suitable subwording technique that fits to the data, two such methods are experimented with namely:

[1]https://github.com/OpenNMT/OpenNMT-py
[2]https://nplt.in/demo/

1. Byte Pair Encoding (BPE)(Sennrich et al., 2015b) : Seperate and independent vocabularies for both source and target are used.

2. SentencePiece (SP) (Kudo and Richardson, 2018): The *unigram* method is used with a character coverage of 0.995 on separate and independent vocabularies for both source and target without any subword regularization.

Table 2: BLEU scores obtained with variable Vocabulary sizes.

	SP		BPE	
Vocab Size	En-As	As-En	En-As	As-En
4K	8.62	14.27	7.00	12.45
8K	8.35	**14.75**	**8.30**	**13.84**
16K	8.38	13.90	7.90	12.94
32K	7.56	12.98	6.74	12.19

From the table 2, it is clearly seen that SP achieves better scores than BPE in both directions. Four vocabulary sizes have been tested and as rightly pointed out in (Gowda and May, 2020), smaller vocabulary size works for low-resource languages while the models tend to degrade towards larger vocabulary size. From the experimental observations, we therefore select the vocabulary size of 8000 for all experiments in this paper.

The model architecture adopted is the Transformer model (Vaswani et al., 2017). We experimented with various sets of parameters, including encoder and decoder layers, attention heads, embedding size, and the number of nodes in the feed-forward layer. Through these experiments, the optimal configuration is identified which includes using 3 encoder and 3 decoder layers, a word vector size of 512, and 2048 nodes in the feed-forward layer. Model training is executed with Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of 2 and Noam decay and 8,000 warm-up steps. The training continued for a total of 600,000 steps, with validation checks performed every 10,000 steps with early stopping having a patience of 4, based on validation perplexity and accuracy criteria.

3.2 Simplified Iterative Back Translation

This paper, describes extensive experiments performed on the simple and easy-to-implement iterative back translation method to investigate how it tunes on the low-resource English-Assamese pair. In the first stage, the standard back translation is performed. In this procedure, a base target→source model is developed on the limited authentic parallel data which is used to translate the available target monolingual data creating a relatively voluminous synthetic data. Iterative back translation which was originally coined by (Hoang et al., 2018), makes use of both source and target monolingual sources turn by turn, discarding the translations from the stage 1 (source translations). We, in our simplified version, make use of only target monolingual data, re-using the full set in every iteration. Also, our method is a simple single stage procedure (see Fig. 1).

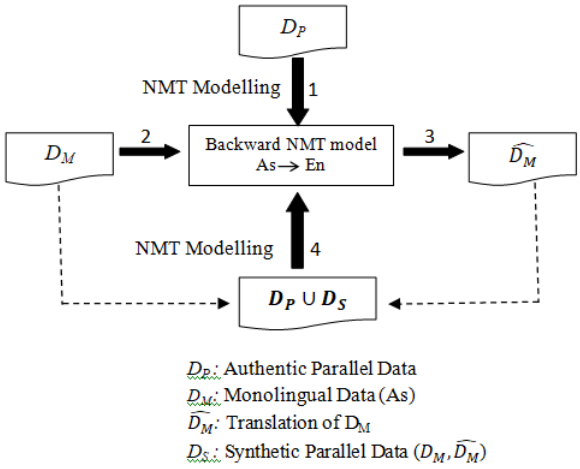

Figure 1: Simple Iterative Back-translation approach. *Steps 2, 3,4 can be repeated to the desired times.*

In the next iteration of re-training, two options are available:

1. using only the synthetic data for re-training and fine-tuning on the original data or

2. merging the original and synthetic parallel data.

We tested the first method but found that fine-tuning reduces the score on BLEU. This could be provoked by the parallel data if it is of pitiable condition. Samanantar, which occupies a major share of our original parallel data, is seen to lag behind in quality which we believe could be the cause for this failure. We, therefore, select the second strategy as suggested by (Sennrich et al., 2016) by re-training on the combination of 0.2M original and

1M synthetic parallel data. Three rounds of iterations are performed and the results seem promising with performance enhancing in every iteration (Table 3).

Table 3: BLEU sores of the Forward and Backward Mode with Iterative Back-translation.

