mySentence: Sentence Segmentation for Myanmar Language using Neural Machine Translation Approach

Thura Aung, Ye Kyaw Thu, and Zar Zar Hlaing

***Abstract*— A sentence is an independent unit which is a string of complete words containing valuable information of the text. In informal Myanmar Language, for which most of NLP applications like Automatic Speech Recognition (ASR) are used, there is no predefined rule to mark the end of sentence. In this paper, we contributed the first corpus for Myanmar Sentence Segmentation and proposed the first systematic study with *Machine Learning based Sequence Tagging* as baseline and *Neural Machine Translation* approach. Before conducting the experiments, we prepared two types of data - one containing only sentences and the other containing both sentences and paragraphs. We trained each model on both types of data and evaluated the results on both types of test data. The accuracies were measured in terms of Bilingual Evaluation Understudy (BLEU) and character n-gram F-score (CHRF^{++}) scores. Word Error Rate (WER) was also used for the detailed study of error analysis. The experimental results show that Sequence-to-Sequence architecture based Neural Machine Translation approach with the best BLEU score (99.78), which is trained on both sentence-level and paragraph-level data, achieved better CHRF^{++} scores (+18.4) and (+16.7) than best results of such machine learning models on both test data.**

***Index Terms*—Sentence segmentation, Neural machine translation, Sequence Tagging**

I. Introduction

SENTENCE Segmentation can be defined as the task of segmenting text into sentences which is an independent unit and is grammatically linked words. Like other languages, Myanmar language has two types of sentences - formal and informal.

In the formal Myanmar language, sentences are grammatically correct and typically end with a pote-ma ("။"). Informal language is more frequently used in daily conversations with others due to its easy flow. Although the grammar used in informal communication in the Myanmar language may not be perfect, people can understand where a sentence ends. Additionally, there are no predefined rules to identify the ending of sentences in informal usages, thus a machine cannot understand itself. Therefore, in some of the applications based on conversations, i.e., Automatic Speech Recognition (ASR), Speech Synthesis or Text-to-Speech (TTS), and chatbots, sentence segmentation is needed to identify the boundaries.

To address this problem, we proposed a Neural Machine Translation (NMT) based sentence segmentation system. Like other NMT systems, for example, Myanmar-Rakhine neural machine translation [1], our approach is translating one sequence to another using deep learning algorithms. In this paper, we chose to use both Sequence-to-Sequence [2] and Transformer [3] architectures which has achieved state-of-the-art quality and efficiency for machine translation and, for supervised machine learning approach, Conditional Random Fields (CRFs), Hidden Markov Model (HMM), Ripple Down Rules based (RDR) were used in the experiments.

Thura Aung is with the Language Understanding Lab., Myanmar. author email: thuraaung.ai.mdy@gmail.com

Ye Kyaw Thu is with the National Electronics and Computer Technology Center (NECTEC), Pathum Thani, Thailand. author email: yekyaw.thu@nectec.or.th

Zar Zar Hlaing is with the Language Understanding Lab., Myanmar. author email: zarzarhlaing.it@gmail.com

Manuscript received January 15, 2023; accepted with minor revision August 19, 2023, revised September 15, 2023; published online Oct, 2023.

II. Related Works

Most Myanmar text segmentation systems use sequence tagging approaches, in which each unit is labeled, and the pairs of unit and label are trained using a supervised learning algorithm. With this approach, Win Pa Pa et al. [4] examined the effectiveness of CRFs for Myanmar word segmentation. Furthermore, there are additional text segmentation approaches for the Myanmar language. Ye Kyaw Thu et al. [5] proposed seven different word segmentation schemes for statistical machine translation systems. However, there were no methodological studies for Sentence Segmentation in informal Myanmar Language.

Previous researchers have worked on sentence segmentation problem by using Rule-based approaches (e.g., Lingua::EN::Sentence [6], which is a perl module for English Sentence Segmentation) and Machine Learning based sequence tagging approaches like CRF [7] and HMM [8]. Sadvilkar et al. also introduced multilingual rule-based sentence segmentation tool called PySBD [9] in which Myanmar Sentence Segmentation is available but it is only useful for formal usages because sentence segmentation is based on the sentence delimiter "။" pote-ma, which is not used in informal communications.

In this paper, for Myanmar Sentence Segmentation, we did experiments not only from a sequence tagging approach but also from a machine translation perspective. For machine translation approach, we converted Myanmar-tagged data into a parallel corpus containing two parallel sequences of words and tags. Similar to Sequence-to-Sequence translation systems, every word se-

quence contained in the parallel corpus is the source, and the targets are the respective tag sequences.

In our study, we used Linear-chain CRFs, HMM, and RDR for Machine Learning (ML) based sequence tagging approach and, for Neural Machine Translation (NMT), Long Short-Term Memory (LSTM) based Sequence-to-Sequence architecture and Transformer architectures were used.

III. Corpus Development

This section describes the information of *mySentence* tagged corpus, as well as an overview of word segmentation and tagged data annotation.

A. Corpus Information

Myanmar NLP researchers are facing many difficulties arising from the lack of resources; in particular parallel corpora are scarce [10]. For this reason, we annotated text data manually with *mySentence* tag information. The myPOS corpus version 3.0 contributed by Zar Zar Hlaing et al. [11] consists of 43,196 meaningful word sequences written in formal and informal formats from various domain areas and the whole corpus has already been word-segmented manually. But not all sequences are used for the experiments as sequences with only one word are ignored except for interjections.

