Impacts of Approaches for Agglutinative-LRL Neural Machine Translation (NMT): A Case Study on Manipuri-English Pair

Gourashyam Moirangthem[a], **Lavinia Nongbri**[a], **Samarendra Salam**[b] and **Kishorjit Nongmeikapam**[a]
[a]Department of CSE, Indian Institute of Information Technology Manipur, Imphal
[b]Department of Mathematics, G.P. Women's College, Imphal

{gourashyam, lavinia, kishorjit}@iiitmanipur.ac.in, samar.crypt@gmail.com

Abstract

Neural Machine Translation (NMT) is known to be extremely challenging for Low-Resource Languages (LRL) with complex morphology. This work deals with the NMT of a specific LRL called Manipuri/Meeteilon, which is a highly agglutinative language where words have extensive suffixation with limited prefixation. The work studies and discusses the impacts of approaches to mitigate the issues of NMT involving agglutinative LRL in a strictly low-resource setting. The research work experimented with several methods and techniques including subword tokenization, tuning of the self-attention-based NMT model, utilization of monolingual corpus by iterative back-translation, embedding-based sentence filtering for back translation. This research work in the strictly low resource setting of only 21204 training sentences showed remarkable results with a BLEU score of 28.17 for Manipuri to English translation.

1 Introduction

Machine translation (MT) is the process of translating text automatically, without human intervention, from one language to another using machines in general and computers in particular. Neural machine translation (NMT) systems, which are based on neural networks, are able to perform on par with human translators (Toral et al., 2018; Popel et al., 2020). But these systems were trained with data sets that have tens or even hundreds of millions of parallel sentences. Large-scale data sets are only accessible for a select few high-resource language pairs.

A language that is considered low-resource or under-resourced is one that lacks electronic tools, resources, linguistic expertise or a distinctive or reliable writing system (orthography) (Krauwer, 2003). In the context of MT, low-resource languages (LRLs) are those that do not have enough parallel data sets to train the deep learning models. Manipuri/Meeteilon is one such low-resource language. It is the state language of the North Eastern Indian state of Manipur and is one of the official languages specified in the Eighth Schedule of the Constitution of India. It is classified as a Tibeto-Burman language and is known to be highly agglutinative (Post et al., 2012). The foundational Natural Language Processing (NLP) works on Manipuri language has been slowly but steadily picking up pace in the last decade or so (Naskar and Bandyopadhyay, 2005; ?; Nongmeikapam et al., 2012b; Meetei et al., 2020; Huidrom and Lepage, 2020; Singh and Singh, 2020; Moirangthem and Nongmeikapam, 2021; Laitonjam and Singh, 2021; Moirangthem and Nongmeikapam, 2021; Rahul et al., 2021; Laitonjam and Singh, 2022; Singh and Singh, 2022b; Maibam and Purkayastha, 2023; Meetei et al., 2023). Although few attempts have been made to collect parallel corpus and to build efficient translation models, the low-resource nature and the complex morphology of the language has hampered the progress to attain tangible successes. Therefore, innovative ways has to be studied to improve the translation accuracy in the low-resource settings.

This work is a serious attempt to study the impacts of experimention with innovative methods to improve the translation accuracy of Manipuri (Mni) to English (En) NMT in a strictly low-resource setting. The paper is organized as follows. Section 2 discusses the literature review of previous works conducted in the area, Section 3 discusses the methods of the experiments conducted and Section 4 discusses the results of the experiments. Finally Section 5 outlines the conclusion and aspects for future work.

2 Literature Review

Machine Learning (ML) has been boosted by the rediscovery of neural networks (Goldberg, 2016). The introduction of NMT has enabled the use of a single large neural net that directly transforms the source sentence into the target sentence (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2014). Different neural architectures have been proposed with the goal of improving efficiency of translation. This includes recurrent networks (Sutskever et al., 2014; Bahdanau et al., 2014; Luong and Manning, 2015), convolutional networks (Kalchbrenner et al., 2016; Gehring et al., 2017; Kaiser et al., 2017) and transformer networks (Vaswani et al., 2017).

