Statistical Machine Translation System Combinations on Phrase-based, Hierarchical Phrase-based and Operation Sequence Model for Burmese and Pa'O Language Pair

Hay Man Htun, Ye Kyaw Thu, Hlaing Myat Nwe, May Thu Win, and Naw Naw

***Abstract*— This paper contributes to the first evaluation of the quality of machine translation between the Myanmar language (Burmese) and Pa'O. We also developed a Burmese-Pa'O parallel corpus (around 18k sentences) based on the Burmese language of the ASEAN MT corpus. The experiments were carried out using three different statistical machine translation approaches: Phrase-based Statistical Machine Translation (PBSMT), Hierarchical Phrase-based Statistical Machine Translation (HPBSMT), Operation Sequence Model (OSM), and making system combination of these three approaches. In addition, three types of segmentation were studied: word segmentation, syllable segmentation, and syllable-SentencePiece. The results show that the OSM approach achieves the highest BLEU, RIBES, and chrF^{++} scores among the three approaches. Our experiments showed that the results of the system combinations of three approaches can achieve significant improvements over each baseline. We also found that syllable segmentation and syllable-SentencePiece are appropriate for translation quality compared with word-level segmentation results.**

***Index Terms*—Myanmar Language (Burmese), Pa'O Language, Statistical Machine Translation (PBSMT, HPBSMT, OSM), Different Word Segmentation Schemes, System Combination**

I. Introduction

THERE are approximately a hundred languages spoken in Myanmar [1]. Burmese is the official language of Myanmar. It is realized as one of the Tibeto Burman groups. A population of 36 million speaks as a mother language. The other languages, dialects, and varieties are spoken by ethnic minorities in Myanmar. In terms of words and writing, ethnic languages can be roughly divided into three categories. They are languages that use their own scripts (e.g Kayah or Red Karen, Shan or Tai), the language using English alphabet or Roman character (e.g Kachin, Rawang), and languages which use Burmese words (e.g Pa'O, Rakhine or Arakanese, Dawei or Tavoyan, Myeik or Beik).

In the current Myanmar education system, the whole country uses Burmese in its curriculum. The people in Myanmar also use Burmese as an official language. From 2014-15, the new education law provides for the teaching of ethnic languages in primary schools. There are still many difficulties, such as the lack of dictionaries for ethnic languages and the lack of digitization. According to the 2014-15 education law, ethnic languages are being introduced through two education channels: teaching them as subjects, 3 to 5 periods every week, and using them orally, as "classroom languages" to explain, the Burmese national curriculum [2]. NLP R&D work is also rare in ethnic languages. Some of the work related to ethnic languages like Kachin, Rawang [3], Rakhine, Dawei, Myeik [4], Kayah [5], Chin [6] and Shan [7] have been done this before. Pa'O is also an ethnic language spoken in Myanmar. The Pa'O language is the native language of Pa'O people, which is used as a primary means of communication for Pa'O people in Myanmar. There are Myanmar people who cannot speak or write Pa'O language, so they have problems in communicating and knowledge sharing with Pa'O people. To overcome the language barrier of communication, we propose a machine translation system between Burmese and Pa'O.

In this paper, we conducted experiments on Statistical Machine Translation (PBSMT, HPBSMT, and OSM), different word segmentation schemes, and system combinations. The state-of-the-art (SoTA) machine translation results between Burmese and Pa'O will be reported.

II. Pa'O Language

Pa'O (also spell Pa-O, Pa-Oh) is a Central Karenic language spoken by half a million Pa'O people in Myanmar [8]. It is also the family of the Tibeto Burman Language. The Pa'O people live mostly in Shan State, Kayin State, Kayah State, Mon State, Bago Division, and Mae Hong Son Province, in northern Thailand. Pa'O people are the seventh-largest ethnic nationality in Myanmar. The Pa'O languages are written using the Burmese script and the same alphabet with the Burmese.

The Pa'O languages mainly use a system of phonetics. In the Pa'O alphabet, "◌့" as "Mine Ngar" and "◌း" as "Mine Paat Ngar" make the original pronunciation a little shorter and longer, giving a special meaning. Moreover, the medial "◌ွ" as "Athat", "◌ျ" as "Yapint"

Hay Man Htun, Hlaing Myat Nwe, May Thu Win, and Naw Naw are with the Faculty of Information Science, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar.

Ye Kyaw Thu is with the National Electronics and Computer Technology Center (NECTEC), Thailand.