We also collected Myanmar sentences and paragraphs from different online resources such as Facebook and Wikipedia and from the short stories available on Facebook pages [14] [15].

TABLE I
Data Resources of the corpus

Data Resources	sentence	paragraph
myPOS ver3.0 [12]	40,191	2,917
Covid-19 Q&A [13]	1,000	1,350
Shared By Louis Augustine Page [14]	547	1,885
Maung Zi's Tales Page [15]	2,516	581
Wikipedia	2,780	1,060
Others	93	672
Total	47,127	8,465

Table I shows resources of data collected to use for building mySentence Corpus for Sentence Segmentation.

B. Word Segmentation

In the Myanmar language, spaces are used only to segment phrases for easier reading. There are no clear rules for using spaces in the Myanmar language. The myPOS version 3.0 corpus has been already word-segmented manually. We used myWord word segmentation tool [16] to do word segmentation on our manually collected extended data and checked word segmentation results manually. We applied the word segmentation rules proposed by Ye Kyaw Thu et al. in myPOS [11] corpus.

The segmented example for the Myanmar sentence (How do you feel ?) is shown as follows:

Unsegmented sentence : ခင်ဗျားဘယ်လိုခံစားရလဲ
Word segmented sentence : ခင်ဗျား|ဘယ်လို|ခံစား|ရ|လဲ

C. Corpus Annotation

After word segmentation, we annotated the word sequences in the corpus into a tagged sequence of words. Each token within the sentence is tagged with one of the four tags: B (Begin), O (Other), N (Next), and E (End). The beginning word which is on the left of the sentence in Myanmar language is tagged B and the ending word of each sentence is tagged E. The three words left to the ending words are tagged N while other words in the sentence are tagged O. The tagging process was done manually for both sentences and paragraphs in the dataset.

TABLE II
Statistics of Tags in the Dataset

Tag	Frequency	Proportion
B	47,264	7.24%
E	48,690	7.33%
N	137,592	20.46%
O	436,942	64.97%

Table II shows the statistics of mySentence tags in the corpus. The tagged example Burmese sentence, (It is my car) is shown as follows:

Untagged sentence : ကျွန်တော့် ကား ပါ
Tagged sentence : ကျွန်တော့်/B ကား/N ပါ/E

If there are more than two /E tags in a sequence, it is considered to be a paragraph. The tagged example Burmese paragraph, (I am bored. I have nothing to do) is shown as follows:

Untagged paragraph : ကျွန်တော် ပျင်း လာ ပြီ ဘာ မှ လည်း လုပ် စရာ မ ရှိ ဘူး
Tagged paragraph : ကျွန်တော်/B ပျင်း/N လာ/N ပြီ/E ဘာ/B မှ/O လည်း/O လုပ်/O စရာ/N မ/N ရှိ/N ဘူး/E

IV. Methodology

In this section, we describe the methodologies used in our paper. Since the Myanmar language is a low-resource language and there is no other available open-source dataset, our mySentence corpus has been created for the sentence segmentation task. In order to compare the performances of the traditional machine learning models and the neural network-based techniques, we used two different approaches - supervised Machine Learning (ML) based sequence tagging and Neural Machine Translation (NMT) for the experiments. We also would like to study the performance differences between two different approaches, i.e, sequence tagging and machine translation.

A. Machine Learning based Sequence Tagging approach

For ML based Sequence Tagging approach, as the baselines, we used Linear-chain CRFs [17], HMM [18] [19] and RDR [20] [21] models.

- **Linear-chain CRFs** are principled probabilistic finite state models, which consider dependencies among the predicted segmentation tags that are inherent in the state transitions of finite state sequence models, on which exact inference over sequences can be efficiently performed. Domain knowledge can be incorporated effectively into segmentation.
- **HMM** is a probabilistic sequence model: given a sequence of words, which computes a probability distribution over possible sequences of labels and chooses the best label sequence for tagging. In our experiment, HMM for sequence tagging, the observation is a sequence of words and is associated with a state sequence of mySentence tags.
- **RDR** is a approach of building knowledge-based system using transformation based learning. It automatically reconstructs transformation rules in the form of Single Classification Ripple Down Rules (SCRDR) tree [22] tree. Figure 1 describes a binary tree of Single Classification Ripple Down Rules.

B. Neural Machine Translation approach

For NMT approach, we used Sequence-to-Sequence [2] and Transformer [3] architectures, which are state-of-the-art in neural machine translation models.

