NEURAL MACHINE TRANSLATION BETWEEN MYANMAR AND ENGLISH LANGUAGES

Nang Zin Min Aye
Assistant Lecturer, Faculty of Computer Science, University of Computer Studies, Yangon, No.(4) Main Road,
Shwe Pyi Thar Township, Yangon, Myanmar,
Yangon, 11411, Myanmar
nangzinminaye@ucsy.edu.mm
http://www.ucsy.edu.mm

Khin Mar Soe
Professor, Faculty of Computer Science, University of Computer Studies, Yangon, No.(4) Main Road, Shwe Pyi
Thar Township, Yangon, Myanmar,
Yangon, 11411, Myanmar
khinmarsoe@ucsy.edu.mm
http://www.ucsy.edu.mm

Abstract

This study aims to present a thorough overview of cutting-edge machine translation solutions and evaluate their effectiveness in translating between Myanmar and English language pairs. Translating Myanmar poses challenges due to its unique language characteristics and limited resources. These challenges include the complex Burmese script, the tonal nature of the language, its agglutinative structure, and the scarcity of parallel corpora for training. The integration of attention mechanisms, Transformer architecture, and transfer learning in Neural Machine Translation (NMT) models demonstrates their potential to deliver accurate, context-aware, and adaptable translations. A key achievement of this study is highlighting the significant roles of hyperparameter optimization and the use of subword units in enhancing the effectiveness of Transformer-based NMT for Myanmar-English and English-Myanmar language pairs, despite resource constraints. The experimental results showed that using a Transformer-optimized model with a 32k Byte Pair Encoding (BPE) subword model resulted in significant improvements in Bilingual Evaluation Understudy (BLEU), Translation Error Rate (TER), and Character n-gram F-score (ChrF) scores compared to the baseline Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) and fine-tuned models such as mBART and mT5. The proposed transformer model achieved impressive BLEU scores of 50.77 for English to Myanmar and 48.92 for Myanmar to English. The comparative studies of various machine translation models have confirmed that both the Transformer and fine-tuned models demonstrate promising results in improving low-resource NMT performance for the Myanmar and English languages.

***Keywords*: Neural Machine Translation(NMT), Myanmar (Burmese), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Transformer, Fine Tuning, mT5, mBART.**

1. Introduction

Neural Machine Translation (NMT) models, particularly those employing the Sequence-to-Sequence (Seq2Seq) architecture with attention mechanisms, have significantly enhanced translations for resource-rich languages by learning from paired input-output data [14]. Transformer models have further advanced translation capabilities by increasing efficiency through better parallelization and mitigating long-term memory issues [2]. However, NMT models face challenges with low-resource languages and infrequent words due to their dependence on large volumes of training data. To overcome these difficulties, researchers are focusing on pre-training, fine-tuning, incorporating linguistic knowledge, and utilizing syntax-aware models. Efforts include creating manual parallel corpora, augmenting data, and employing pre-trained language models. These strategies aim to narrow the performance gap between high-resource and low-resource languages, making translations more reliable and accessible.

In the realm of low-resource languages such as Myanmar, NMT systems trained on Myanmar-English datasets frequently encounter reduced translation quality due to data scarcity, linguistic complexity, and domain-specific challenges. Selecting the right subword models is critical for language pairs like Myanmar-English because subword modeling helps to overcome the fixed-vocabulary limitations of NMT systems. Methods like Byte Pair

Encoding (BPE) allow models to learn representations for a more extensive vocabulary by encoding rare and unknown words as sequences of subword units [20], [24]. Although most research on subword models has concentrated on high-resource languages [21], [22], there has been a lack of recommendations regarding the best subword model types for Myanmar-English translation. Character-based models, despite being simple and memory-efficient, often provide limited semantic information, making them less effective for both English and Myanmar languages.

Researchers have investigated translating between Myanmar and English using hybrid approaches that combine statistical and neural methods to address specific linguistic challenges [25], [3]. They have also utilized data augmentation techniques such as back-translation and the creation of parallel data from monolingual sources to enlarge training datasets [19]. NMT models, particularly those incorporating attention mechanisms, have shown promise in this area [28], [29], [30]. Subword segmentation techniques like Byte Pair Encoding (BPE) and the unigram language model are crucial in low-resource contexts, where the choice of segmentation level and vocabulary size greatly influences MT performance [13]. Despite various successful strategies for low-resource NMT, their application to the Myanmar-English language pair has been insufficiently explored due to the complex script of Myanmar, rich morphology, and limited availability of high-quality parallel corpora. This research addresses these issues by optimizing NMT models with subword tokenization techniques, fine-tuning pre-trained multilingual models such as Multilingual Denoising Pre-training for Neural Machine Translation (mBART) [27] and Massively Multilingual Pre-trained Text-to-Text Transformer (mT5) [12], and identifying significant research gaps in data scarcity and model optimization. The contributions include experimental validation and the development of tailored evaluation metrics, aiming to enhance translation quality and provide a robust foundation for future advancements in low-resource NMT for Myanmar and English.

The primary objective of this research is to identify the most effective subword model type, balancing translation quality. This involves conducting experiments and empirical analyses to determine the optimal subword modeling approach for maximizing translation performance. Initial improvements were made to the baseline LSTM-RNN [5] and Transformer [2] NMT models by optimizing hyperparameters and applying BPE and Unigram tokenization for both languages. Additionally, pre-trained multilingual models such as mBART [27] and mT5 [12] were fine-tuned, and their performance was compared with the optimized Transformer models and the baseline LSTM-RNN. The experimental results indicated that using a Transformer-optimized model with a 32k Byte Pair Encoding (BPE) subword model significantly enhanced Bilingual Evaluation Understudy (BLEU) [11], Translation Error Rate (TER) [17], and Character n-gram F-score (ChrF) [16] scores compared to the baseline Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) and fine-tuned models utilizing mBART and mT5. Comprehensive analyses of all NMT systems were conducted to assess their performance.