Corresponding Authors: haymanhtun@utycc.edu.mm and yktnlp@gmail.com

Manuscript received July 16, 2021; accepted October 13, 2021; revised October 20,2021; published online October 31, 2021.

and "ြ" as "Layit or Rayit" pronunciations have different pronunciations in some places. In the Pa'O script, the medial "ြ" in "က" as "Ka", "ပ" as "Pa" and "ဗ" as "Ba" alphabets has a "Layit" pronunciation and the medial "ြ" in "ခ" as "Kha" and "ဖ" as "Pha" alphabets has a "Rayit" pronunciation. Thus, for example, in the Pa'O script, "ကြ" is pronounced as "Kla" and "ခြ" is pronounced as "Khra". Similarly, "ပြ" is pronounced as "Pla" and "ဖြ" is pronounced as "Phra". "ြ" as "Layit" or "Rayit" sounds like "ကြ" as "Kla", "ပြ" as "Pla", "ခြ" as "Khra" and "ဖြ" as "Phra" are the most common sounds and alphabets in the Pa'O language and script. Compared to the Burmese language, the speech of the Pa'O language is likely to be closer to the written form. Although the Pa'O alphabets are similar to the Burmese alphabets, some alphabets have different pronunciations. The alphabets "ရ", "သ" are pronounced as "Ya", "Tha" in the Burmese script, but they are pronounced as "Ra", "Sa" in the Pa'O script.

In summary, there are grammatical differences, and the most significant differences between Pa'O and Burmese languages are in their pronunciations and their vocabularies. The basic components of Pa'O language are 33 consonants, 8 independent vowels, 3 medial diacritics, 16 dependent vowels, and 3 tones. The word order of Pa'O sentence is Subject-Verb-Object (SVO), which has the same order as the English language. Some example parallel sentences of Burmese (bm) and Pa'O (po) are as follows:

bm: သူ သစ်ပင် တွေ ပန်းပင် တွေ စိုက်တယ် ။
po: ဝွဲ့ ဆို့ သေင်းမွၣ်း ဖုံး ကဝ်းမွၣ်း ဖုံး ။
English: He planted trees and flowers.

bm: မင်း အကြိုက်ဆုံး ရာသီဥတု က ဘာလဲ ။
po: နာ့ အကျိုက်သွတ် ရားသီဥတု နဝ့် တမဲ့်ဟောင်း ။
English: What is your favorite weather?

bm: ခင်ဗျား ပြောခဲ့ သလို ကျွန်တော် ရှင်းပြ ခဲ့တယ် ။
po: နာ့ ကဒေါ့ အတွိုင့် ဓွေ သျင့်ပျ ဗားဒျား ။
English: I explained as you said.

In the above examples, the Burmese and Pa'O words that have same meaning but have different spellings such as "သူ" vs "ဝွဲ့" ("He" in English), "သစ်ပင်တွေ" vs "သေင်းမွၣ်းဖုံး" ("trees"), "ပန်းပင်တွေ" vs "ကဝ်းမွၣ်းဖုံး" ("flowers"), "စိုက်တယ်" vs "ဆို့" ("planted"), "မင်း အကြိုက်ဆုံး" vs "နာ့ အကျိုက်သွတ်" ("your favorite"), "ရာသီဥတု" vs "ရားသီဥတု" ("weather"), "က" vs "နဝ့်" ("is"), "ဘာလဲ" vs "တမဲ့်ဟောင်း" ("What"), "ခင်ဗျား" vs "နာ့" ("you"), "ပြောခဲ့" vs "ကဒေါ့" ("said"), "သလို" vs "အတွိုင့်" ("as"), "ကျွန်တော်" vs "ဓွေ" ("I") and "ရှင်းပြ ခဲ့တယ်" vs "သျင့်ပျဗားဒျား" ("explained").

III. Methodology

In the methodology section, we describe the methodologies of the statistical machine translation used in the experiments for this paper.

A. Phrase-Based Statistical Machine Translation (PBSMT)

A PBSMT translation model is based on phrasal units [9]. The phrase translation model is based on the noisy channel model. To find best translation $\hat{t}$ that maximizes the translation probability $\mathbf{P}(t|s)$ given the source sentences; mathematically. Here, the source language is Burmese and the target language is Pa'O. The translation of a Burmese sentence s into a Pa'O sentence t is modeled as equation 1.

$$\hat{t} = argmax_t \mathbf{P}(t|s) = argmax_t \mathbf{P}(s|t)\mathbf{P}(t) \quad (1)$$