- **Sequence-to-Sequence** model consists of two models, i.e., the encoder and decoder. Both encoder and decoder are recurrent neural networks (RNNs). RNN architectures, which keep the relationships of every data point within a sequence throughout the time axis, are particularly used for sequence data. In our experiments, we used LSTM-based *Sequence-to-Sequence* encoder-decoder architecture. The encoder model transforms an encoding form, from the input sequence to extract the key features related to the machine translation task, as the context vector. The decoder generates the desired output sequence from the encoded input sequence. Figure 2 shows how a Myanmar word sequence - "ကျွန်တော် လုပ် စရာ မ ရှိ ဘူး" ("I have nothing to do" in English.) was translated into mySentence tags - "B O N N N E" using Sequence-to-Sequence architecture.
- **Transformer** model is based on the attention mechanism and feed-forward artificial neural network that includes encoder and decoder models. These encoder and decoder models consist of positional encoding, multi-head attention, and feed-forward neural networks. In the encoder, there are two sub-layers in each stack of the N layers. The initial sub-layer is a multi-head self-attention layer that is responsible for encoding the relevant parts of the source sequence at each translation. Positional encoding is also included in the *Transformer* model to encode the source word embeddings to know the word-order position in a sentence. The decoder similarly has a stack of N layers, but unlike the encoder, it has an extra sub-layer for executing attention over the encoder output. Figure 3 illustrates the architecture used to train the transformer model.

V. Experiments

A. Experimental Setup

TABLE III
Dataset Split for Experiments

	sent	**sent+para**
training	40,000	47,000
development	2,414	3,079
test	4,712	5,512

From the *mySentence* corpus, as shown in Figure 4, we prepared two types of data - one containing sentence-only data and the other with sentence+paragraph data, to train ML models and NMT models. And we split both types of data into training, development and test data as shown in TABLE III. Here, "sent" is the abbreviation for sentence-level data and "para" for paragraph-level data.

ML models and NMT were trained on both sentence-only and sentence+paragraph-level. For the evaluation, the models were tested on both sentence-only test data and sentence+paragraph-level test data.

Before the dataset splitting, the format of mySentence tagged datasets were converted at first. For training the CRFs model, to be able to use the software, the datasets were converted into word-tag pair columns. For the NMT approach, the tagged datasets were formatted and aligned into parallel data, one containing word-sequences and another one containing tag-sequences for word-to-tag translation.

B. Software

The following open source software and frameworks were used for the experiments of both machine learning and machine translation approaches:

- **CRFSuite** (Version 0.12) [23] is an open source tool (https://github.com/chokkan/crfsuite) for training and testing CRFs models. This tool was chosen because of its speed compared to other CRFs toolkits.
- **Jitar** [24] (Version 0.3.3) is a simple sequence labeling tool based on trigram Hidden Markov Model (HMM). The idea was first introduced by Thorsten Brants et al. [25].
- **Marian** [26] framework was used for the machine translation experiments. It is a popular self-contained machine translation toolkit focusing on efficiency for research and development. In our experiments, two Sequence-to-Sequence and two Transformer models were built for word-to-tag translation.