Phase	En → As	As → En
Baseline	9.14	14.95
Iteration-1	14.39	17.36
Iteration-2	14.66	**19.33**
Iteration-3	**15.52**	18.64

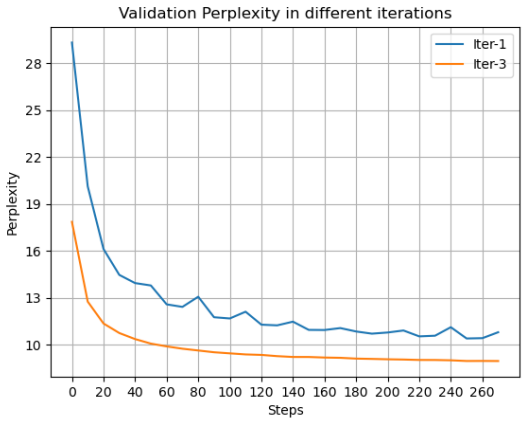

Figure 2: Perplexity of the Validation set during training.

In Table 3, baseline is the score obtained with the model trained on the original parallel data. The backward (As→En) base model is used to translate the Assamese monolingual data to produce the first set of synthetic parallel data. The synthetic and original parallel data are merged and the model for Iteration-1 is trained on it. Instead of training it from scratch, the model is initialised with the parameter values of the best checkpoint obtained for the baseline model. On testing, a significant increase in BLEU score is noticed for Iteration-1 in both directions as the augmented data is able to satisfy the model to a fair extent. The forward model score is increased by +5.25 and the backward model by +2.41. The Iteration-1 backward model is then again used to re-translate the monolingual target data and the same procedure of data merging and model retraining is followed. The backward model for Iteration-2 is highly benefitted with an increase by +1.97 BLEU while a slight

Table 4: Domain-wise BLEU scores for En→As.

Domains	Base Model	Iteration-3
Agriculture	8.67	10.56
Education	9.89	11.78
Law	5.70	8.17
Administration	11.51	14.82
Technology and Climate	7.96	9.19
Health	10.65	9.55
Tour	9.13	12.50
Overall	9.14	15.52

increase is noticed for the forward model. Same trend is seen for Iteration-3 in the forward direction. The domain-wise score break-up is shown in Table 4. The domains highly benefited are Law, Administration and Tourism. This trend can be attributed to the fact that the monolingual IndicCorp data are mined from news articles which are rich in these domains which might be the probable cause for the significant improvement in those domains. Also, Figure 2 shows the perplexity of the validation set during training in the first and third iterations. The graph shows that, as the iteration progresses, the model converges faster indicating the impact of data quality on model training.

Also, shown in Table 5 are some test set sample translations by the base model(trained without monolingual data), iteration-1(Iter-1) model and iteration-3 (Iter-3) model (trained with monolingual data). The first set includes translations for short sentence, the second for mid-length sentence and the third set is for longer sentences with Named Entities. For a quantitative analysis, we have included three kinds of quality analysis metrics:

1. *LaBSE:* The LaBSE (Feng et al., 2020) score indicates the cosine similarity between the source and target sentence embeddings by the LaBSE model. Higher the value, closer is the source embedding to the target embedding.

2. *Adequacy:* It indicates whether the translation produced is semantically adequate i.e., captures the meaning of the source sentence.

3. *Fluency:* It shows if the translation is aligned to the common grammatical usage by a native speaker of the language.

Table 5: *Translation Samples:* Base Vs Iteration-1 Vs Iteration-3 Models.

Model	Translation	LaBSE score	Adequacy	Fluency
En	Our Assam is an agricultural state.			
Base	আমাৰ অসম কৃষি ৰাজ্য।	0.918	3	4
It-1	আমাৰ অসম এখন কৃষি ৰাজ্য।	0.921	4	5
It-3	আমাৰ অসম এখন কৃষিপ্ৰধান ৰাজ্য।	**0.942**	5	5
En	Rice cultivation was introduced in Assam about three thousand years ago.			
Base	অসমত প্ৰায় তিনি হাজাৰ বছৰ পূৰ্বে চাউলৰ খেতি কৰা হৈছিল।	0.866	2	5
It-1	অসমত প্ৰায় তিনি হাজাৰ বছৰ পূৰ্বে চাউলৰ খেতিৰ সূচনা হৈছিল।	0.895	4	5
It-3	প্ৰায় তিনি হাজাৰ বছৰ পূৰ্বে অসমত ধান খেতিৰ প্ৰচলন হৈছিল।	**0.904**	5	5
En	In 1836, American Baptist missionaries led by Dr. Nathan Brown and Oliver T. Carter arrived in Assam and established their centers first at Shadia and later at a place name Jaipur.			
Base	১৮৩৬ চনত ডা. নাথন বাদামী আৰু জলফাইৰ টি কেটাৰৰ নেতৃত্বত আমেৰিকা বাপ্তিষ্ট মিছনেৰী ডা. নাথান বাদামী আৰু অলিভ টি কেৰেটে অসমৰ প্ৰথম স্থানত নিজৰ কেন্দ্ৰ স্থাপন কৰে আৰু পাছত জয়পুৰৰ নাম দিয়ে।	0.619	1	3
It-1	১৮৩৬ চনত ডan নাথান ব্ৰাউন আৰু অলিভাৰ টি কেৰ নেতৃত্বত আমেৰিকান বাপ্তিষ্ট মিছনেৰীসকলে অসমত উপস্থিত হয় আৰু পিছলৈ জয়পুৰ নামৰ ঠাইত নিজৰ কেন্দ্ৰ স্থাপন কৰে।	0.715	2	4
It-3	১৮৩৬ চনত ড০ নাথান ব্ৰাউন আৰু অলিভৰ টি কাৰ্টাৰৰ নেতৃত্বত আমেৰিকানসকলে অসমত উপস্থিত হয় আৰু ইয়াৰ কেন্দ্ৰ প্ৰথমে ছাদিয়াত আৰু পাছত জয়পুৰ নামেৰে স্থাপন কৰে।	**0.716**	2	4

