GUIT-NLP's submission to Shared Task: Low Resource Indic Language Translation

Mazida Akhtara Ahmed, Kuwali Talukdar, Parvez Aziz Boruah
Shikhar Kumar Sarma and **Kishore Kashyap**
Dept. of Information Technology, Gauhati University
Guwahati, Assam, India
14mazida.ahmed@gmail.com, kuwalitalukdar@gmail.com, parvezaziz70@gmail.com,
sks001@gmail.com, kb.guwahati@gmail.com

Abstract

This paper describes the submission of the GUIT-NLP team in the "Shared Task: Low Resource Indic Language Translation" focusing on three low-resource language pairs: English-Mizo, English-Khasi, and English-Assamese. The initial phase involves an in-depth exploration of Neural Machine Translation (NMT) techniques tailored to the available data. Within this investigation, various Subword Tokenization approaches, model configurations (exploring differnt hyper-parameters etc.) of the general NMT pipeline are tested to identify the most effective method. Subsequently, we address the challenge of low-resource languages by leveraging monolingual data through an innovative and systematic application of the Back Translation technique for English-Mizo. During model training, the monolingual data is progressively integrated into the original bilingual dataset, with each iteration yielding higher-quality back translations. This iterative approach significantly enhances the model's performance, resulting in a notable increase of +3.65 in BLEU scores. Further improvements of +5.59 are achieved through fine-tuning using authentic parallel data.

1 Introduction

Work on Machine Translation (MT) involving indigenous languages is on the rise to provide such languages a global existence rather than limiting its scope to regional geographical boundaries. But such a work is quite challenging owing to its typical characteristic being limited (low) resourced as NMT models are data hungry which tend to degrade with limited data input. Established methods like Back Translation (Sennrich et al., 2015b), Transfer Learning (Kim et al., 2019; Zoph et al., 2016), Multilingual Neural Translation (MNT) (Lakew et al., 2018; Ngo et al., 2020), Dual Learning (He et al., 2016; Wang et al., 2018) and such do exist to tackle the low-resource challenge. With the monolingual and limited parallel data provided to the teams to work with, Back Translation (BT) seemed to be an appropriate choice. In BT, a target to source model translates the target side monolingual data to generate a substantial amount of synthetic parallel data which could be augmented with the limited authentic parallel data to increase the volume of training data. Previous experiments (Sennrich et al., 2015a; Edunov et al., 2018), (Poncelas et al., 2018; Wu et al., 2019) have shown improved results in such scenarios.

The general NMT pipeline comprises of various stages like tokenization, subword tokenization, NMT model training, inference and post-editing. It should be noted that several methods are available for every stage making it difficult for the researcher to select the one that would suit the data best as each method has its own influence on the model performance. We, therefore, perform an initial investigation on two popular subword tokenization methods to find the best choice. The rest of the paper is organized as follows: Section 2 describes the methods applied for the task, Section 3 presents the experimental setup and the results obtained for the three language pairs: English-Mizo, English-Khasi and English-Assamese. Section 4 concludes the paper.

2 Methodology

The following section describes the methodology used for the task for each of the language pairs.

2.1 Data Exploration

In this section, we delve into the data used for the task, which encompasses two primary categories:

1. *Parallel Data:* This data category consists of two distinct, non-overlapping sets, specifically the training and validation set.

2. *Monolingual Data:* This category encompasses an extensive corpus with monolingual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 935–940
December 6–7, 2023.

Language Pair	Train Set	Dev. Set
English-Mizo	50,000	1,500
English-Khasi	24,000	1,000
English-Assamese	50,000	2,000

Table 1: Parallel Data Statistics.

Language	Sentences*(in millions)*
Mizo	1.9
Khasi	0.18
Assamese	2.6

Table 2: Monolingual Data Statistics.

sentences. It is imperative to underscore that all participating teams are expressly instructed to rely exclusively on the provided data, refraining from any utilization of external resources.

Upon conducting a preliminary manual analysis of the data, several noteworthy observations have come to light:

(i) Instances exist within the corpus wherein sentences commence with multiple spaces.

(ii) Instances within the corpus also manifest where multiple spaces occur between words.

(iii) The corpus exhibits a mixture of both tokenized and untokenized sentences.