Few progress have recently been made regarding the research in Manipuri NMT also. Singh and Bandyopadhyay (2010c) reported an example-based MT system for Manipuri and English. Statistical models were also explored (Singh and Bandyopadhyay, 2010b,a; Achom et al., 2018; Rahul et al., 2021). Attempts to integrate reduplicated multiword expressions (RMWEs) into the Phrase Based Statistical Machine Translation (PBSMT) to improve translation quality has also been reported (Singh and Bandyopadhyay, 2011). Studies on the Name Entity Recognition (NER), Morpheme Identification etc. of Manipuri to assist downstream NLP pipelines have been worked upon (Nongmeikapam et al., 2011, 2012a). Investigations were also reported with different supervised and unsupervised methods of Statistical Machine Translation (SMT) and NMT (Singh and Singh, 2020; Singh et al., 2021; Singh and Singh, 2022b). Laitonjam and Singh (2021) studied the normalization of the morphological inflection issue of Manipuri, and to induce inter-language connecting points between Manipuri and English. Singh and Singh (2022a) used Long Short-Term Memory (LSTM) based many-to-many multilingual NMT system that is infused with cross-lingual features. Moirangthem et al. (2022) worked on devising an efficient Manipuri-English automatic sentence aligner based on embeddings. Huidrom and Lepage (2022) studied on a pretrained word embedding for Manipuri. Meetei et al. (2023) reported the study on Multimodal Machine Translation (MMT). Maibam and Purkayastha (2023) attempted with a factored model of SMT with a part-of-speech (POS) tag as a factor to incorporate linguistic information about the languages followed by hand-coded reordering.

Improving the MT systems by using the monolingual data to solve the parallel sentences dependency issue has been reported (Wu and Wang, 2007). Many academics have looked into using back-translation (BT) as an alternative to using monolingual data (Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011; Lambert et al., 2011; Sennrich et al., 2015; Edunov et al., 2018; Poncelas et al., 1804; Edunov et al., 2018; Rubino et al., 2020). The reported MT systems that exploited monolingual data were already trained using large parallel corpus. For low-resource languages like Manipuri, back-translation has found to be attempted by a few (Achom et al., 2018; Singh et al., 2021; Laitonjam and Singh, 2021; Singh and Singh, 2022b). However, there have been reports of noisy data issues that require post-processing which makes the work less viable (Singh and Singh, 2020). Currey et al. (2017) showed that low resource language pairs can also be improved with synthetic data where the source is simply a copy of the monolingual target data. Hoang et al. (2018); Cotterell and Kreutzer (2018); Dou et al. (2020) tried to implement iterative procedure that enhances the back-translation and final systems' quality over time.

3 Experimentation

Experiments were carried out to get the best translation accuracy for the Mni to En translation. Figure 1 summarizes and illustrates the final overall system architecture. The methodologies used to construct the NMT systems can be divided into three main categories, which are covered in the subsections that follow: 1) Data preparation for NMT, 2) Building optimal NMT model and 3) Taking advantage of Monolingual data.