2. Methodologies

2.1. Neural Machine Translation

Machine translation (MT) has significantly advanced through three main approaches: rule-based (RBMT), statistical-based (SMT), and neural machine translation (NMT). RBMT depends on human-crafted linguistic rules and dictionaries, ensuring domain-specific accuracy but struggling with scalability and adaptability due to the labor-intensive maintenance required for extensive rule sets. SMT employs a data-driven approach, using large bilingual corpora to statistically infer translation patterns, offering greater flexibility but often losing coherence in longer texts. NMT represents a major breakthrough, utilizing deep learning and sequence-to-sequence models with attention mechanisms to model translation as an end-to-end process. By capturing entire sentence contexts, NMT improves coherence and contextual understanding, overcoming the limitations of RBMT and SMT. Despite its high computational and data requirements, NMT excels in handling linguistic ambiguities and producing more accurate, contextually appropriate translations, making it the preferred method for machine translation.

2.1.1. LSTM-RNN with Attention Mechanism

LSTM-RNN with attention mechanisms has significantly improved sequence modeling tasks, including machine translation [5]. LSTM networks solve the vanishing gradient problem of traditional RNNs by using gating units that allow information to pass across multiple time steps, effectively capturing long-range dependencies. In machine translation, LSTM-based sequence-to-sequence models encode the source language into a fixed-size vector representation and generate the target language based on this representation. However, these models often struggle with long sequences, leading to performance degradation. Attention mechanisms enhance these models by allowing them to focus on relevant parts of the input sequence during decoding, dynamically assigning more weight to specific input tokens. This improves the handling of long sequences and results in more accurate translations. The baseline NMT model employed a two-layer LSTM architecture for both the encoder and decoder, incorporating attention mechanisms. Despite their effectiveness, LSTM-based architectures have been largely replaced by Transformer-based models, which offer superior translation quality and efficiency and have become the preferred choice for state-of-the-art machine translation systems.

2.1.2. Transformer

The Transformer architecture significantly transformed neural machine translation (NMT) by moving away from recurrent neural network (RNN)-based models and introducing a self-attention mechanism paired with a feed-forward neural network [2]. This change aimed to improve computational efficiency and facilitate training on extensive datasets. The encoder-decoder framework of the Transformer comprises layers with distinct sub-layers: the multi-head self-attention mechanism enables concurrent attention to various parts of input sequences, enhancing comprehensive context comprehension. The feed-forward network further refines outputs to abstract and transform the encoded data. Similarly structured decoder layers include mechanisms to focus on input sequences and utilize encoder outputs for precise translation. Residual connections and layer normalization ensure stable training and smooth information flow. This symmetrical design enhances the ability of the Transformer to generate high-quality translations by effectively capturing dependencies and contextual nuances within sequences. It has become the cornerstone for subsequent models and architectures in NLP, with variants like BERT (Bidirectional Encoder Representations from Transformers) [10] and GPT (Generative Pre-trained Transformer) [1] further extending its capabilities.

The self-attention mechanism, the core operation of the Transformer model, maps query vectors (Q) and a set of key-value pairs (K, V) to an output. This process calculates the output matrix by applying a softmax function to the dot product of the query and key vectors, scaled by the square root of the dimensionality (d_k).

$$Attention(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V} \qquad (1)$$

To overcome the limitations of using a single attention head, the Transformer model utilizes multi-head attention. This mechanism enables the model to simultaneously attend to information from various representation subspaces at different positions. Each attention head independently computes its own query, key, and value matrices, which are then concatenated and linearly transformed to generate the final output. Figure 1 illustrates the detailed internal representation of the Transformer architecture.

$$MultiHead(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}\ (head_1\ , \ldots, head_h\ )W^O \qquad (2)$$

where, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$.