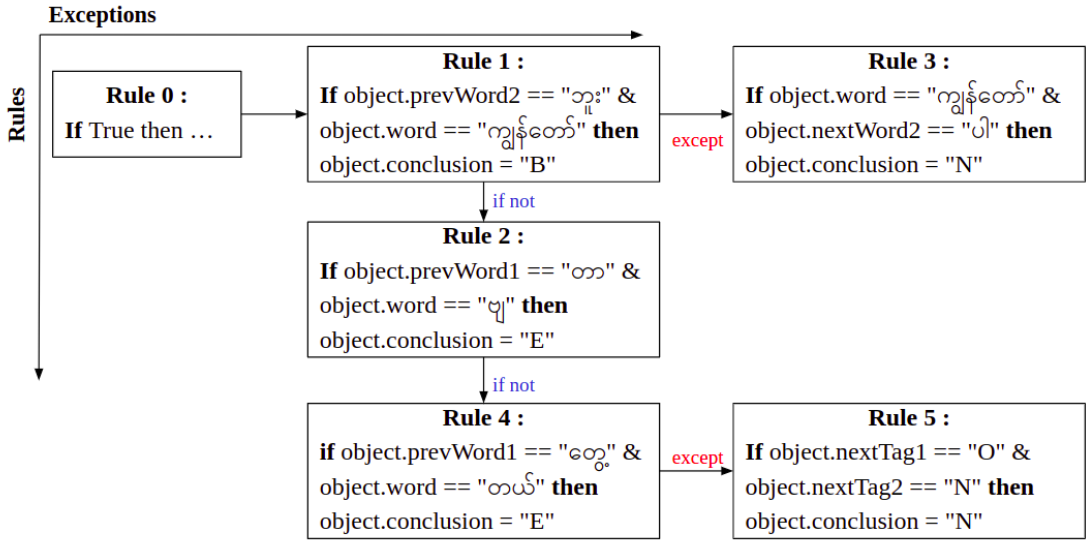

Fig. 1: A binary tree of Single Classification Ripple Down Rules

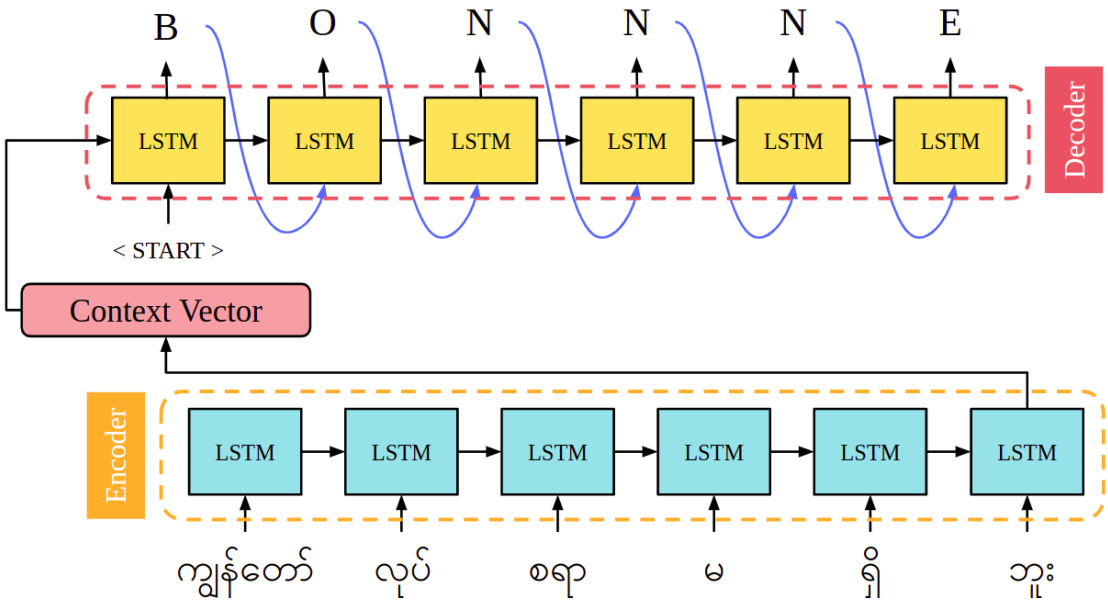

Fig. 2: Example of Sequence-to-Sequence model [2] for Neural Machine Translation based sentence segmentation

TABLE IV
HYPERPAMETERS FOR NMT MODELS

Hyperparameter	Sequence-to-Sequence	Transformer
Maximum length	200	200
Minibatch	64	1000
Early stopping	10	10
Dropout rate	0.3	0.3

C. Training

ML based sequence tagging (baseline) models were trained on both sentence-level only and sentence+paragraph-level tagged data with the default settings in the respective software.

NMT models were also trained on both sentence-level only and sentence+paragraph-level parallel data. The hyperparameters in Table IV were used to train Sequence-to-Sequence and Transformer models on one "NVIDIA GeForce" GPU with "CUDA" version 11.7.

D. Evaluation

Trained models were tested on both sentence-level test data and sentence+paragraph-level test data. We use three criteria - two to measure the evaluation of experimental results and Word Error Rate (WER) [27] to calculate the rate of error for error analysis. Character n-gram F-score (CHRF^{++}) scores [28] [29] for accuracy and Bilingual

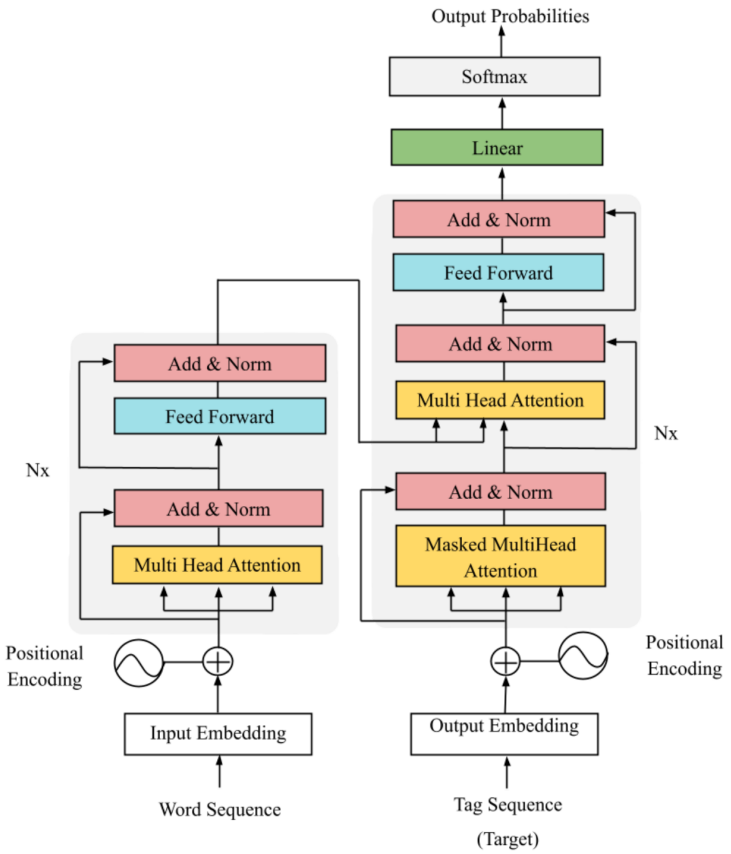

Fig. 3: Transformer architecture [3] for Neural Machine Translation based sentence segmentation