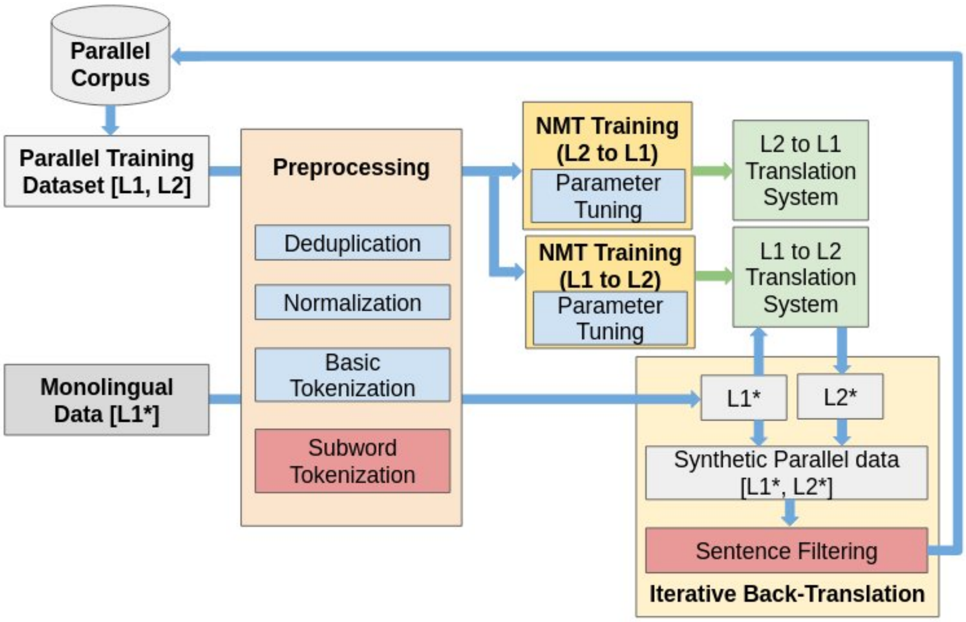

Figure 1: Complete overview of the proposed architecture

3.1 Data Preparation for NMT

The parallel corpus for training, validation and testing was taken from the repository provided by the LRILT task, WMT 23 [1]. The statistics of the corpus collected is illustrated in Table 1. The data acquired was in two parts as 1) Mni-En parallel data for training, validation and testing; and 2) a large monolingual corpus of Mni. Mni script in both the corpus was in Bengali script. Firstly, the parallel sentences were checked manually for the quality of misalignment and other possible syntactic and semantic errors by using random sampling. Random sampling is used to study the quality of the corpus because it is not feasible to manually verify all the sentences. According to the findings about the alignments, spellings and quality of translations of the corpus, automatic cleaning of the corpus is performed. The important steps in data cleaning may be discussed as follows:

Deduplication The corpus was analyzed for duplicates in this step. The monolingual corpus of 2144897 Mni sentences was found to include a major chunk of duplicates. The duplicate sentences were automatically removed using Python which reduced the number of sentences to 298662. The parallel training data also had duplicates. Out of the 21687 parallel sentences, there were 781 duplicate sentences in En side and 483 duplicate sentences in Mni side as reported by the python program. The duplicate sentence pairs were also removed from the corpus.

Normalization The normalization step is done to make the whole dataset consistent for MT pipeline. For this, the removal of control characters, byte order mark, zero width joiner, zero width non joiner etc. and replacement of the zero width space, no break space etc. with space was performed for both the languages. For En, the normalization step includes another step of lower-casing where all the capitalized texts are converted to lowercase. This step is important as the capital letter and small letter signify different characters in computer programming, although the difference seems to be minor to human.

Basic Word Tokenization Tokenization has been an important part of NMT pipeline. The tokenization of the sentences into proper words is important for NMT work as any character joined to a word will be taken as a new word different to the original one by the translation system. For example, 'go' and 'go!' will be interpreted as different words. Hence, the texts in both the languages are first tokenized into words not polluted by any preceding or

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

Data	Sentences acquired	After preprocessing
Training (Mni-En)	21687	21204
Validation (Mni-En)	1000	1000
Testing (Mni-En)	1000	1000
Monolingual (Mni)	2144897	298662

Table 1: Statistics of corpus acquired from LRILT task, WMT 23

succeeding punctuation. For En, the text is tokenized using the regular Latin punctuation. For Mni, in addition to the above, the Meetei Mayek fullstop punctuation mark '꯫' is also used for the purpose.

3.2 Building Optimal NMT Model

The NMT model used for training the Mni-En translation is adapted from the works of Vaswani et al. (2017). This is the self-attention based encoder-decoder model commonly known as Transformer. A base model was developed using Python and Tensorflow. The model architecture can be simplified and summarized into Encoder and Decoder as illustrated in Figure 2. In simple terms, the Encoder finds the relationships between the tokens of the input sequence. In essence, the encoder transforms the embedding space of the input tokens by weighing the vectors in accordance with their significance to the meaning of the sentence. Decoder takes the translated sentence as the first input and applies attention to it. Then it combines the result with the encoder's output in another attention mechanism. The Decoder learns to relate the target embedding space with the input embedding space, so that it finds a basis transformation between both vector spaces.