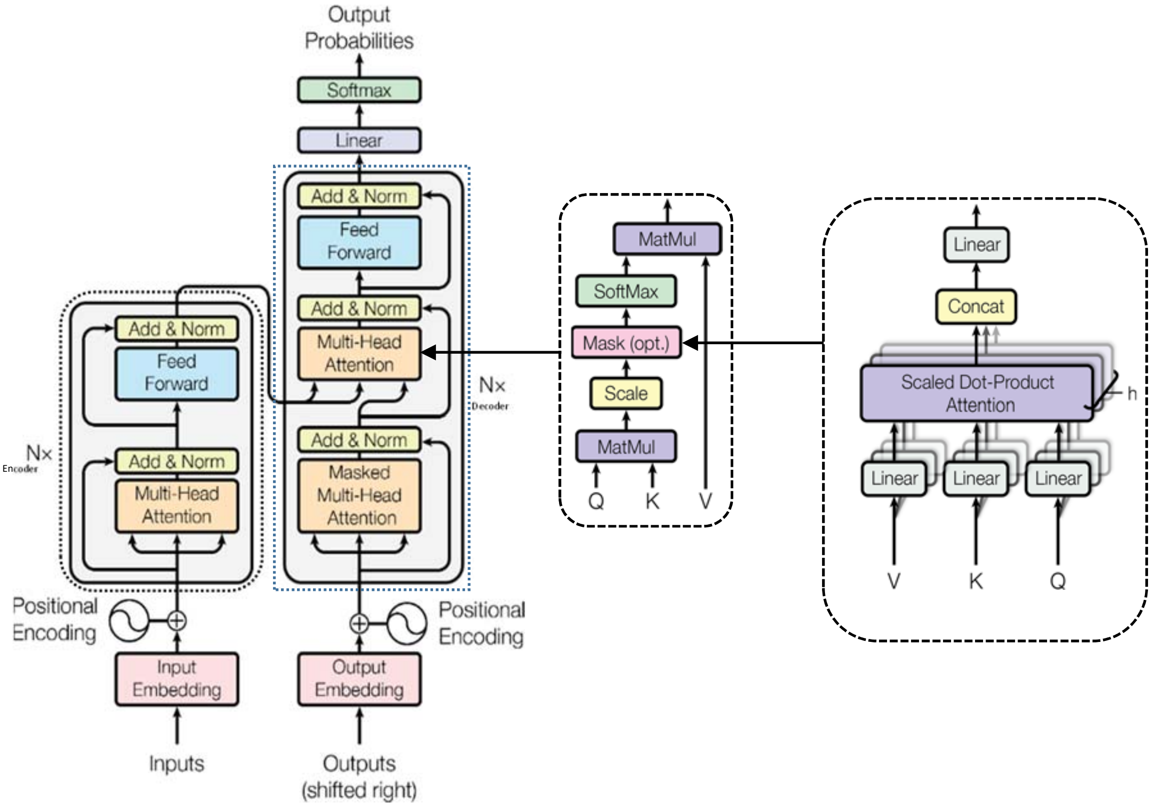


Figure 1. The internal representation of the Transformer architecture

2.1.3. Multilingual Denoising Pre-training for Neural Machine Translation (mBART)

mBART, an extension of the BART model specifically designed for bilingual machine translation tasks, employs a Seq2Seq transformer architecture featuring 12 layers each for both the encoders and decoders, amounting to approximately 680 million parameters [27]. It excels in capturing cross-lingual dependencies by fine-tuning on diverse multilingual corpora that are tokenized into subword units. The training objectives include masked language modeling, sentence permutation, and translation tasks, with optimization achieved through hyperparameters such as layer configurations, attention heads, and learning rates using an Adam optimizer. For bilingual translation, mBART is fine-tuned on parallel corpora to maximize translation accuracy, adjusting hyperparameters as necessary. This approach eliminates the need for retraining from scratch and instead focuses on optimizing existing capabilities for specific translation needs, demonstrating the effectiveness of mBART in bilingual machine translation applications.

2.1.4. Massively Multilingual Pre-trained Text-To-Text Transformer (mT5)

mT5 is a transformer-based model developed for multilingual machine translation and other text-to-text tasks across more than 100 languages [12]. It operates within a unified framework that maintains task consistency by converting input text into the desired output format. Utilizing encoder-decoder layers with advanced attention mechanisms, mT5 excels at producing accurate and contextually relevant translations. Trained on the vast Common Crawl dataset, mT5 leverages a rich multilingual corpus that facilitates cross-lingual transfer, allowing it to adapt even with limited parallel data for certain languages. The model is available in various sizes, from mT5-Small to mT5-XXL, accommodating different computational resources and specific application requirements. Fine-tuning for specific tasks further improves its performance, especially for low-resource languages, enabling high-quality translations in few-shot or zero-shot learning scenarios. However, ensuring high-quality training data and addressing potential biases are crucial for maximizing the effectiveness of mT5 across diverse applications.

3. Experimental Setup

3.1. Datasets and Preprocessing

A parallel corpus is an essential resource for NLP, particularly for machine translation systems. The Myanmar language is considered resource-constrained, and the availability of Myanmar-English parallel corpora remains limited. This research utilizes datasets from two sources: the Asian Language Treebank (ALT) corpus [26] and the University of Computer Studies, Yangon (UCSY) corpus [30]. The ALT corpus, part of the Asian Language Treebank project, includes a 20k Myanmar-English parallel dataset covering diverse topics such as Wikipedia articles, news, travel, food, culture, and general information, providing extensive language coverage. The UCSY corpus, compiled by the NLP Lab at UCSY, contains a larger 200k Myanmar-English parallel dataset from various sources [30]. It encompasses news, travel, food, cultural content, general information, educational materials, legal and government documents, and literary works, offering a comprehensive resource for natural language processing tasks.

NMT models were utilized with bilingual datasets from the ALT and UCSY corpora, collectively referred to as the ALT+UCSY corpus, without incorporating additional extensive monolingual language models. The training datasets included parallel source and target sentences, each on a separate line with tokens separated by spaces. The ALT+UCSY dataset was randomly divided into training, validation, and testing subsets, with both the validation and testing datasets containing 2000 parallel sentences each, designated as test set 1. To further illustrate the effectiveness of the optimized transformer models, another publicly available open-source parallel dataset, known as TALPCo (Translation and Language Processing Corpora) [18], containing 1,369 parallel sentences for Myanmar-English language pairs, was used as test set 2. The TALPCo dataset provides a diverse range of data types, including formal news articles, legal documents, scientific and technical texts, literary works, conversational data, educational materials, web content, cultural and historical texts, business documents, and healthcare texts.

Due to the absence of word boundary markers in Burmese script, syllable-level tokenization was determined to be the most effective method for Myanmar language translation. This approach was validated by the Syllable-NMT model, which utilizes an attention mechanism [29]. Consequently, a syllable-level tokenizer was employed during the preprocessing step [4], as illustrated in Table 1. To improve the translation quality of the NMT model, low-quality segments were filtered out from the datasets. This data filtering process involved removing misaligned sentences, empty segments, and duplicate entries. The statistics for our experimental parallel datasets are presented in Table 2.

Original Sentence:	ကျွန်တော်မနက်ဖြန်အိမ်မှာရှိချင်မှရှိမယ်။
	I may be away from home tomorrow.
Syllable Level Segmented Sentence:	ကျွန် တော် မ နက် ဖြန် အိမ် မှာ ရှိ ချင် မှ ရှိ မယ် ။

Table 1. Example of syllable-level segmented sentence

Datasets	Parallel Sentences	ALT+ UCSY Corpus	No. of Training Sentences	No. of Validation Sentences	No. of Testing Sentences
ALT	20,106	224,645	220,646	2000	2000
UCSY	204,539				
TALPCo	1369	-	-	-	1369

Table 2. Statistic of parallel sentences for ALT+UCSY corpus and open test set (TALPCo)

3.2. Subword Tokenization

Using subwords instead of entire words in machine translation helps address hardware limitations on vocabulary size by breaking down words into smaller, manageable units. SentencePiece [23], a versatile language-independent subword tokenizer designed for NMT tasks, is notable for its ability to train directly from raw sentences without pre-segmented data, making it adaptable to various languages and domains. It effectively handles languages with complex morphologies and limited linguistic resources, which is crucial for improving translation accuracy. In the context of Myanmar-to-English translation, where there is no consensus on the optimal subword segmentation, this study evaluates techniques such as BPE and unigram models. It optimizes hyperparameters for BPE models with different vocabulary sizes (8k, 16k, 32k), demonstrating their impact on enhancing translation performance across LSTM-RNN and Transformer architectures. This research highlights the importance of subword models in refining NMT systems, showing significant improvements in translation accuracy through structured evaluations.

3.3. Model Implementations

Four distinct models were employed for translating between English and Myanmar: an LSTM with attention model as the baseline, fine-tuning models using mBART and mT5, and the optimized Transformer-based model as the proposed model. All models were trained using an Nvidia Tesla K80 GPU. The baseline LSTM-RNN model and the proposed optimized Transformer NMT models were implemented using the PyTorch-based OpenNMT toolkit [7], while the fine-tuned models were executed with the PyTorch-based framework Fairseq [15]. The performance variations with different subword tokenization methods, BPE (8k, 16k, and 32k), and unigram in LSTM-RNN and Transformer models were examined.

