Enhancing Myanmar-English Machine Translation Using Transformer Architecture

Nang Zin Min Aye
Natural Language Processing Laboratory
University of Computer Studies, Yangon
Yangon, Myanmar
nangzinminaye@ucsy.edu.mm

Khin Mar Soe
Natural Language Processing Laboratory
University of Computer Studies, Yangon
Yangon, Myanmar
khinmarsoe@ucsy.edu.mm

***Abstract*— Myanmar-English and English-Myanmar machine translation are challenging but important research areas in bridging communication gaps and facilitating access to information for speakers of the Myanmar language. Sustained research and ongoing innovation are essential to enhance the quality and accessibility of machine translation for these languages. Neural Machine Translation (NMT) models, especially based on the attention mechanism and transformer architecture, have shown promise in machine translation. These models have the capacity to grasp complex linguistic structures and manage long-range connections within language. The Transformer relies completely on attention mechanisms and has established new performance records in neural machine translation. It surpasses other sequence-to-sequence models in terms of its results. While the Transformer model has proven effective in well-resourced settings, its suitability for low-resource language pairs remains a subject of debate. Furthermore, it faces challenges when dealing with languages that have complex word forms, as well as when the data spans multiple domains. In such cases, the lack of an extensive parallel dataset presents a substantial barrier. In this research, we investigate various NMT models: the attention-based Long Short Term Memory (LSTM) encoder-decoder and Transformer based NMT models, for translating between Myanmar and English language pairs. Our findings indicate a clear advantage for the Transformer based model, surpassing LSTM by a margin of 10.91 BLEU scores for Myanmar to English and 14.11 BLEU scores for English to Myanmar.**

Keywords—NMT, Transformer, LSTM, Myanmar, English

I. INTRODUCTION

This paper presents the achievement of cutting-edge results in English to Myanmar and Myanmar to English neural machine translation using the transformer architecture, surpassing the performance of attention-based LSTM encoder-decoder model, all without the need for extensive monolingual language models. The core concept of Neural Machine Translation lies in leveraging the insights extracted from pre-existing corpora to interpret new text. Therefore, the primary prerequisite for Machine Translation (MT) is a parallel corpus. A parallel corpus serves as a valuable resource that offers substantial benefits to various natural language processing tasks, with MT systems deriving significant advantages from its use. The Myanmar language is often classified as a language with scarce linguistic resources. The availability of parallel corpora between Myanmar and English is still limited. The dataset provided by WAT2019 was used in this translation task. This dataset comprises parallel corpora originating from two separate domains: the Asian Language Treebank ALT corpus and the UCSY corpus. The ALT corpus, sourced from the Asian Language Treebank project, comprises 20K Myanmar-English parallel dataset from Wiki news. In contrast, the UCSY corpus, assembled by the NLP Lab at the University of Computer Studies, Yangon, contains 200K Myanmar-English parallel dataset gathered from diverse sources, including news articles and textbooks [3].

Our baseline model used two-layer Long Short-Term Memory (LSTM) for both the encoder and decoder with attention mechanisms [2]. In this research, our objective is to present the outcomes of English to Myanmar and Myanmar to English translation by employing both the attention-based LSTM encoder-decoder model and Transformer architectures. The translation performance of LSTM and Transformer models will be evaluated and compared subsequently. The paper is structured as follows: section 2 offers the context relevant to our work. Section 3 details the experimental setup and section 4 encompasses a discussion of the experiment results. Finally, conclusions are presented in section 5.

II. NEURAL MACHINE TRANSLATION

Machine Translation (MT), also referred to as automated translation, employs machine learning methods to transform text from one source language into another target language. Neural machine translation (NMT) has emerged as the cutting-edge technology in MT. The NMT system includes an Encoder-Decoder architecture with attention mechanisms, trained via MLE (Maximum Likelihood Estimation). This study explores two architecture variants: the recurrent LSTM-based with attention model and the Transformer model. In both methodologies, the encoder transforms the source sentence into hidden state vectors, which the decoder then employs to forecast characters in the target language. [8]. The next sections provide brief descriptions of these two model architectures.

A. Recurrent Neural Network with attention

In the field of machine translation, an RNN-based model takes a sentence in the source language as input and produces a corresponding sentence in the target language as its output. It sequentially processes the input sentence, updating its internal state with each word, and produces the output sentence word by word, taking into account the linguistic context and relationships between words in both languages. However, traditional RNNs have limitations in handling long-range dependencies in language, which can be a challenge in translation tasks. As a result, more advanced RNN variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are often used to address these issues and improve the quality of machine translation [4]. RNN with attention in MT is the integration of an attention mechanism into a RNN-based machine translation model. This attention mechanism significantly improves the ability of model ability to handle long-range dependencies and capture context effectively during translation.