According to the works of Van Biljon et al. (2020), the optimal size of the NMT model for low-resource settings are not necessarily large. Thus the base model (NMT_b) is created in a lightweight fashion by following their work to save temporal and computing resources for the incremental study of several experimentation factors. It has 2 transformer self-attention heads with 3 layers in encoder and decoder. The size of rnn hidden states is 256 and the size of hidden transformer feed-forward is 512. The model uses Adam optimization. The starting learning rate is 1 with "noam" learning rate decay method. The number of model training steps is set to 100000. After the base model is created, experiments were conducted to study the trend of improvement in translation accuracy using two major strategies viz. Sub-word tokenization and parameter tuning, which are discussed below.

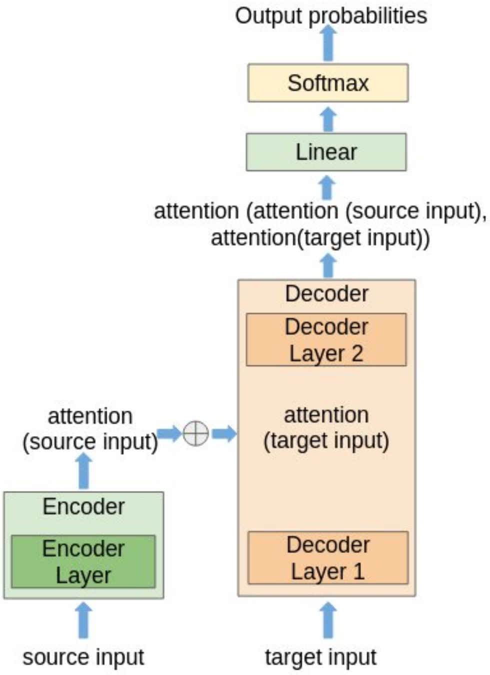

Figure 2: NMT model, adapted from (Vaswani et al., 2017)

3.2.1 Sub-Word Tokenization

NMT requires a limited-size vocabulary for computational cost and substantial number of examples to estimate word embeddings. Three advantages are frequently cited in favor of subwords: shorter encoding of frequent tokens, compositionality of subwords, and the ability to deal with unknown words (Wolleb et al., 2023). The more frequent the token, the better its representation. NMT shows weakness in translating low-frequency words (Koehn and Knowles, 2017). Therefore it is desired that a vocabulary contains a series of well represented high-frequency tokens. As Manipuri language is highly agglutinative, the number of unique words is disproportionately large which is counterproductive for NMT systems as explained above.

Subword tokenization is basically the splitting words into subwords. In this step, the word tokens are further tokenized into subwords. For example, the word ”Tokenization” may be split into subwords as ”Token” and ”ization”. There are several methods for subword tokenization. In this work, the Byte-Pair-Encoding (BPE) based subword-tokenization has been explored.

Byte Pair Encoding(BPE)-based Tokenizer BPE is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data. It was first described in (Gage, 1994), and it has been used as a popular data tokenization technique (Tennage et al., 2018). It begins by treating each word as a sequence of characters and iteratively combines most commonly occurring character pair into one. The algorithm stops after a controllable number of operations, or when no token pair appears more than once. BPE ensures that the most common words are represented in the vocabulary as a single token while the rare words are broken down into two or more subword tokens and this is in agreement with what a subword-based tokenization algorithm does.

3.2.2 NMT Model Parameter Tuning

The base model (NMT_b) with the added subword tokenization layer is used for further experiments to study if making the model larger or smaller through parameter tuning would improve the translation accuracy. Experiments were conducted by gradually changing the parameters in the direction of increasing translation accuracy. Parameters that were tweaked in the Transformer model may be named as RNN size, Feed forward layer size, encoder decoder layer number, number of heads etc. Parameters of the model giving the best accuracy in the experimentation are chosen for the next improvement strategy.