For the baseline LSTM attention model, a 2-layer LSTM with an attention mechanism was utilized. The model featured 500 hidden units, a dropout ratio of 0.3, and a mini-batch size of 64. Each LSTM-RNN model employed unigram and BPE subword models with vocabulary sizes of 8k, 16k, and 32k. All LSTM models were trained for 100,000 steps, and the translation beam size was fixed at 5.

Choosing a random search strategy for hyperparameter optimization (HPO) in Transformer models, especially for Neural Machine Translation (NMT), acknowledges the challenge of long training times associated with these models. Random search offers an efficient way to explore the hyperparameter space without the exhaustive demands of grid search [9]. The focus on shallow Transformer architectures aligns with previous research indicating their effectiveness in improving translation quality, particularly in resource-constrained NMT scenarios. The variations tested, such as adjusting neuron count per layer and modifying layer depth, aim to strike an optimal balance between model complexity and performance. Additionally, exploring different dropout rates helps assess the impact of regularization on model robustness. Short training cycles of 5,000 steps enable quick evaluation of a wide range of hyperparameter configurations. Once an optimal setting is identified for a parameter, it remains fixed for subsequent evaluations, streamlining the optimization process. The summary of the random search results, highlighting the optimal hyperparameters that offer valuable insights to improve the performance of Transformer models in NMT tasks, includes both the encoder and decoder comprising six attention blocks, with a feedforward dimension of 2,048 and an embedding size of 512. Eight attention heads were employed, and Adam with Noam decay served as the optimizer [6]. The batch size was set at 2,048, and a cross-entropy loss function was used, along with the Adam optimizer, a learning rate of 2, label smoothing of 0.1, dropout of 0.1, and attention dropout of 0.1. Each Transformer model also utilized unigram and BPE subword models with vocabulary sizes of 8k, 16k, and 32k. Similar to the LSTM models, the Transformer models were trained for 100,000 steps.

The mBART-25 large model was fine-tuned for a bilingual setting without pre-training. The fine-tuning parameters included a maximum iteration update of 40,000, a maximum input and output text token length of 1,024, an initial learning rate of 3e-05, the Adam optimizer, a dropout probability of 0.3, and an attention dropout of 0.1. This configuration was designed to enhance the performance of the model specifically for English-Myanmar and Myanmar-English translation tasks.

For the mT5 base model, fine-tuning was conducted for English-Myanmar and Myanmar-English machine translation using the following parameters: a maximum text length (input and output) of 256, a learning rate of 5e-05, and 6 epochs. The MT5Tokenizer and MT5ForConditionalGeneration from the Hugging Face Transformers library were utilized for fine-tuning.

In all NMT systems, the optimal checkpoints were selected for translating the test datasets, and the translated hypothesis data were evaluated using BLEU [11], ChrF [16], and TER [17] metrics. Additionally, the translated data underwent statistical significance testing with the compare-mt tool, employing paired bootstrap resampling with 1,000 resamples, at a threshold of $p < 0.05$.

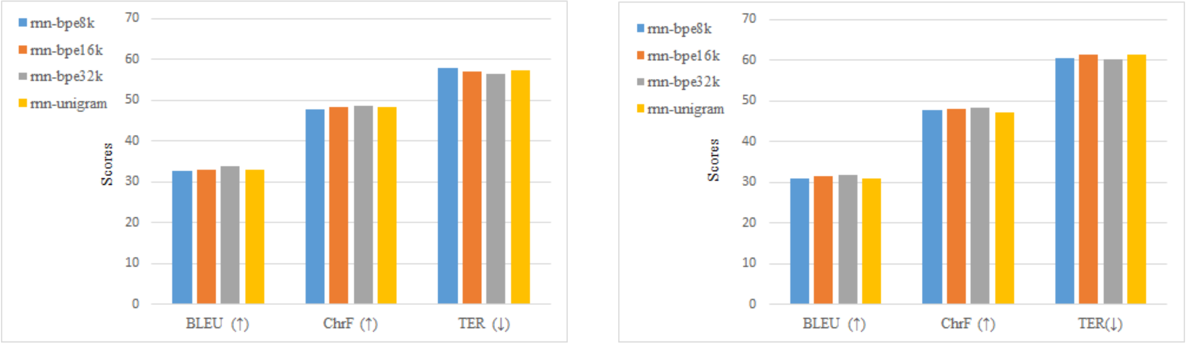
4. Findings and Analysis

The evaluation of different LSTM-RNN models with varying subword tokenization methods for English-Myanmar and Myanmar-English translation tasks indicates that the rnn-bpe32k model consistently outperforms the other configurations. It achieves the highest BLEU and ChrF scores and the lowest TER in both translation directions, signifying superior translation quality and fewer errors. These results highlight the effectiveness of using a larger BPE vocabulary size (32k) in enhancing the performance of LSTM-RNN models for neural machine translation tasks between English and Myanmar, as shown in Figure 2.

Similarly, the evaluation of different Transformer models with various subword tokenization methods for the same translation tasks demonstrates that the trans-bpe32k model consistently outperforms the others. It achieves the highest BLEU and ChrF scores and the lowest TER in both translation directions, reflecting superior translation quality and fewer errors. This underscores the effectiveness of using a larger BPE vocabulary size (32k) to enhance the performance of Transformer models for neural machine translation between English and Myanmar, leading to significant improvements in translation accuracy and overall quality, as shown in Figure 3.