Both Adequacy and Fluency are ranked in the scale from 1-5. Higher score indicates better quality. Deeper explorations are conducted in line to this experiment to find the effect of original and synthetic data balancing as well as backward model enhancement by cleaner data as discussed in the following experiments.

3.3 Effect of Quantity and Quality of Synthetic Data

A common apprehension regarding back translation is that the model might be negatively affected by an unbalanced proportion of original and synthetic data as the translations are prone to (serious) errors. To investigate in this direction, the synthetic data is progressively incremented in every iteration instead of using the full set at once. Starting with equal sized original and synthetic data (1:1) in iteration-1 and doubling the synthetic data in the second iteration, table 6 describes the results obtained. No clear trend is seen in both directions on comparing with the scores in Table 3.

Again, (Poncelas et al., 2018) has demonstrated that the quality of the synthetic data affects the model training. They have used intermediate translations generated by partially convergent models to report the results. We use translations produced by fully converged models which are then filtered and their quality estimated. For filtration we use the filtration procedure as described in data processing. For quality estimation of the translations the LaBSE module is utilized. The translations below the selected threshold value (here, 0.7) are dropped. Target→source models are re-trained on these relatively better translations which enhances the target model which in turn, is expected to reproduce better translations. Table 6 reports the results obtained in various iterations. A positive trend is seen from the results indicating the success of our hypothesis.

4 Conclusion

From the experiments conducted we find that Iterative Back translation stands out as an effective approach, particularly for low-resource NMT scenarios. Secondly, it is evident that the utilization of higher-quality synthetic data significantly enhances model performance in both directions, underscoring the critical importance of data quality in NMT training. And finally, our analysis reveals no distinct, consistent trend on the perfor-

Table 6: BLEU scores for Iterative BT with Original:Synthetic data balancing and LaBSE cleaned synthetic data

	En → As		**As → En**	
Phase	Orig:Syn	LaBSE	Orig:Syn	LaBSE
It-1	13.91 (1:1)	15.59	19.24 (1:1)	17.94
It-2	14.23 (1:2)	**15.61**	18.36 (1:2)	**20.25**

mance of Back translation models concerning the proportional inclusion of synthetic data alongside the original parallel dataset. However, it is important to acknowledge that these trends cannot be generalized as every language has their own peculiarities which makes them distinct from one another. We plan to further investigate these directions in a bunch of low-resource language pairs to arrive at a general conclusion.

References

Idris Abdulmumin, Bashir Shehu Galadanci, and Abubakar Isa. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers 3*, pages 355–371. Springer.

Mazida Akhtara Ahmed, Kishore Kashyap, and Shikhar Kumar Sarma. 2023. 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.

Kartheek Akella, Sai Himal Allu, Sridhar Suresh Ragupathi, Aman Singhal, Zeeshan Khan, Vinay P Namboodiri, and CV Jawahar. 2020. Exploring pairwise nmt for indian languages. *arXiv preprint arXiv:2012.05786*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. *arXiv preprint arXiv:2004.02334*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

Aditya Jain, Shivam Mhaskar, and Pushpak Bhattacharyya. 2021. Evaluating the performance of back-translation for low resource english-marathi language pair: Cfilt-iitbombay@ loresmt 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 158–162.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, 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.

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.

Akshara Kandimalla, Pintu Lohar, Souvik Kumar Maji, and Andy Way. 2022. Improving english-to-indian language neural machine translation systems. *Information*, 13(5):245.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

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*.

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.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

Alberto Poncelas, Dimitar Shterionov, Andy Way, G Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

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. IEEE.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

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.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216.