After having these disparities removed from the data, the sets are tokenized with Moses (Koehn et al., 2007) tokenizer for English, Mizo and Khasi as they share the same Roman script while Assamese is tokenized with IndicNLP[1]. Prior to tokenization of the English, Mizo and Khasi text, all characters are normalized to lowercase for consistency. With no difference in case for Assamese, this step is not required for the language. Additionally, a fundamental filtering routine is applied as part of the data preprocessing process as described below:

(a) *Removal of Empty Lines:* (Source, target) pairs containing empty lines on either the source or target side are systematically eliminated from the dataset.

(b) *Elimination of Duplicate Lines:* (Source, target) pairs characterized by duplicate lines in both the source and target segments are systematically removed. Duplicate content can introduce redundancy and skew the training process, hence necessitating their exclusion.

(c) *Relative Length-Based Filtering:* To maintain a balanced and coherent dataset, pairs where the length of the target sentence significantly exceeds that of the source sentence (or vice versa), exceeding a predetermined threshold (typically set at twice the length), are judiciously omitted.

2.2 Subword Tokenization

In the context of developing NMT models for low-resource Indian languages, subword tokenization emerges as a critical technique as it addresses out-of-vocabulary (OOV) challenge, morphological richness, facilitates cross-lingual transfer of knowledge, reduces the vocabulary size substantially. Two popular schemes are explored namely:

1. *Byte Pair Encoding (BPE):* BPE (Sennrich et al., 2015c) is a data compression technique designed to systematically merge the most common pair of character sequences. Consequently, frequent substrings are unified into single symbols, while rare words are segmented into smaller constituents. BPE is experimented in two forms:

   (a) Independent Vocabulary: This involves creating separate and independent subword vocabularies for both the source and target languages.

   (b) Shared Vocabulary: When dealing with closely related languages a shared subword vocabulary is a popular choice as it aligns (sub)words from source and target sentences into the same embedding space so as to strengthen the semantic correlation between them.

2. *Sentencepiece* (Kudo and Richardson, 2018): Though Sentencepiece (SP) has the capability to directly train subword models from raw text, eliminating the need for prior tokenization, we pre-tokenize it as (Kudo and Richardson, 2018) has shown better results with tokenized input. Also, SP supports subword regularization, which dynamically enhances the training data with on-the-fly tokenization during NMT model training. This process

[1]https://anoopkunchukuttan.github.io/indic_nlp_library/

contributes to the construction of a robust and accurate model, and it is not tied to any specific architectural configuration. Our experimentation with SentencePiece, implemented with independent vocabularies for English and Mizo, involves two main approaches:

(a) *With subword regularization:* With this method the model encounters different variations of subword splitting of the same word which could in turn be beneficial in producing a robust model for agglutinative languages. We have set the number of nbest candidates to 16 and the smoothing parameter to 0.1.

(b) *Without subword regularization.*

2.3 Using Monolingual Data

A relatively large monolingual data have been provided which could be made to use in various ways like constructing embeddings or for data augmentation. Back translation (Sennrich et al., 2015a) is a popular data augmentation method that exploits the target side monolingual data to create synthetic parallel corpus. Back Translation uses a base target→source model (initially trained on the limited genuine bitext) to translate the target side monolingual data. The synthetic data thus generated can serve as supplementary resource and could be explored in various ways.

Re-training the model on the manifold synthetic data is expected to boost up the model producing better translations. Two obvious assumptions can be made on the performance of an NMT model for low-resource scenario:

1. Data augmentation could boost up the model.
2. Also, more error-free the training data is, better is its performance.

Based on these assumptions and inspired by the previous reports on back-translation with iterations such as (Cotterell and Kreutzer, 2018; Hoang et al., 2018), we use an innovative twist to improvise model by using back translated data iteratively rather than using all in one go. In every iteration, the model is trained with increased data back translated by the previous iterations's improved model along with the original bitext thereby producing better translations for the next iteration. As the synthetic data is prone to error which could in turn hamper model performance (Poncelas et al., 2018), we add the back translated data proportionate to the size of the genuine bitext. Also, the trained model is followed by finetuning on the genuine bitext for further improvement (Tonja et al., 2023). Our method could be summarized by the following algorithm:

Algorithm 1 An innovative usage of Back Translation

Require: Authentic parallel corpus(S_0, T_0), target monolingual corpus(M), number of splits (n)
$M_0 \leftarrow Train_{(Target \rightarrow Source)}(S_0, T_0)$
$C_1, C_2, ..., C_n \leftarrow Split(M, n)$
such that $|C_i| \propto |S_0|$
$i \leftarrow 1$
while $i \leq n$ **do**

$$(S_i, T_i) = (S_0, T_0) \bigcup (M_{i-1}(\sum_1^i C_i), \sum_1^i C_i)$$

$$M_i \leftarrow Train_{(Target \rightarrow Source)}(S_i, T_i)$$

$$M_i \leftarrow Finetune M_i(S_0, T_0)$$

$i \leftarrow i + 1$
end while

2.4 Post-Editing

The predicted translations (for English, Mizo and Khasi) have been post-edited in the following ways:

1. *Truecasing:* A truecaser model has been trained on the training set with the Moses' truecaser script.
2. Capitalizing the first character of every prediction.
3. As the text in the test set is not completely detokenized with several punctuation markers space separated, adjustments have been made to replicate the reference translations.

3 Experiments and Results

Experimental Setup: All the experiments have been conducted on the opensource NMT toolkit, OpenNMT (Klein et al., 2017). Subword vocabulary size is kept at 8000. The Transformer (Vaswani et al., 2017) has been customized to work on the small-scale dataset by simplifying the standard model. After conducting experiments

Table 3: Experimentation setup for English-Assamese

Model	En/Dec Layers	Attention Heads	Dimensions	Batch Size
Model 1	6	8	512	512
Model 2	3	4	256	256
Model 3	6	4	256	256

En/Dec : Encoder/Decoder

Table 4: Experimentation setup for English-Khasi

Model	Batch Size	BPE Vocab Size	En/Dec Layer	Attention Heads
M_1	256	6000	3	4
M_2	512	6000	3	4

En/Dec : Encoder/Decoder

with various parameter sets (including encoder and decoder layers, heads, embedding size, and feed-forward nodes), we have determined that the optimal configuration for English-Mizo data consists of 3 encoder and 3 decoder layers, a word vector size of 512, and 2048 nodes in the feed-forward layer. For English Assamese pair, three models have been built with varying hyperparameters and training is performed in both the directions. For English Khasi, two models have been built and trained. The model descriptions for English-Assamese and English Khasi are shown in Table 3 and Table 4 respectively. All the models are trained using the Adam optimizer with an initial learning rate of 2, incorporating Noam decay and 8,000 warm-up steps. The training process continues for 200,000 steps, with validation performed every 10,000 steps. Additionally, checkpoints are saved at 10,000-step intervals, and early stopping is implemented with a patience of 4 based on validation perplexity and accuracy.

Checkpoint Selection: Throughout training, checkpoints are saved every 10,000 steps. Among all the checkpoints generated, the model with the best validation perplexity and validation accuracy is chosen as the model for testing purposes.

3.1 Results

In Table 5 we report our results on the initial experiments using various subword tokenization schemes for English-Mizo. Our results have been evaluated by four evaluation metrics as provided by the organizers. It is clear from the results that Byte Pair Encoding using independent vocabularies works best for this data. Hence, for all the future experiments, BPE with independent vocabularies is selected as the standard format. Also, it should be noted that we have reported the results obtained with BPE (shared vocabulary) as the primary results for both En → Mizo and Mizo → En directions.

Table 5: Results obtained with various Subword mechanisms (English-Mizo).

Method	BLEU	CHRF	TER	RIBES
English → Mizo				
SP_{wo_reg}	22.63	44.93	58.07	0.75
SP_{w_reg}	23.78	48.06	58.07	0.75
BPE_{sh}	23.29	46.72	59.93	0.75
BPE_{ind}	**25.58**	**48.19**	**57.35**	**0.76**
Mizo → English				
SP_{wo_reg}	20.65	40.98	72.8	0.67
SP_{w_reg}	18.51	41.32	73.7	0.67
BPE_{sh}	18.81	40.33	73.65	0.66
BPE_{ind}	**20.95**	**41.38**	**72.43**	**0.67**

SP_{wo_reg} : SentencePiece without Subword regularization

SP_{w_reg} : SentencePiece with Subword regularization

BPE_{sh} : Byte Pair Encoding with shared vocabulary

BPE_{ind} : Byte Pair Encoding with independent vocabulary

Table 6 summarizes the result obtained by our method of using proportionate back translated data which is in turn generated by the model developed in the previous iteration. The baseline scores are obtained by using 1M back translated data (translated by SP_{w_reg} model) which acheives a BLEU of value of 16.77 for En → Mz. In the 1st iteration, equal size of back translated data is added to the genuine bitext and the model is trained from scratch. It is able to achieve a BLEU score of 20.42. This shows the negative impact of adding a large size synthetic data, which is not error-free, relative to the authentic parallel data. Also, a significant improvement is noticed after fine-tuning on the given authentic data. Similar results are also seen in the 2nd iteration. The successive improvement is a successful implementation of our novel usage of back translation method.