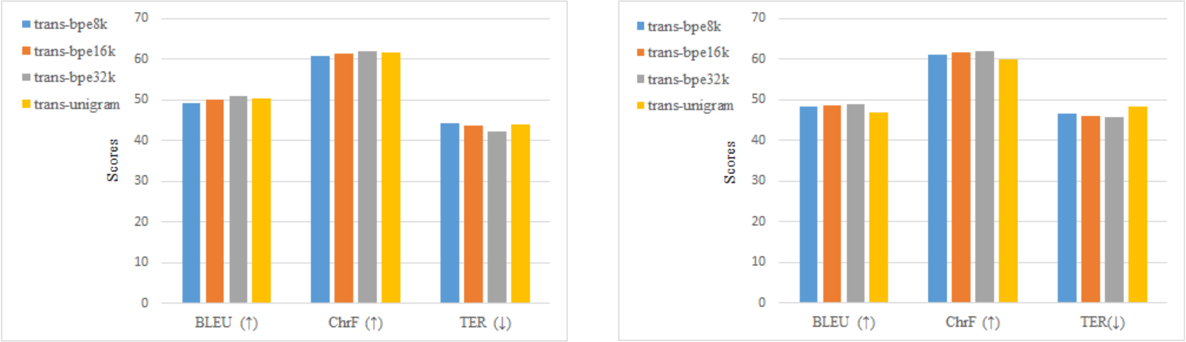
Figure 4 presents a comparison of four architectures: LSTM-RNN-32k, Transformer-32k, Fine-tuned mBART, and Fine-tuned mT5 on English-Myanmar and Myanmar-English translation tasks. In test set 1, the Transformer-32k architecture significantly outperforms others, achieving the highest BLEU scores (50.77 for English-Myanmar and 48.92 for Myanmar-English) and ChrF scores (61.84 and 61.96 respectively) while maintaining the lowest TER scores (42.29 and 45.74). This indicates superior translation accuracy and fluency. The LSTM-RNN-32k model performs moderately well, but lags behind the Transformer in all metrics. Fine-tuned mBART shows a balanced performance with BLEU and ChrF scores in the mid-range and lower TER scores compared to the LSTM-RNN-32k. However, the fine-tuned mT5 model performs the least effectively, with the lowest BLEU and ChrF scores and the highest TER, suggesting it is the least reliable for translations between these languages. The Transformer-32k model demonstrates the most robust performance, making it the preferred choice for high-quality translations in both directions between English and Myanmar.

The performance of different translation models on the TALPCo test set 2 between English and Myanmar languages, as shown in Figure 5, highlights the superiority of the Transformer-32k model. For English-Myanmar translation, the Transformer-32k achieves the highest BLEU score of 31.5 and the highest ChrF score of 53.2, alongside the lowest TER of 54.57, indicating better translation quality and fluency. Similarly, for Myanmar-English translation, it attains a BLEU score of 22.6345, a ChrF score of 44.92, and a TER of 65.9, outperforming other models. The fine-tuned mBART and mT5 models show moderate performance, with mT5 performing slightly better in Myanmar-English translation with a BLEU score of 22.5386 and a TER of 67.06. The LSTM-RNN-32k model lags significantly, especially in Myanmar-English translation, with the lowest BLEU score of 18.4397 and the highest TER of 75.22, indicating less accurate translations. The Transformer-32k model stands out as the most effective for both translation directions, providing the best balance of high BLEU and ChrF scores with low TER.



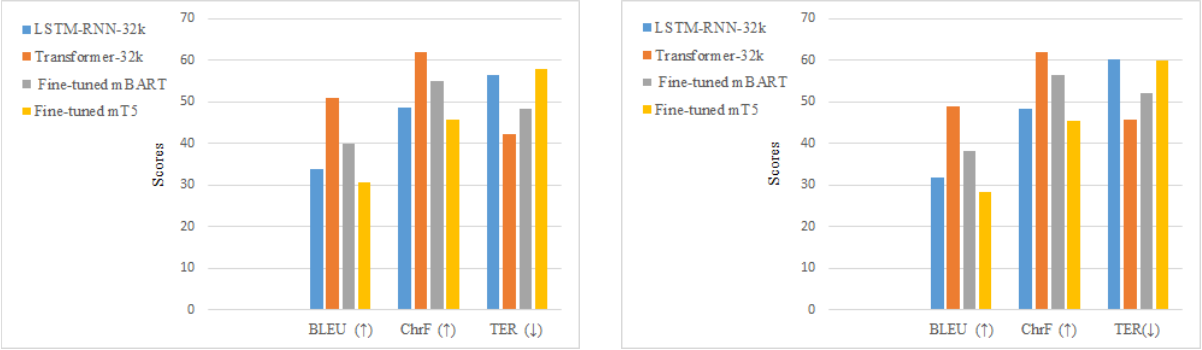
(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 2. Performance of baseline LSTM-RNN architecture on test set 1



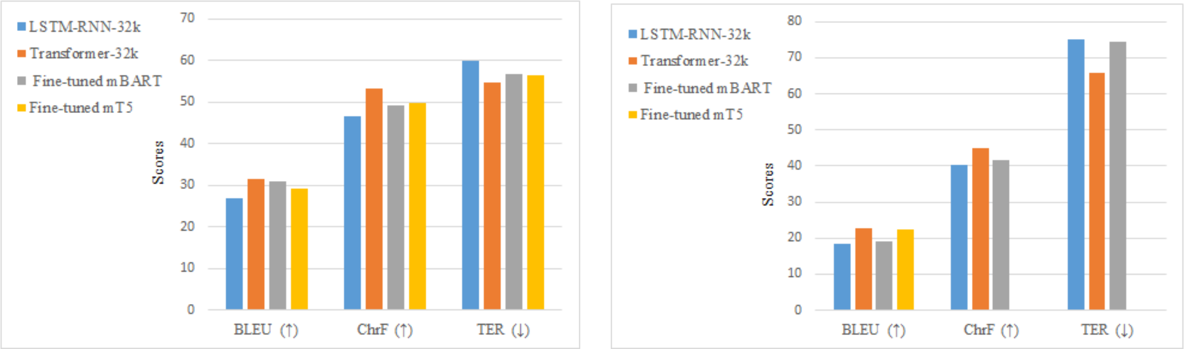
(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 3. Performance of proposed Transformer architecture on test set 1



(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 4. Performance of different translation models on test set 1 between English and Myanmar languages



(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 5. Performance of different translation models on test set 2 TALPCo (open evaluation) between English and Myanmar languages