3.3 Taking Advantage of Monolingual Data

In the constrained data setting of a very small parallel corpus, attempts were made to utilize the huge monolingual corpus in improving the translation accuracy. Two main methods were employed in utilizing the monolingual corpus viz. Iterative back-translation and Automatic sentence filtering by embeddings-based scoring.

3.3.1 Iterative Back-Translation

Back Translation is the process of augmenting the parallel training corpus with the translations of the target language sentences, which is basically generating synthetic parallel data using the existing translation system (Edunov et al., 2018). Iterative Back-translation is used to generate increasingly better synthetic parallel data from monolingual data (Hoang et al., 2018).

For the iterative back-translation of Language-1 (L_1) and Language-2 (L_2), the models for L_1 to L_2 translation ($NMT_1{}^2$) and L_2 to L_1 translation ($NMT_2{}^1$) is trained first with available parallel corpus. Then the Language-1 monolingual data ($L_1{}^*$) is used to translate synthetic Language-2 ($L_2{}^*$) sentences using the trained translator model ($NMT_1{}^2$). This synthetic parallel data [$L_1{}^*$, $L_2{}^*$] is added to the original parallel data for training to improve the accuracy of $NMT_1{}^2$ and $NMT_2{}^1$ models. If improvement is seen, the process is repeated till the convergence condition is reached or till the computing and time resource is viable. This can be illustrated in Figure 1 and Algorithm 1.

Algorithm 1 Iterative Back-Translation.

```
Require: Parallel Data [$L_1$, $L_2$], Monolingual L1 Data ($L_1{}^*$),NMT models $NMT_1{}^2$ and $NMT_2{}^1$
Training Data ($D^p$) $\leftarrow [L_1, L_2]$
repeat
    Train $NMT_1{}^2$ using $D^p$
    Use $NMT_1{}^2$ to translate ($L_2{}^* \leftarrow L_1{}^*$)
    Filter $[L_1{}^*, L_2{}^*]$ using similarity score
    $D^p \leftarrow [L_1, L_2] \cup [L_1{}^*, L_2{}^*]$
    Train $NMT_2{}^1$ using $D^p$
until Convergence
Ensure: Updated NMT models $NMT_1{}^2$ and $NMT_2{}^1$
```

It has also been pointed out in the earlier work of Edunov et al. (2018), that in resource poor settings back-translations are of not much use. Hence, the main issue is how to enable the iterative back-translation method in resource poor settings. This is tried to be solved by using quality sentence filtering method which is explained below.

3.3.2 Automatic Sentence Filtering by Embedding-based Scoring

Sentence filtering is the method to choose better translations among the back-translated synthetic sentence pairs. An automatic sentence filtering methodology was devised. The methodology applied is based on the works of Moirangthem et al. (2022) which used embeddings based cosine distance for sentence alignments for Mni-En sentence pairs. A sentence similarity scoring method was devised to filter better translated parallel sentences among the noisy back-translated sentences. The scoring function was based on normalized cosine distance between multilingual sentence embeddings. Cosine similarity states that to find the similarity between two points or vectors A & B considering two axis X and Y, the angle between them must be found out which is given by Equation 1.

$$\begin{aligned} cosine\_similarity = cos(\theta) &= \frac{A.B}{\|A\| . \|B\|} \\ &= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \end{aligned} \tag{1}$$

Having the Cosine Similarity, the Cosine Distance is simply given by the following Equation 2

$$cosine\_distance = 1 - cosine\_similarity \tag{2}$$

When distance is less, the similarity is more which means points are near to each other. Contrarily, when the distance is more, two points are dissimilar or far away from each other. Using this scoring function, the percentage of sentence similarity was calculated, using which the better translated sentence pairs could be selected for iterative back translation.