Encoder (Source Language)

The encoder processes the source sentence word by word and comprises two LSTM layers. The first LSTM layer takes word embeddings x_i as input and produces hidden states h_i^1 and cell states c_i^1. The second LSTM layer takes the output of the first layer as input and produces hidden states h_i^2 and cell states c_i^2 . This two-layer LSTM encoder captures the contextual information of the source sentence.

Attention Mechanism

The attention mechanism calculates alignment scores $score_{ij}$ between the source and target words. The scores are calculated using the second-layer hidden states of the encoder h_i^2 and the previous decoder hidden state s_{j-1}. Softmax is applied to these scores to obtain attention scores a_{ij} , which represent the relevance of source words for generating the target word.

Decoder (Target Language)

The decoder generates the target sentence word by word and also comprises two LSTM layers. The first LSTM layer in the decoder takes the embeddings of previously generated target words y_{j-1}, the previous decoder hidden state s_{j-1}^1, and the context vector c_j as input. It produces new hidden states s_j^1 and cell states c_{j-1}^1. The second LSTM layer in the decoder takes the output of the first layer as input and produces the final decoder hidden states s_j^2 and cell states c_j^2. This two-layer LSTM decoder generates the target sentence while considering the context provided by the attention mechanism.

Generating Target Words

The decoder uses the final decoder hidden states s_j^2 and the context vector c_j to calculate the probability distribution over the target vocabulary for the next word. A softmax layer (Decoder Output Layer) is applied to compute these probabilities. This probability distribution is used to predict the next target word y_j in the sequence.

This architecture leverages two-layer LSTMs in both the encoder and decoder to capture sequential information and context effectively. The attention mechanism allows the model to focus on relevant parts of the source sentence during translation, resulting in improved translation quality. While RNN with attention has been effective, it has largely been replaced by Transformer-based models.

B. Transformer

Neural Machine Translation employs the Transformer architecture, a novel framework primarily relying on self-attention mechanism and multi-head attention, and has excelled in delivering cutting-edge performance across a variety of language pairs. [4]. Our proposed NMT model is based on Transformer architecture.

Self-Attention Mechanism

The core of the Transformer is the self-attention mechanism, which computes attention scores for each word in the input sequence.

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

Here, Q, K, and V represent query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The softmax operation normalizes the attention scores, and the values are weighted by the scores. Transformers often employ multi-head attention, where multiple sets of Q, K, and V matrices are used to capture different types of relationships in parallel. Transformers lack inherent positional information, so positional encodings are added to word embeddings to convey word order. Transformers consist of multiple layers of encoders and decoders. In the decoder, masking ensures that the model only attends to previous positions during autoregressive generation. The model is trained using cross-entropy loss to minimize the difference between predicted and actual target sequences [1].

C. Research in Myanmar-English Languages Pairs

Although machine translation has made substantial advancements in recent years, translating between English and Myanmar remains a challenging task due to linguistic differences, script variations, and limited parallel corpora. Myanmar script is complex and composed of characters and diacritics, making it challenging for automatic translation systems to handle. Effective machine translation models often require large parallel corpora for training. In the case of English-Myanmar, limited parallel data is available compared to more widely studied language pairs. In [9], the authors have explored hybrid approaches in English-Myanmar translation. The authors in [10] combined statistical methods with neural approaches to address specific linguistic challenges. Due to the scarcity of parallel data, researchers have explored techniques like data augmentation and synthesis to expand the size of training dataset. This includes back-translation, parallel data creation from monolingual data, and other methods [11]. NMT models, especially based on the attention mechanism, has shown promise in English-Myanmar machine translation [12]. In [13], the choice of segmentation level depends on the specific NMT model being used. Attention-based NMT benefits from word-level segmentation, while the Transformer model works better with sub-word level segmentation. This highlights the importance of aligning preprocessing choices with the characteristics of the NMT model to optimize performance in machine translation tasks.

III. Experimental Setting

A. Datasets and Preprocessing

We employed the NMT models on the bilingual datasets provided by ALT and UCSY corpus, without relying on additional extensive monolingual language models. The training datasets comprise parallel source and target sentences, with each sentence presented on a separate line, and tokens separated by spaces. The dataset was randomly split into training, validation, and testing subsets. For translation, we implement the Moses format for English text, employing word-level tokenization [7]. In the Myanmar language, there are no markers to indicate word boundaries. Based on the analysis, the appropriate tokenization for Myanmar language translation is at the syllable level. We can observe this clearly through Syllable-NMT, which relies on an attention mechanism [2]. Therefore, we utilize the syllable-level tokenizer for the Myanmar language [5]. Improving the translation quality of the output NMT model can be achieved by filtering out low-quality segments from the datasets. Our data filtering process involves the removal of misalignments, empty segments, and duplicate entries. Table 1 presents the statistics for our experimental parallel datasets.

Table 1. Statistic of parallel sentences.