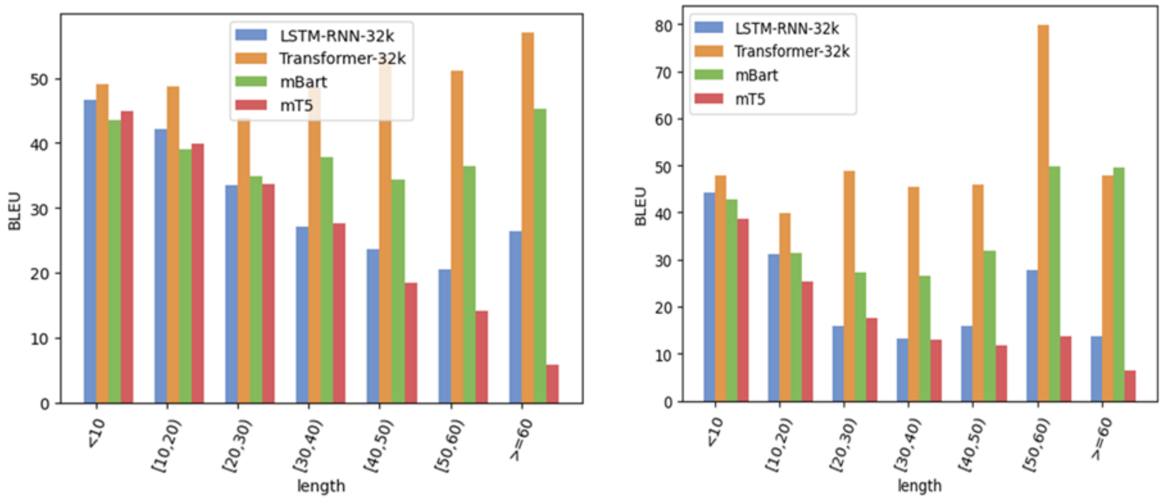
4.1. Experimental Analysis

An evaluation of well-performing models was carried out using the compare-mt tool [8] to assess their performance. Both sentence length buckets analysis and sentence-level analysis were performed to examine the specifics of all NMT systems translating between the Myanmar and English languages.

4.1.1. Sentence Length Buckets Analysis

Buckets categorize sentences based on their lengths, such as short (<10 words), medium ([10,20) to [40,50) words), and long ([50,60) and =60 words) sentences, to analyze translation quality using the BLEU metric. This approach allows for a segmented evaluation that considers how well translation systems perform across different sentence structures. By calculating BLEU scores for each bucket, compare-mt provides insights into how translation quality varies with sentence length, helping to identify strengths and weaknesses in machine translation outputs across various text complexities and linguistic nuances, as shown in Figure 6.

Based on the results highlight distinct performance, LSTM-RNN-32k and Transformer-32k excel with shorter sentences, while mBART demonstrates effectiveness in mid-length sentences. mT5 consistently outperforms other models across all buckets, especially in longer sentences, showcasing its robustness in handling complex sentence structures in Myanmar-English translation tasks. These findings underscore the importance of selecting appropriate models based on the expected length and complexity of source sentences, with mT5 emerging as a strong candidate for achieving higher translation accuracy across a broad spectrum of sentence lengths. mBART and mT5 show competitive performance in mid-length sentences, but their scores tend to fluctuate more compared to Transformer-32k. This suggests that while they can achieve high BLEU scores in specific ranges, they may require more fine-tuning or adjustments to maintain consistency across varied text complexities.



(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 6: Sentence length buckets analysis for English-Myanmar and Myanmar-English machine translations

LSTM-RNN-32k performs competitively in shorter sentences, but its scores diminish as sentence length increases, particularly in longer sentence categories. This indicates potential limitations in handling more complex sentence structures compared to the Transformer-based models. mT5 shows significant struggle with very long sentences (>=60 words), reflected in its notably lower BLEU score (5.8978). This highlights a potential area for improvement or specific adaptation when handling lengthy texts in English-Myanmar translation tasks. Transformer-32k stands out as the most consistent performer, making it a strong choice for applications requiring reliable translation quality across varying sentence lengths. Depending on specific requirements (e.g., focus on mid-length vs. longer sentences), mBART and mT5 could be suitable with additional adjustments to optimize performance in their respective strong areas. LSTM-RNN-32k remains a viable option for scenarios where shorter texts predominate, given its competitive performance in that category. Transformer-32k appears as the most versatile and robust model for English-Myanmar translation based on the provided BLEU scores, offering consistent performance across various sentence lengths and demonstrating capability across different text complexities.

In comparing Myanmar-English translation models by sentence length buckets, it is evident that Transformer-32k consistently performs the best across various lengths, especially excelling in shorter and longer sentences. For sentences with fewer than 10 words, Transformer-32k achieves the highest BLEU score of 47.83, closely followed by LSTM-RNN-32k at 44.32, while mBART and mT5 also show strong but slightly lower performance. In the [10, 20) words range, Transformer-32k continues to lead with a BLEU score of 39.93, with LSTM-RNN-32k scoring 31.22, and mBART and mT5 scoring lower, indicating some difficulty with moderate-length sentences. For [20, 30) words, Transformer-32k maintains its strength with a BLEU score of 48.86, mBART performs decently at 27.27, and LSTM-RNN-32k and mT5 show lower scores, suggesting challenges with longer sentences in this range. In the [30, 40) words bucket, Transformer-32k leads again with 45.36, followed by LSTM-RNN-32k (13.17), while mBART and mT5 show similar performance, indicating struggles with longer and more complex sentences. For [40, 50) words, Transformer-32k and LSTM-RNN-32k perform comparably around 45, but mBART shows a higher score of 31.85, and mT5 lags behind, indicating difficulty with longer sentences. In the [50, 60) words range, Transformer-32k achieves an exceptionally high BLEU score of 79.92, indicating strong performance with longer sentences, with mBART also performing well at 49.86, while LSTM-RNN-32k and mT5 show lower scores. For sentences with exactly 60 words, mBART leads with a BLEU score of 49.64, followed by Transformer-32k at 47.93, while mT5, despite its strengths in other buckets, struggles the most with very long sentences, indicating challenges in maintaining translation quality at maximum sentence length. Overall, Transformer-32k consistently performs well across various sentence length buckets, mBART demonstrates competitive performance in mid-length to longer sentences but varies more compared to Transformer-32k, mT5 shows strengths in specific buckets like [50, 60) words but struggles with very long sentences, and LSTM-RNN-32k generally shows competitive performance but tends to lag behind Transformer-32k and mBART in longer sentence categories.