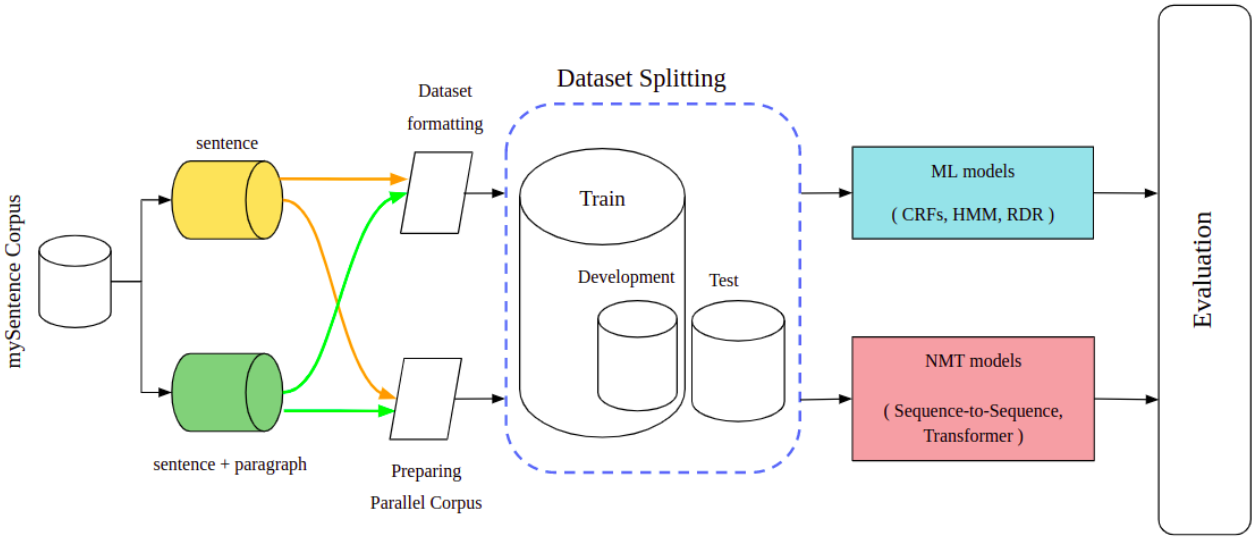

Fig. 4: Overview of Experimental Setup for Myanmar sentence segmentation with two approaches - Machine Learning based Sequence Tagging as a baseline and Neural Machine Translation

Evaluation Understudy (BLEU) [30] for the adequacy of the translation were used for evaluation of the results.

chrF has calculated the F-score of the machine-translated output with respect to the reference translation (target) based on character n-gram. The general formula for calculating chrF score is as Eq.1:

$$CHRF_{\beta} = (1+\beta^2)(\frac{CHRP.CHRR}{\beta.CHRP + CHRR}) \quad (1)$$

where CHRP and CHRR stand for character n-gram precision and recall arithmetically averaged over all n-grams.

BLEU score measures the n-gram precision with respect to the comparison of hypothesis and reference files. The higher BLEU score means the better the translation model is. BLEU score can be calculated by the following Eq. 2:

$$BLEU_n = BP \prod_{i=1}^{n} (precision) \quad (2)$$

where BP is brevity-penalty and is calculated as Eq. 3:

$$BP = \min(1, \frac{\text{output length}}{\text{reference length}}) \quad (3)$$

Word Error Rate (WER) is the percentage of words, which are to be inserted, deleted or substituted to convert the hypothesis translation of the source to the reference translation (target) of it. WER can be computed using this Eq. 4:

$$WER = \frac{(I + D + S) \times 100}{N} \quad (4)$$

where I, D, S and C refers to the number of insertions, deletions, substitutions and correct words respectively. N refers to the number of words in the reference translation (target), which can be calculated as follows:

$$N = S + D + C \quad (5)$$

VI. Result and Discussion

In this section, we describe and discuss the evaluation results of the experiments for both baselines and neural machine translation approach and an overview of error analysis in detail.

A. Evaluation Results

The chrF and BLEU score results both for ML based Tagging experiments with CRFs, HMM, RDR and for NMT experiments with Sequence-to-Sequence and Transformer models, trained either on sentence only or on both sentence and paragraph level data are shown in TABLE V, VI, VII and VIII respectively. *sent* stands for sentence-only and *sent+para* stands for sentence+paragraph-level.

Each *sent* and *sent+para* model for both baseline and NMT approach is tested on sentence-only test data as well as sentence+paragraph-level test data. Using each type of test data, the models in each approach are compared with one another in the same approach. Bold numbers indicate the highest scores in each comparison.

TABLE V
Experimental results of sent-level models in terms of chrF^{++} scores

Approach	models	sent	sent+para
Baseline	sent-CRF	76.23	74.57
	sent-HMM	74.33	73.54
	sent-RDR	**76.48**	**74.68**
NMT	sent-s2s	**94.79**	**90.46**
	sent-T	79.93	72.65

TABLE VI
Experimental results of sent+para-level models in terms of chrF^{++} scores

Approach	models	sent	sent+para
Baseline	sent+para-CRF	75.34	**76.01**
	sent+para-HMM	74.31	74.89
	sent+para-RDR	**75.91**	75.83
NMT	sent+para-s2s	**94.31**	**92.71**
	sent+para-T	93.56	87.16

Base on the results from TABLE V and VI, "sent-RDR" with 76.48 on sentence-only test data and "sent+para-CRF" with 76.01 on sentence+paragraph-level test data have highest chrF scores compared to other ML based Sequence Tagging (baseline) models. However, to compare the baseline approach with NMT approach, "sent-s2s" with 94.79 on sentence-only test data and "sent+para-s2s" with 92.71 on sentence+paragraph-level test data have highest chrF scores for each test data.