The English-Assamese and English-Khasi experiments have been conducted using various configuration of the Transformer model as shown in Table 3 and Table 4 respectively. This is done to find the optimal model configuration for the languages. Though English and Khasi share the same script, the morphologies are completely different

Table 6: English-Mizo BLEU scores with our *novel* usage of Back Translation (BT)

Method	BT Data Size	En->Mz	Mz ->En
Baseline	1M	16.77	14.40
1st Iter.	50K	20.42	16.20
	FineTuned	26.01	20.06
2nd Iter.	100K	22.04	18.19
	FineTuned	**26.63**	**20.81**

Table 9: Results for English Khasi pair.

English → Khasi				
Model	**BLEU**	**CHRF**	**RIBES**	**TER**
M_1	10.41	33.31	0.63	71.67
M_2	10.27	32.63	0.63	70.71
Khasi → English				
M_1	8.74	30.54	0.63	79.64

and as Table 5 clearly manifests, appropriate hyper-parameter values can bring about significant impact in the performance. The results obtained for English → Assamese is shown in Table 7 and Assamese → English is shown in Table 8. From both the tables, we see that Model 1 has shown the best results in both English → Assamese and Assamese → English translation. We, therefore, select the results obtained for Model 1 as the primary score. For English-Khasi, the results for model (M_1) is submitted as the primary score.

Table 7: Results for English → Assamese

Model	BLEU	CHRF	TER	RIBES
Model 1	4.89	25.16	87.21	0.46
Model 2	4.27	24.59	90.13	0.43
Model 3	3.75	22.65	93.57	0.42

Table 8: Results for Assamese → English

Model	BLEU	CHRF	TER	RIBES
Model 1	5.5	25.81	80.1	0.56
Model 2	4.7	24.96	81.53	0.55
Model 3	4.14	23.73	83.41	0.53

4 Conclusion

In this study, we have provided a comprehensive overview of our Neural Machine Translation (NMT) system developed for three language pairs: English-Assamese, English-Khasi, and English-Mizo, encompassing both translation directions. Our research delved into the intricacies of model configurations (Transformer layers, heads, batch sizes, etc.) and subword tokenization schemes (Byte Pair Encoding and SentencePiece and its variants). Through rigorous experimentation, we identified and adopted the optimal configurations for each language pair.

Challenged by the inherent scarcity of data in these low-resourced language pairs, we innovatively leveraged monolingual data to augment our translation models. We have presented a novel variation of a well-established technique for addressing the challenges of low-resourced NMT systems: *Back Translation*. This adaptation yielded remarkable results, surpassing the performance of conventional Back Translation methods by a substantial margin.

5 Limitation

We use the standard tokenization implementation (Moses) for English, Mizo and Khasi. Though Moses seems to work fine for English, certain disparities (associated with language-specific characters) are observed for Mizo and Khasi, both morphologically rich languages. Similar observations are also noted for Assamese. Using a customized tokenizer for these languages is believed to enhance the results which needs further investigation.

The dataset given was too small for Neural Machine Translation trainingespecially for Khasi. Though Back Translation is a well known method for low-resource setting, merely translating and using it as a pseudo-parallel corpus may not help as the monolingual data quality also has an impact. We have not used any mechanism to judge the quality. With our method, we iteratively use incremented back translations which is observed to boost the model. But the translation data is proportional to the original parallel corpus size which hinders leveraging fully the large monolingual corpus. We would like to explore ways to fully exploit the large availability of monolingual corpus for data augmentation or linguistic embellishments. Monolingual data usage is not explored (due to time constraint as we joined late) for the English-Assamese and English-Khasi which we plan to investigate in future. Our overall system lags in producing correct translations for long sentences. Semi-automatic post editing is utilized which needs further investigations in automatising the process.

References

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

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.

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

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

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*.

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*.

Surafel M Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2018. Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen. 2020. Improving multilingual neural machine translation for low-resource languages: French, english-vietnamese. *arXiv preprint arXiv:2012.08743*.

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

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. *CoRR*, abs/1508.07909.

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

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.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.