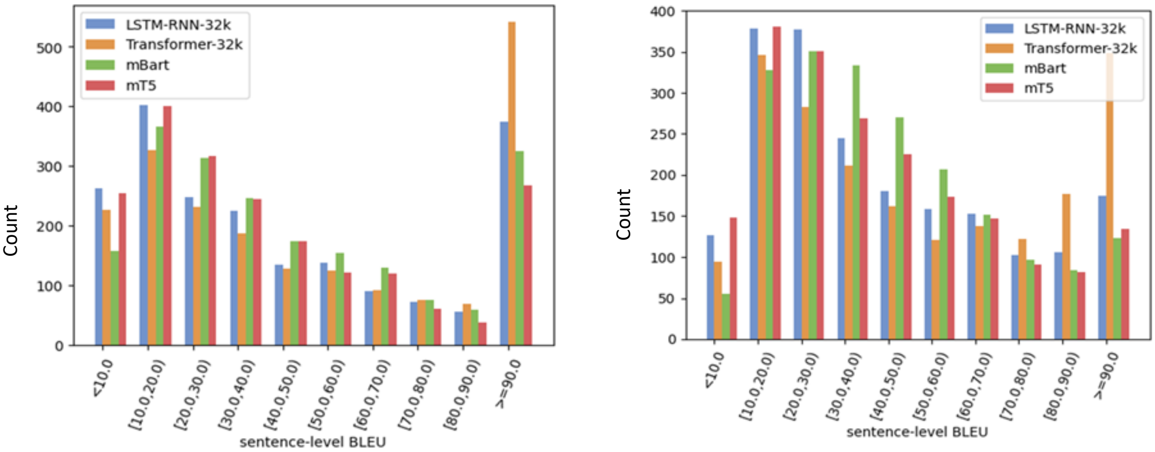
4.1.2. Sentence-level BLEU Analysis

Sentence-level BLEU analysis in compare-mt provides a detailed evaluation of translation quality by assessing each translated sentence individually rather than aggregating scores over an entire corpus. This method highlights variations in model performance across different sentence lengths, complexities, and content types, identifying specific strengths and weaknesses. By grouping sentences into buckets based on predefined criteria, such as length, this analysis reveals how well models handle various challenges in translation. It is particularly useful for comparative evaluation, allowing for a detailed comparison of how different models translate the same sentences. The insights gained from sentence-level BLEU analysis help in diagnosing model issues, guiding targeted improvements, and offering a nuanced understanding of translation quality variations across different sentence types. The detailed analysis for both translations is shown in Figure 7.

The analysis of sentence-level BLEU scores for English-Myanmar machine translation highlights significant variations in model performance across different quality buckets. mT5 demonstrates challenges with lower-quality translations, leading in the <10.0 BLEU bucket with 148 sentences, while LSTM-RNN-32k follows closely with 127. In the [10.0, 20.0) BLEU range, all models exhibit substantial counts, with mT5 having the highest at 381, indicating frequent production of translations in this lower-quality spectrum. mBART and mT5 show consistency in the [20.0, 30.0) BLEU bucket with 351 sentences each, suggesting moderate performance, whereas mBART leads in the [30.0, 40.0) BLEU range with 334 sentences, showcasing better performance in mid-level quality translations compared to other models. Transformer-32k excels in higher-quality ranges, notably leading in the [80.0, 90.0) and ≥90.0 BLEU buckets with 176 and 349 sentences, respectively, indicating superior translation quality in these categories. Overall, while mBART shows competitive performance across various BLEU score ranges, Transformer-32k consistently produces translations of higher quality, highlighting its effectiveness in English-Myanmar machine translation tasks.

The sentence-level BLEU analysis for Myanmar-English machine translation reveals distinct patterns in model performance across various BLEU score buckets. In the lowest quality bucket (<10.0 BLEU), LSTM-RNN-

32k has the highest count with 262 sentences, followed closely by mT5 with 255, indicating challenges in producing adequate translations in this range. In the [10.0, 20.0) BLEU bucket, mBART leads with 366 sentences, followed closely by mT5 with 401, suggesting moderate quality translations. Transformer-32k shows consistent performance across several buckets, notably leading in the ≥90.0 BLEU bucket with 542 sentences, indicating superior translation quality in high BLEU score ranges. mBART and mT5 demonstrate competitive performance across mid-range BLEU buckets ([20.0, 30.0) to [60.0, 70.0)), although Transformer-32k generally maintains higher counts in most quality buckets. Overall, Transformer-32k stands out for producing translations of higher quality in Myanmar-English machine translation, especially in the highest BLEU score categories, while mBART and mT5 show competitive performance across various mid-range buckets but struggle more in the lower and higher ends of the spectrum.



(a) English-Myanmar Machine Translation (b) Myanmar-English Machine Translation

Figure 7: Sentence-level BLEU analysis for English-Myanmar and Myanmar-English machine translations

5. Conclusion

In this study, four NMT models—LSTM-RNN, Transformer, fine-tuned mBART, and mT5—were evaluated for English-Myanmar and Myanmar-English translation tasks. The Transformer-32k model consistently outperformed the others across various evaluation metrics. It was found that the optimized Transformer-32k model achieved superior BLEU, ChrF, and TER scores for all translation pairs, surpassing the baseline LSTM-RNN and fine-tuned models. The proposed transformer model achieved impressive BLEU scores of 50.77 for English to Myanmar and 48.92 for Myanmar to English. The Transformer-32k models, particularly those with 32k BPE subword tokenization sizes, demonstrated strong capabilities in capturing intricate language patterns and minimizing errors, making them well-suited for real-world applications that demand high translation quality. Future advancements in these models will benefit from concentrated efforts in fine-tuning, data augmentation, and the refinement of evaluation metrics to better address diverse linguistic and contextual challenges in machine translation. Additionally, ongoing research will focus on developing improved fine-tuning and prompting strategies to enhance knowledge transfer from multilingual pre-trained models.

Acknowledgment

I am profoundly thankful to everyone who played a role in the successful completion of our publication paper. I extend special gratitude to my supervisors for their unwavering dedication and hard work, and to the reviewers and editors for their invaluable feedback. I also appreciate the support and resources provided by our institution. Thank you all for your contributions and support.