3.4 Model Evaluation

The performance of the models was tested using the 1000 parallel test data from the LRILT task, WMT 23. The performance evaluation is done using the popular metric of Evaluation Understudy (BLEU), which is de-facto standard in measuring translation output. It indicates how similar the candidate text is to the reference texts, with values closer to one representing more similar texts. It works by counting n-grams in the generated sentence to n-grams in the reference sentence. BLEU score calculation can be stated as below (Papineni et al., 2002):

$$BLEU(N) = BP.exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{3}$$

where BP is brevity penalty and N is the number of n-grams. BP is given by the following formula:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{\frac{1-r}{c}} & \text{if } c \leq r \end{cases} \tag{4}$$

where c is predicted length which is the number of words in the predicted sentence and r is target length which is the number of words in the target sentence.

4 Result and Analysis

Each step of the experiment conducted was evaluated using the evaluation method explained in 3.4. The effects of different experiments conducted are discussed separately as below.

4.1 Tokenization Effect

Subword Tokenization proved to be one of the most important steps in improving the NMT performance for the agglutinative LRL of Manipuri language. It's effect is illustrated in Table 2.

Method	Mn-En BLEU
NMT_b	9.61
NMT_bpe+BPE	18.11

Table 2: Effects of Sub-word tokenization.

The method helps to break down the agglutinative Mni words and thus increase the frequency of subwords. This in turn helps reducing the number of vocab and to better handle the unknown/<UNK> words. The application of sub-word tokenization improved the

Sl.No.	RNN size	Transformer ff	Enc/Dec layers	Head	Mn-En BLEU
1	256	512	3	2	18.11
2	256	2048	3	16	25.15
3	512	4096	4	16	27.28

Table 3: Effects of Parameter tuning.

No. of Iteration	Sim. score	No. of parallel sent.	Mn-En BLEU
1	40%	23219	26.66
1	45%	15826	26.92
1	50%	7310	**27.34**
1	55%	2730	25.72
1	60%	850	25.35
2	55%	28468	27.56
2	60%	22401	**27.67**
2	65%	5799	27.16
2	70%	2331	27.12
3	55%	42772	27.71
3	60%	23219	**28.17**
3	65%	15826	27.74
3	70%	7310	27.68

Table 4: Effects of similarity score on back-translation by parallel sentence filtering.

translation accuracy from the base score without sub-word tokenization by a BLEU score of 8.95 for Mni to En translation. This is in line with the expected outcome as discussed in Section 3.2.1 that tokenizing the agglutinative words will greatly help improving the NMT translation accuracy.

4.2 Parameter Tuning Effect

The effects of parameter tuning is illustrated in Table 3. The first attempt of increasing the size of transformer feed forward from 512 to 2048 and the number of attention heads from 2 to 16 helped gain a BLEU score of 7.04. Again, increasing the RNN size from 256 to 512, transformer feed forward from 2048 to 4096 and increasing the encoder/decoder layer from 3 to 4 helped increase the BLEU score by another 2.13 BLEU score. The optimal tuning improved the accuracy with a BLEU of 9.17 for Mni to En translation. Thus, the experiment with parameter tuning indicated that making the model larger improved the translation accuracy in the case of NMT model involving agglutinative LRL of Manipuri.

4.3 Monolingual Data Effect

After having the subword tokenization integrated, and a tuned NMT model was prepared, attempts were made to implement back-translation. Using the Mni to En NMT model, the Mni monolingual corpus is translated back to En. A preliminary attempt was made to use the back-translated parallel sentences for training the models. This backfired and instead of improving the translation accuracy, the method greatly reduced the accuracy of translation. Upon manual inspection by random sampling, it was found as expected that most of the translated sentences were too much noisy and are not good enough for training the models to learn anything useful. It is learned from the experiment that, utilization of back-translation in low-resource setting is a hard task. Innovative ways of parallel sentence filtering was required.

The experimentation with parallel sentence filtering using the embedding based similarity score indicated a slow improvement in translation accuracy. Thus iterative back-translation was conducted by selecting parallel sentences using a minimum score. The effects of similarity score over the size of synthetic parallel corpora and ultimately the translation accuracy on applying iterative back-translation is illustrated in Table 4. As expected (Moirangthem et al., 2022), the similarity score and the number of parallel sentences are inversely proportional, whereby choosing a small score increases the number of parallel sentences and vice versa. It was found that, under the constrained condition, increasing the similarity score improved the BLEU score but plateaued at a certain point, and then declines. The maxima is different for different iterations. For example, in the first iteration 50% similarity score gave the best results, but in the third iteration 60% gave the best result. It can be inferred from the findings that for each iteration the generated translations has lesser noise thereby more sentences can be used for next iteration.