Datasets	Parallel Sentences	No. of Training Sentences	No. of Validation Sentences	No. of Testing Sentences
ALT	18088	204539	2000	2000
UCSY	202627			

B. Training

We trained all our models using the Nvidia Tesla K80 GPU. We developed an attention-based English-Myanmar NMT model that operates at word to syllable levels in both language directions. This system was trained using the PyTorch OpenNMT toolkit[1] as our baseline. The model is structured with a two-layer LSTM featuring 500 hidden units in both the encoder and decoder. We selected the Adam optimizer with β1 = 0.9, β2 = 0.98, a batch size of 64, a dropout rate of 0.3, and a learning rate set to 1.0. Additionally, the vocabulary size for each language is 50,000 words. We trained the model for 100,000 steps and saved checkpoints every 10,000 steps. We constructed a transformer-based English-Myanmar NMT model that operates at the word to syllable level in both translation directions. For vocabulary construction, we utilized SentencePiece to learn a BPE (Byte Pair Encoding) vocabulary with a size of 50,000 from both English and Myanmar parallel datasets [6]. Table 2 provides the detailed hyper-parameters used in the transformer based training.

Table 2. Hyper-parameters used for transformer model

Hyper-parameter	**Value**
layers in the encoder/decoder	6
heads for transformer self-attention	8
size of hidden transformer feed-forward	2048
learning rate warm-up steps	4000
Adam optimizer	β1=0.9, β2=0.998
label smoothing	0.1
Word embedding size	512
batch size	2048
dropout	0.1
learning rate decay	0.5
learning rate	0.1

IV. EXPERIMENTAL RESULTS

To evaluate the translations produced by all NMT models, we utilized a metric called BLEU (Bilingual Evaluation Understudy). The quality of translation is assessed by comparing the output generated by a machine with that of a professionally translated human reference. The performance of each NMT model is detailed in Table 3. The translation quality of the English-Myanmar NMT based on the Transformer architecture surpasses that of an attention mechanism. The Transformer outperforms the attention-based NMT model by achieving a 14.11 BLEU score improvement in the English-to-Myanmar direction and a 10.91 BLEU score improvement in the Myanmar-to-English direction. The Transformer architecture is exclusively centered around attention mechanisms and attains state-of-the-art outcomes in both directions of English-Myanmar neural machine translation.

Table 3. Comparing BLEU Scores of RNN and Transformer models

Model	**En →My**	**My →En**
RNN with attention	32.12	30.43
Transformer	**46.23**	**41.34**

V. CONCLUSION

In this research paper, we assess the effectiveness of Neural Machine Translation (NMT) models, specifically comparing the attention-based Long Short Term Memory (LSTM) encoder-decoder with the Transformer architecture in the context of Myanmar to English translation. Our observations reveal that the Transformer-based model excels in cases where the NMT baseline models performance low. We believe there is still room for improvement by more effectively utilizing parallel data resources, combining these diverse data sources more efficiently, and developing better approaches like transfer learning and fine-tuning on pretrained models for this language pair.

REFERENCES

[1] V. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 30-40.

[2] 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.

[3] 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, 2018, pp. 464-465, doi: 10.1109/GCCE.2018.8574614.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory in Recurrent Neural Networks," in Proceedings of the Neural Information Processing Systems (NeurIPS), 1997, pp. 673-679.

[5] Ding, C., Aye, H. T. Z, Pa, W. P., Nwet, K. T., Soe, K. M., Utiyama, M., and Sumita, E. (2019). "Towards Burmese (Myanmar) Morphological Analysis: Syllablebased Tokenization and Part-of-Speech Tagging", ACM TALLIP, Vol. 19, Issue 1, Article No. 5.

[6] 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, Brussels, Belgium, 2018, pp. 1912-1922.

[7] P. Koehn et al., "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, Prague, Czech Republic, 2007, pp. 177-180.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104-3112.

[9] Y. K. Thu et al., "Hybrid Statistical Machine Translation for English-Myanmar: UTYCC Submission to WAT-2021," in Workshop on Asian Translation, 2021, doi: 10.18653/v1/2021.wat-1.7.

[10] Marie, Benjamin et al. "Combination of Statistical and Neural Machine Translation for Myanmar-English." Pacific Asia Conference on Language, Information and Computation (2018).

[11] Chen, Peng-Jen et al. "Facebook AI's WAT19 Myanmar-English Translation Task Submission." Conference on Empirical Methods in Natural Language Processing (2019).

[12] Y. M. Shwe Sin, K. M. Soe. "Attention-Based Syllable Level Neural Machine Translation System for Myanmar to English Language Pair." International Journal on Natural Language Computing (2019).

[13] Y. M. Shwe Sin et al. "UCSYNLP-Lab Machine Translation Systems for WAT 2018." Pacific Asia Conference on Language, Information and Computation (2019).

[1] https://github.com/OpenNMT/OpenNMT