Conflict of interest

The authors have no conflicts of interest to declare.

References

[1] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving language understanding by generative pretraining," in Proceedings of the 2018 Conference on Neural Information Processing Systems (NeurIPS), Montreal, Canada, pp. 5513–5523, December 2018.

[2] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems 30, 2017. doi:10.48550/arXiv.1706.03762.

[3] B. Marie, A. Fujita, and E. Sumita, "Combination of statistical and neural machine translation for myanmar-english," In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation, Hong Kong, 2018. https://aclanthology.org/Y18-3007.

[4] C. Ding et al., "Towards burmese (myanmar) morphological analysis: syllable-based tokenization and part-of-speech tagging", ACM Trans. Asian Low-Resour. Lang. Inf. Process. 19, 1, Article 5, May 2019. https://doi.org/10.1145/3325885

[5] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proceedings of the International Conference on Learning Representations (ICLR), 2015. doi: 10.48550/arXiv.1409.0473.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, doi: 10.48550/arXiv.1412.6980.

[7] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," In Proceedings of ACL 2017, System Demonstrations, pp 67–72, Vancouver, Canada, 2017. https://aclanthology.org/P17-4012.pdf

[8] G. Neubig et al., "compare-mt: A tool for holistic comparison of language generation systems," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp 35–41, Minneapolis, Minnesota, 2019. https://aclanthology.org/2006.amta-papers.25.pdf

[9] J. Bergstra, Y. Bengio, "Random search for hyper-parameter optimization" Journal of Machine Learning Research, pp. 281–305, 2012.

[10] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186, June 2019. doi: 10.48550/arXiv.1810.04805.

[11] K. Papineni, S. Roukos, T. Ward and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318, July 2002. https://doi.org/10.3115/1073083.1073135

[12] L. Xue et al., "mT5: A massively multilingual pre-trained text-to-text transformer," In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 483–498, October 2020. doi: 10.48550/arXiv.2010.11934.

[13] Lankford, Seamus, H. Afli, and A. Way. "Transformers for Low-Resource Languages: Is F'eidir Linn!.", Proceedings of Machine Translation Summit XVIII: Research Track 2021, March 2024. doi: 10.48550/arXiv.2403.01985.

[14] M. Luong, H. Pham and C. D. Manning, "Effective approaches to attention-based neural machine translation", August 2015. doi: 10.48550/arXiv.1508.04025.

[15] M. Ott et al., "fairseq: a fast, extensible toolkit for sequence modeling," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota, 2019. https://aclanthology.org/N19-4009.pdf

[16] M. Popović, "ChrF: character n-gram F-score for automatic MT evaluation," In Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395, Lisbon, Portugal, September 2015. https://aclanthology.org/W15-3049.pdf

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231, Cambridge, Massachusetts, USA, 2006. https://aclanthology.org/2006.amta-papers.25.pdf

[18] Nomoto, Hiroki, K. Okano, D. Moeljadi and H. Sawada, "TUFS asian language parallel corpus (TALPCo)," Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing, pp. 436-439, 2018. https://anlp.jp/proceedings/annual_meeting/2018/pdf_dir/C3-5.pdf

[19] P. Chen et al., "Facebook AI's WAT19 myanmar-english translation task submission," Conference on Empirical Methods in Natural Language Processing, 2019, doi: 10.48550/arXiv.1910.06848.

[20] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp. 1715–1725, 2016. doi: 10.48550/arXiv.1508.07909.

[21] S. Ding, A. Renduchintala and K. Duh, "A call for prudent choice of subword merge operations in neural machine translation," May 2019. doi: 10.48550/arXiv.1905.10453.

[22] T. Gowda, J. May, "Finding the optimal vocabulary size for neural machine translation," April 2020. doi: 10.48550/arXiv.2004.02334.

[23] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, pp. 66–71, 2018. doi: 10.48550/arXiv.1808.06226.

[24] T. Kudo, "Subword regularization: improving neural network translation models with multiple subword candidates," Annual Meeting of the Association for Computational Linguistics, 2018. doi: 10.48550/arXiv.1804.10959.

[25] Y. K. Thu et al., "Hybrid statistical machine translation for english-myanmar: UTYCC submission to WAT-2021." Workshop on Asian Translation, 2021. doi: 10.18653/v1/2021.wat-1.7.

[26] Y. K. Thu, W. P. Pa, M. Utiyama, A. Finch, and E. Sumita, "Introducing the asian language treebank (ALT)," In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA). https://aclanthology.org/N19-4007.pdf

[27] Y. Liu et al., "Multilingual denoising pre-training for neural machine translation," Trans. Assoc. Comput. Ling. 8 (2020), pp. 726-742, 2020. https://doi.org/10.1162/tacl_a_00343

[28] Y. M ShweSin, T. M. Oo, H. M. Mo, W. P. Pa, K. M. Soe and Y. K. Thu, "UCSYNLP-Lab machine translation systems for WAT 2018," Pacific Asia Conference on Language, Information and Computation, 2018. https://aclanthology.org/Y18-3016

[29] Y. M. ShweSin and K. M. Soe, "Attention-based syllable level neural machine translation system for myanmar to english language pair," International Journal on Natural Language Computing, 2019. https://aircconline.com/ijnlc/V8N2/8219ijnlc01.pdf

[30] Y. M. ShweSin, K. M. Soe and K. Y. Htwe. "Large scale myanmar to english neural machine translation system" 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, Japan, pp. 464-465, 2018. doi: 10.1109/GCCE.2018.8574614.

Authors Profile

Nang Zin Min Aye is currently pursuing a Ph.D. degree at the University of Computer Studies, Yangon, Myanmar. She received a Master's degree (M.C.Sc.) in Computer Science from the University of Computer Studies, Yangon. Her research interests include Natural Language Processing, Artificial Intelligent and Machine Translation. She can be contacted at email: nangzinminaye@ucsy.edu.mm.

Dr. Khin Mar Soe received a Ph.D. (Information Technology) degree from the University of Computer Studies, Yangon, Myanmar. She is a professor at the University of Computer Studies, Yangon. Her research interests are in the areas of Natural Language Processing, Part-of-Speech Tagging, Machine Translation and Myanmar Name Entity Recognition. She can be contacted at email: khinmarsoe@ucsy.edu.mm.