Three iterations were experimented using the methodology explained above, which proved to improve the translation accuracy of Mni to En translation by margin of 0.89 BLEU score. It can be noted that, the process of training the back-translation of the monolingual corpus, selecting parallel sentences according to similarity score and performing the next back-translation for different similarity scores consumes a lot of time and computing resources. The experimentation proved that iterative back-translation can, however minutely, improve the translation accuracy. But, comparing with human translation in the strictly low resource settings, it seems that the trade-off between the two need to be carefully calculated.

A final summary of the trend of increase in performance for the different methodologies can be referred from Table 5. The trend of increase in BLEU score and gain in BLEU for each additional methods is illustrated in Figure 3. It can be seen that, the application of BPE helped gain the BLEU score by 8.95 from the base model. Increasing the NMT model size helped the most gain in BLEU score by 9.17. And iterative-back translation by sentence filtering helped gain a BLEU score of 0.89.

Methodology	Mn-En BLEU	Gain
NMT_b	9.16	-
$NMT_b + BPE$	18.11	8.95
$NMT_{pt} + BPE$	27.28	9.17
$NMT_{pt} + BPE + BT$	28.17	0.89

Table 5: Overall trend of performance gain.

5 Conclusion

The research work has studied various methodologies like subword tokenization, parameter tuning, iterative back translation using embeddings based sentence similarity score for the Transformer based NMT model in low-resource setting involving agglutinative Manipuri language. The impacts of each methodology and their advantages and disadvantages are discussed. The work helped achieve a BLEU score of 28.17 for Mni to En translation in a strictly constrained low-resource environment of LRILT Task, WMT-23.

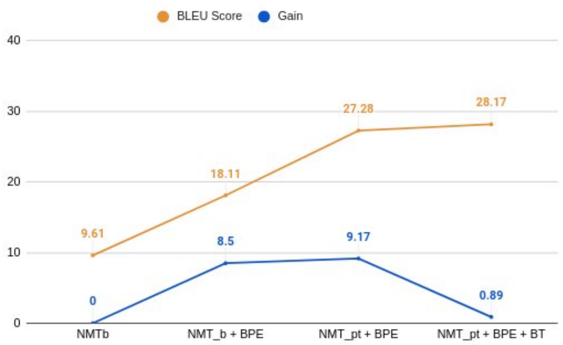

Figure 3: Trend of increase in BLEU score and gain in BLEU for each additional methods

The work also exposed several areas for future research. Tokenization for Manipuri should be studied further. Apart from BPE, other subword tokenization methods may be studied and experimented. The effect of iterative back-translation is needed to be studied along with better performing NMT models which are already trained with more parallel sentences.

Acknowledgement

This research work was supported by grant from Ministry of Electronics & Information Technology (MeitY), Govt. of India vide approval no. 11(1)/2022-HCC(TDIL)-Part(4).

References

Amika Achom, Partha Pakray, and Alexander Gelbukh. 2018. Addressing the issue of unavailability of parallel corpus incorporating monolingual corpus on pbsmt system for english-manipuri translation. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 299–319. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *EACL 2009 Fourth Workshop on Statistical Machine Translation*, pages 182–189. ACL.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the sixth workshop on statistical machine translation*, pages 330–336.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

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

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the second conference on machine translation*, pages 148–156.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*.

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

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.

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.

Rudali Huidrom and Yves Lepage. 2020. Zero-shot translation among indian languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 47–54.

Rudali Huidrom and Yves Lepage. 2022. Introducing em-ft for manipuri-english neural machine translation. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–6.

Lukasz Kaiser, Aidan N Gomez, and Francois Chollet. 2017. Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Steven Krauwer. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, pages 8–15.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-english machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88.