TABLE VII
Experimental results of sent-level models in terms of BLEU scores

Approach	models	sent	sent+para
Baseline	sent-CRF	88.33	84.76
	sent-HMM	88.41	**85.77**
	sent-RDR	**88.49**	84.71
NMT	sent-s2s	**99.78**	**90.93**
	sent-T	65.09	42.49

TABLE VIII
Experimental results of sent+para-level models in terms of BLEU scores

Approach	models	sent	sent+para
Baseline	sent+para-CRF	88.00	87.90
	sent+para-HMM	**88.23**	**88.24**
	sent+para-RDR	88.21	87.16
NMT	sent+para-s2s	**99.38**	**94.21**
	sent+para-T	95.67	69.88

Looking at the results in the TABLE VII and VIII, for baseline models, "sent-RDR" on sentence-only test data and "sent+para-HMM" on sentence+paragraph-level test data has higher BLEU scores of 88.49 and 88.24 than other

ML models. For the NMT approach, "sent-s2s" with 99.38 on sentence-only test data and "sent+para-s2s" with 94.21 on sentence+paragraph-level test data have the highest result compared to other models.

From the machine translation perspective, BLEU scores of both Sequence-to-Sequence models (sent-s2s and sent+para-s2s) are high meaning that the adequacy of Sequence-to-Sequence models are good. But Transformer models (sent-T and sent+para-T) have lower BLEU scores than Sequence-to-Sequence models and even lower than baseline machine learning models while testing on sentence+paragraph-level test data. With more training data and hyperparameter tuning, we believe that the performances of Transformer models will be improved since we used a batch size of 1,000 for the transformer models because of our training hardware resources. Ali Araabi et al. [31] discussed optimizing Transformers for low-resource neural machine translation and showed that with the appropriate settings, model performance can be increased substantially.

After testing on both sentence-only test data and sentence+paragraph-level test data, although the differences of CHRF and BLEU scores between each ML model are small, it is clear that the proposed word-to-tag NMT approach with Sequence-to-Sequence architectures achieved significantly better scores than the best ML models.

B. Error Analysis

The SCLITE (score speech recognition system output) program from the NIST scoring toolkit (Version 2.4.11) is used to align the machine-translated hypothesis tags with error-free reference tags and calculate the word error rate (WER). This program shows the recognition rate at the sequence level and word-level and also gives the confusion pairs.

For WER calculation, the SCLITE scoring method first aligns the hypothesis and reference sequences and then calculates a minimum Levenshtein distance which weights the cost of correct words (C), insertions (I), deletions (D), substitutions (S), and the number of words in the reference (N).

To know the counts of I, D, C, and S for the tag sequence "B O N N N E", at first, the output (hypothesis) sequence is compared to the reference sequence. Then, WER is calculated based on the counts.

```
Scores: (#C #S #D #I) 4 2 0 0
REF: B O N N N E
HYP: B N N N N N
Eval:     S       S
```

For this example, there is no deletions (D=0) or insertions (I=0) and only two substitutions (N=>O) and (O=>N) are happened so the number of correct word C is 4. Using WER equation, the SCLITE program calculated the WER value for the given example as 16.67%.

We compared WER value of each model in both Machine Learning based Sequence Tagging (baseline) approach and Neural Machine Translation approach, tested on two different test data - sentence-only and sentence+paragraph-level.

From TABLE IX and X, where bold numbers indicate the lowest percentages in each comparison, "sent-RDR" and "sent+para-RDR" have lowest percentage of Word Error Rate (WER) with 6.8% and 7.3% respectively on sentence-only test data. On sentence+paragraph-level test data, although "sent-RDR" has lowest error rate percentages with 10.5%, "sent+para-CRF" is lower than other models with 8.2% WER. Compared to the WER percentages of proposed NMT models with that of the baseline models, "sent-s2s" and "sent+para-s2s" have the lowest error rates on both sentence-only test data (0.2% and 0.5%) and sentence+paragraph-level test data (7.2% and 5.3%).

TABLE IX
WORD ERROR RATE (WER) FOR SENT-LEVEL MODELS

Approach	models	sent	sent+para
Baseline	sent-CRF	7.0%	10.7%
	sent-HMM	8.4%	11.2%
	sent-RDR	**6.8%**	**10.5%**
NMT	sent-s2s	**0.2%**	**7.2%**
	sent-T	30.5%	47.0%

TABLE X
WORD ERROR RATE (WER) FOR SENT+PARA-LEVEL MODELS

Approach	models	sent	sent+para
Baseline	sent+para-CRF	7.5%	**8.2%**
	sent+para-HMM	8.7%	9.7%
	sent+para-RDR	**7.3%**	8.4%
NMT	sent+para-s2s	**0.5%**	**5.3%**
	sent+para-T	4.1%	26.8%

We also made manual error analysis on the results of the experiments and we found that dominant errors are different. There are two frequent major error patterns: "long sequence error" and "Generalization error".

Long sequence error happens when the input Myanmar sentence or paragraph is a very long and complicated sequence. Many deletions (D) errors were found in this type of error pattern.