B. Hierarchical Phrase-Based Statistical Machine Translation (HPBSMT)

A hierarchical model differs from a classical PBSMT model in terms of rule expressivity [10]. The synchronous context-free grammars (SCFG) rules are allowed to contain one or more non-terminals. Each terminal acting as a variable can be expanded into other expressions using the SCFG grammar. An example of hierarchical phrase-based grammar rules between Burmese and Pa'O from an HPBSMT model is as follows:

[X][X] ဘယ်သူ လဲ [X] ||| [X][X] ပါ့မဲ့်ဲ့ [X]
[X][X] ဘယ်သူ လဲ [X] ||| [X][X] ပါ့မဲ့်ဲ့ ဖုံးဟောင်း [X]
[X][X] ဘယ်သူ လဲ [X] ||| [X][X] ပါ့မဲ့်ဲ့ ဟောင်း [X]
[X][X] ဘယ်သူ လဲ [X] ||| [X][X] ပါ့မဲ့်ဲ့ အီ [X]
[X][X] ဘယ်သူ လဲ [X] ||| [X][X] ပါ့မဲ့်ဲ့ဟောင်း [X]

Each line in the example of hierarchical phrase-based grammar rules describes one translation rule. It consists of two components separated by three bars (|||): the source string (Burmese) and three variables ([X]) in the source side, the target string (Pa'O) and three variables ([X]) in the target side. Here, the Burmese word "ဘယ်သူ လဲ" means "who?". HPBSMT approach is particularly applicable to language pairs that require long-distance re-ordering during the translation process [11].

C. Operation Sequence Model (OSM)

OSM is an N-gram-based translation and reordering model that represents aligned bilingual corpus as a sequence of operations and learns a Markov model over the resultant sequences [12]. The operation types are (i) generate (generation of a sequence of source and target words), (ii) insert gap (insertion of gaps as explicit target positions of reordering operations), and (iii) jump (forward and backward jump operations which perform the actual reordering) [13]. The probability of a sequence of operations is defined according to an N-gram model, i.e., the

probability of an operation depends on the n-1 preceding operations. Let O = o_1, … ,o_N be a sequence of operations as hypothesized by the translator to generate a word-aligned bilingual sentence pair < F;E;A >; the model is then defined as equation 2.

$$\mathbf{P}_{osm}(F,E,A) = \mathbf{p}(o_1,...,o_N) = \prod_i \mathbf{p}(o_i|o_{i-n+1}...o_{i-1}) \tag{2}$$

where n indicates the amount of context used, F defines the source sentences, E defines the target sentences and A defines the word-alignment function between E and F. The following shows an example translation process of Pa'O sentence "ဝွဲမူး ကျိုက်ဒျား ကဝ်းရား" into Burmese "သူမ ပန်းသီး ကြိုက်တယ်" ("She likes apples" in English) with the OSM.

Source: ဝွဲမူး ကျိုက်ဒျား ကဝ်းရား
Target: သူမ ပန်းသီး ကြိုက်တယ်

Operation 1: Generate (ဝွဲမူး, သူမ)
Operation 2: Insert Gap
Operation 3: Generate (ကဝ်းရား, ပန်းသီး)
Operation 4: Jump Back (1)
Operation 5: Generate (ကျိုက်ဒျား, ကြိုက်တယ်)

In this example, the Pa'O word "ဝွဲမူး" and the Burmese word "သူမ" mean "she","ကျိုက်ဒျား" and "ကြိုက်တယ်" mean "likes", and "ကဝ်းရား" and "ပန်းသီး" mean "apples".

IV. Experiments

A. Burmese-Pa'O Parallel Corpus

We used 18,354 Burmese sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [14], which is a parallel corpus in the travel domain. It contains six main categories and they are people (greeting, introduction, and communication), survival (transportation, accommodation, and finance), food (food, beverage, and restaurant), fun (recreation, traveling, shopping and nightlife), resource (number, time and accuracy), and special needs (emergency and health). Manual translation into Pa'O language was done by native Pa'O monks from Taunggyi, Shan State, Myanmar, and native Pa'O students from Myanmar universities. Word segmentation for Pa'O was done manually and there are exactly 82,782 words in total. We used 14,000 sentences for training, 2,500 sentences for the development or tuning process, and 1,854 sentences for evaluation respectively.