Lenin Laitonjam and Sanasam Ranbir Singh. 2022. A hybrid machine transliteration model based on multi-source encoder–decoder framework: English to manipuri. *SN Computer Science*, 3(2):1–18.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Sixth Workshop on Statistical Machine Translation*, pages 284–293.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79.

Indika Maibam and Bipul Syam Purkayastha. 2023. Reordering of source side for a factored english to manipuri smt system. *International journal of electrical and computer engineering systems*, 14(3):285–292.

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

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Sivaji Bandyopadhyay, Mihaela Vela, and Josef van Genabith. 2020. English to manipuri and mizo post-editing effort and its impact on low resource machine translation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 50–59.

Gourashyam Moirangthem, Lavinia Nongbri, Ningthoujam Johny Singh, and Kishorjit Nongmeikapam. 2022. Embeddings-based parallel corpus creation for english-manipuri. In *International Conference on Communication and Intelligent Systems*, pages 489–502. Springer.

Gourashyam Moirangthem and Kishorjit Nongmeikapam. 2021. A back-transliteration based manipuri meetei mayek keyboard ime. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 1–6. IEEE.

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2005. Use of machine translation in india: Current status. In *Proceedings of Machine Translation Summit X: Posters*, pages 465–470.

Kishorjit Nongmeikapam, Vidya Raj RK, Yumnam Nirmal, and Sivaji Bandyopadhyay. 2012a. Manipuri morpheme identification. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 95–108.

Kishorjit Nongmeikapam, Vidya Raj RK, Oinam Imocha Singh, and Sivaji Bandyopadhyay. 2012b. Automatic segmentation of manipuri (meiteilon) word into syllabic units. *arXiv preprint arXiv:1207.3932*.

Kishorjit Nongmeikapam, Tontang Shangkhunem, Ngariyanbam Mayekleima Chanu, Laisuhram Newton Singh, Bishworjit Salam, and Sivaji Bandyopadhyay. 2011. Crf based name entity recognition (ner) in manipuri: A highly agglutinative indian language. In *2011 2nd National Conference on Emerging Trends and Applications in Computer Science*, pages 1–6. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alberto Poncelas, Dimitar Shterionov, Andy Way, G Wenniger, and Peyman Passban. 1804. Investigating backtranslation in neural machine translation. 2018.

M Popel, M Tomková, J Tomek, Łukasz Kaiser, Jakob Uszkoreit, and Ondrej Bojar. 2020. Z. ˇzabokrtskỳ. transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the seventh workshop on statistical machine translation*, pages 401–409.

Laishram Rahul, Loitongbam Sanayai Meetei, and HS Jayanna. 2021. Statistical and neural machine translation for manipuri-english on intelligence domain. In *Advances in Computing and Network Communications*, pages 249–257. Springer.

Raphael Rubino, Benjamin Marie, Raj Dabre, Atushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2020. Extremely low-resource neural machine translation for asian languages. *Machine Translation*, 34(4):347–382.

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

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

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

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

Telem Joyson Singh, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2021. English-manipuri machine translation: An empirical study of different supervised and unsupervised methods. In *2021 International Conference on Asian Language Processing (IALP)*, pages 142–147. IEEE.

Thoudam Doren Singh and Savaji Bandyopadhyay. 2010a. Statistical machine translation of english-manipuri using morpho-syntactic and semantic information. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Student Research Workshop*.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010c. Manipuri-english example based machine translation system. *Int. J. Comput. Linguistics Appl.*, 1(1-2):201–216.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Pasindu Tennage, Achini Herath, Malith Thilakarathne, Prabath Sandaruwan, and Surangika Ranathunga. 2018. Transliteration and byte pair encoding to improve tamil to sinhala neural machine translation. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 390–395. IEEE.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *arXiv preprint arXiv:2004.04418*.

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.

Benoist Wolleb, Romain Silvestri, Giorgos Vernikos, and Ljiljana Dolamic Andrei Popescu-Belis. 2023. Assessing the importance of frequency versus compositionality for subword-based tokenization in nmt. *arXiv preprint arXiv:2306.01393*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21:165–181.