Long sequence error:
Source (my): သူ သည် မြန်မာ နိုင်ငံ ၏ ၁၉၄၇ ခုနှစ် ဖွဲ့စည်းအုပ်ချုပ်ပုံ အခြေခံ ဥပဒေ အရ နိုင်ငံ ၏ ဝန်ကြီးချုပ် ရာထူး ကို ၁၉၄၈ ခုနှစ် ၄ ဇန်နဝါရီ လ ဇန်နဝါရီ လ ၄ ရက် မှ ၁၉၅၆ ခုနှစ် ဇွန် ၁၂ ရက် ထိ တစ်ဖန် ၁၉၅၇ ခုနှစ် ဖေဖော်ဝါရီ ၂၈ မှ ၁၉၅၈ ခုနှစ် အောက်တိုဘာ ၁၈ ထိ နှင့် နောက်ဆုံး ၁၉၆၀ ပြည့် နှစ် ဧပြီ ၄ ရက် မှ ၁၉၆၂ ခုနှစ် မတ် ၂ ရက် ထိ ထမ်းဆောင် ခဲ့ သူ ဖြစ် သည် (Under the 1947 constitution of Myanmar, he held the position of Prime Minister of the state from January 4, 1948 to June 12,

1956, again from February 28, 1957 to October 18, 1958 and finally from April 4, 1960 to March 2, 1962 to March 2, 1962.)

Scores: (#C #S #D #I) 1 0 62 0
REF: b O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O N N N E
HYP: b *
Eval: D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D D

Sometimes our models automatically translated some words as “O” tags, which is the most common tag 64.97% in our corpus according to Table III. This type of error was defined as a generalization error because it happens mostly in machine translation models and the system tagged the unseen word in the generalized manner. NMT models with high BLEU scores concern more about the contextual role of each token than the position. Therefore, as a result of machine translation, generalization error happens mostly in the NMT approach.

Generalization error:
Source (my): သိပ် မ ငြင်း ပါ နဲ့ ကွယ် အန်ကယ့် ကို တစ် နည်း နည်း နဲ့ ကျေးဇူးဆပ် ခွင့် တော့ ပေး မှ ပေါ့ မ ဟုတ် ဘူး လား (Don’t deny it too much. Shouldn’t you let me to repay you in some ways ?)

Scores: (#C #S #D #I) 18 0 4 4
REF: b o N N N E B o o o o o o o N N N E B n n e
HYP: b o O O O O O o o o o o o o O O O O N n n e
Eval: S S S S S S S S S S

However, those two types of common error patterns even happen only when the sentence input, i.e., word sequences, are too long and too complicated with unknown vocabularies for the systems. After we made an analysis of confusion pairs, some of the confusion pairs are related to long sequence and generalization errors. Here, from TABLE XI which is the top 10 confusion pairs of the “sent+para-s2s” model tested on sentence-only data for word-to-tag machine translation, the confusion pairs of “o ==> n”, “o ==> e” and “o ==> b” are happened because of the generalization error.

TABLE XI
THE TOP 10 CONFUSION PAIRS OF “SENT+PARA-S2S” MODELS TESTED ON SENTENCE-ONLY DATA

Freq	Confusion Pair (REF==>HYP)
126	o ==> n
52	o ==> e
51	o ==> b
15	b ==> n
12	n ==> o
8	n ==> b
6	b ==> o
3	e ==> o
2	n ==> e
2	e ==> b

VII. CONCLUSION

This paper contributes the first corpus with a total size of around 55K for Myanmar sentence segmentation and the first systematic study of evaluation performances not only from ML-based Sequence Tagging but also from NMT approaches. Our proposed NMT approach with Sequence-to-Sequence architecture outperformed the Machine Learning based Sequence Tagging models and Transformer architecture. Although text segmentation for Myanmar language is mostly approached as a sequence tagging problem, we showed NMT with Sequence-to-Sequence architecture can give higher CHRF and BLEU scores with lower WER percentages than ML based Sequence Tagging approaches. In the near future, we plan to extend our tagging data and hold new experiments with other approaches and models using the extended corpus to get better performances for Myanmar sentence segmentation, especially for informal usages. We also release the corpus that was used in this experiment. (https://github.com/ye-kyaw-thu/mySentence)

REFERENCES

[1] Thazin Myint Oo, Ye Kyaw Thu and Khin Mar Soe, “Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)”, In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 80–88, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 80-88, 2019.
[2] I. Sutskever, O. Vinyals, Q. V. Le, “Sequence to sequence learning with neural networks”, In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14, MIT Press, Cambridge, MA, USA, pp. 3104–3112, December 8-13, 2014.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser, and I. Polosukhin. “Attention is all you need”, In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.
[4] Win Pa Pa, Ye Kyaw Thu, A. Finch, E. Sumita, “Word Boundary Identification for Myanmar Text Using Conditional Random Fields”, In T. Zin, JW. Lin, JS. Pan, P. Tin, M. Yokota, (eds) Genetic and Evolutionary Computing. GEC 2015. Advances in Intelligent Systems and Computing, vol 388. Springer, Cham, 2016.
[5] Ye Kyaw Thu, A. Finch, Y. Sagisaka, E. Sumita “A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation”, In Proceedings of the 11th International Conference on Computer Applications (ICCA 2013), Yangon, Myanmar, pp. 167-179, February 26 27, 2013.
[6] Lingua::EN::Sentence: https://metacpan.org/release/KIMRYAN/Lingua-EN-Sentence-0.29/view/lib/Lingua/EN/Sentence.pm
[7] K. Tomanek, J. Wermter, and U. Hahn, “Sentence and token splitting based on conditional random fields,” in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, vol. 49, p. 57, 2017.
[8] B. Jurish and K.-M. Würzner, “Word and Sentence Tokenization with Hidden Markov Models.,” JLCL, vol. 28, no. 2, pp. 61–83, 2013.

[9] N. Sadvilkar, M. Neumann, "PySBD: Pragmatic Sentence Boundary Disambiguation", In Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), Association for Computational Linguistics, pp. 110-114, November, 2020.
[10] Ye Kyaw Thu, V. Chea, A. Finch, M. Utiyama and E. Sumita, "A Large-scale Study of Statistical Machine Translation Methods for Khmer Language", In Proceedings of 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, pp. 259-269, October 30 - November 1, 2015.
[11] Zar Zar Hlaing, Ye Kyaw Thu, T. Supnithi and P. Netisopakul, "Improving Neural Machine Translation with POS-tag features for low-resource language pairs," Heliyon, vol. 8, August 2022. https://doi.org/10.1016/j.heliyon.2022.e10375
[12] Ye Kyaw Thu, "myPOS : Myanmar Part-of-Speech Corpus", GitHub Link: https://github.com/ye-kyaw-thu/myPOS
[13] NHK World-Japan, "Corona Virus Questions and Answers in Burmese", December 2022, Link: https://www3.nhk.or.jp/nhkworld/my/news/qa/coronavirus/
[14] Shared By Louis Augustine: https://www.facebook.com/sharedbylouisaugustine
[15] Maung Zi's Tales: https://www.facebook.com/MaungZiTales
[16] Ye Kyaw Thu, "myWord: Syllable, Word and Phrase Segmenter for Burmese", GitHub Link: https://github.com/ye-kyaw-thu/myWord, September 2021
[17] J. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 282–289, 2001.
[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", In Proceedings of the IEEE, vol. 77, pp. 257–286, 1989.
[19] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Prentice Hall PTR, Upper Saddle River, NJ, USA, 1 st edition, 2000.
[20] Scheffer, Tobias,"Algebraic Foundation and Improved Methods of Induction of Ripple Down Rules", pp. 23–25, 1996.
[21] Dat Q. Nguyen, Dai Q. Nguyen, D. D. Pham and S. B. Pham, "RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger", In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, Association for Computational Linguistics, pp. 17–20, 2014.
[22] D. Richards, "Two decades of Ripple Down Rules research", Knowledge Eng. Review.24, pp. 159–184, 2009.
[23] Okazaki, Naoaki, "CRFsuite: a fast implementation of Conditional Random Fields (CRFs)", 2007.
[24] de Kok, Daniël, "Jitar: A simple Trigram HMM part-of-speech tagger", 2014, [accessed 2016].
[25] T. Brants,"TnT: A Statistical Part-of-speech Tagger" , In Proceedings of the Sixth Conference on Applied Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 224–231, April 2000.
[26] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++". In Proceedings of ACL 2018, System Demonstrations, pp. 116–121, Melbourne, Australia. Association for Computational Linguistics, 2018.
[27] A. Ali, S. Renals, "Word Error Rate Estimation for Speech Recognition: e-WER", In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol.2: Short Papers, pp. 20-24, Melbourne, Australia, July, 2018.
[28] CHRF^{++}: http://www.statmt.org/wmt17/pdf/WMT70.pdf
[29] 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, Lisboa, Portugal, September 17-18, 2015.
[30] K. Papineni, S. Roukos, T. Ward, W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center, 2001.
[31] A. Araabi, C. Monz, "Optimizing Transformer for Low-Resource Neural Machine Translation". In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3429-3435, Barcelona, Spain (Online), January 2020.

Thura Aung is a member of Language Understanding Lab., Myanmar. He is currently studying B.Eng. in Software Engineering at the Faculty of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand. He is interested in the research areas of Artificial Intelligence (AI), Natural Language Processing (NLP), and Software Engineering.

Ye Kyaw Thu is a Visiting Professor of Language & Semantic Technology Research Team (LST), Artificial Intelligence Research Unit (AINRU), National Electronic & Computer Technology Center (NECTEC), Thailand and Affiliate Professor at Cambodia Academy of Digital Technology (CADT), Cambodia. He is also a founder of Language Understanding Lab., Myanmar. His research lie in the fields of artificial intelligence (AI), natural language processing (NLP) and human-computer interaction (HCI). He is actively co-supervising/supervising undergrad, masters' and doctoral students of several universities including Assumption University (AU), Kasetsart University (KU), King Mongkut's Institute of Technology Ladkrabang (KMITL) and Sirindhorn International Institute of Technology (SIIT).

Zar Zar Hlaing is a member of the Language Understanding Lab in Myanmar. She is currently working as a Machine Learning and NLP Engineer. She earned her Ph.D. in Information Technology from the School of Information Technology at King Mongkut's Institute of Technology Ladkrabang (KMITL) in Bangkok, Thailand. She holds a B.C.Sc. and a B.C.Sc. (Hons) in computer science from the University of Computer Studies in Monywa, as well as an M.C.Sc. in computer science from the University of Computer Studies in Mandalay. Her research interests include Artificial Intelligence (AI), Natural Language Processing (NLP), Language Acquisition, and Text Analysis.