B. Word Segmentation

There is no space in both Burmese and Pa'O text but spaces are used to separate phrases. There are no clear rules for using spaces, and thus spaces may (or may not) be inserted between words, phrases, and even between root words and their affixes. Although Burmese sentences of the ASEAN-MT corpus are already segmented, we have to consider some rules for manual word segmentation of Pa'O sentences. We defined Pa'O "word" to be meaningful units and affixes, prefixes, root word, and suffixes are separated such as "အံး ဒျား", "အံး ထွျလဲဉ်း", "အီး အံး ဒျား". Here, "အံး" ("eat" in English) is a root word and the others are prefixes and suffixes for past and future tenses. Similar to Burmese, Pa'O plural nouns are identified by the following plural form of the noun. We also put a space between the noun and the following plural form of the noun, for example, a Pa'O word "မူးပေး ဖုံး" ("girls") is segmented as two words "မူးပေး" ("girl") and the plural form of the noun "ဖုံး". In Pa'O grammar, particles describe the type of noun and are used after the number or text number. For example, a Pa'O word "အယွဲးနိဗီး" ("two coins") is segmented as three words "အယွဲး" ("coin"), text number "နိ" ("two") and a particle "ဗီး". In our manual word segmentation rules, compound nouns are considered as one word and thus, a Pa'O compound word "ရွဉ်" + "အိတ်" ("money" + "bag") is written as one word "ရွဉ်အိတ်" ("wallet"). Pa'O adverb words such as "ဖွး" ("really"), "မွိုင်မွိုင်" ("quickly") and "ဖွးမဲ့ဲင်း" ("very") are also considered as one word. The following is an example of word segmentation for a Pa'O sentence in our corpus and the meaning is "She tells a story to her children."

Unsegmented sentence:
po: ဝွဲမူးခြုင်းနယ်လိုပေးဖုံးငဝ်းခြုင်းတပုဒ်။

Segmented sentence:
po: ဝွဲမူး ("She") ခြုင်းနယ် ("tells") လိုပေး ဖုံး ("children") ငဝ်းခြုင်း တ ပုဒ် ("a story") ။ (".")

In this example, "လိုပေးဖုံး" ("children") is a plural noun of "လိုပေး" ("child") and a plural form of the noun "ဖုံး" are segmented as two words. Three Pa'O words, "ငဝ်းခြုင်း" ("story"), text number "တ" and a particle "ပုဒ်" are segmented as three words.

C. Syllable Segmentation

In general, like Burmese, Pa'O words are written with multiple syllables and most of the syllables are written with more than one character. If we only focus on consonant-based syllables, the structure of the syllable can be described with Backus normal form (BNF) as follows:

Syllable := CMV[CK][D]

Here, it defines the consonants as C, medials as M, vowel as V, vowel killer character as K, and diacritic characters as D. Burmese syllable segmentation can be done with a rule-based approach, finite-state automaton (FSA), or regular expressions (RE) (https://github.com/ye-kyaw-thu/sylbreak). In Pa'O, the characters used are almost identical to Burmese as we mentioned in Section II. Therefore, in our experiment, Pa'O syllable segmentation can be done with an RE-based Burmese

syllable segmentation tool named “sylbreak”. Syllable segmentation can solve the problem of the Out Of Vocabulary (OOV) in SMT between Burmese and Pa’O language pairs. The following is an example of syllable segmentation for a Pa’O sentence in our corpus and the meaning is “She tells a story to her children.”

Syllable segmented Pa’O sentence:
po: ဇဝ့ မူး ခြဲင်း နယ် လို့ ပေး ဖုံး ဝဝ်း ခြဲင်း တ ပ့ဒ်။

There is no ambiguity in word or syllable segmentation in both Burmese and Pa’O.

D. SentencePiece

SentencePiece is a language-independent subword tokenizer and detokenizer designed for Neural-based text processing, including NMT [15]. It provides open-source C++ and Python implementations for subword units. While existing subword segmentation tools assume that the input is pre-tokenized into word sequences, SentencePiece can train subword models directly from raw sentences, which allows us to make a purely end-to-end and language-independent system. The following is an example of syllable-SentencePiece segmentation for a Pa’O sentence in our corpus and the meaning is “She tells a story to her children.”

SentencePiece segmented Pa’O sentence:
po: ▁ဇဝ့ ▁မူး ▁ခြဲင် း ▁နယ် ▁လို့ ▁ပေး ▁ဖုံး ▁ဝဝ်း ▁ခြဲင် း ▁တ ▁ပ ့ ဒ် ▁ ။

E. Moses SMT System

We used Moses toolkit [16] for training the PBSMT, HPBSMT, and OSM statistical machine translation systems. The standard word alignment toolkit GIZA++ [17] aligned the word segmented source language and the word segmented target language. The alignment was symmetrized by grow-diag-final and heuristic [9]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [18]. The language modeling toolkit KenLM [19] was used for training the 5-gram language model with modified Kneser-Ney discounting [20]. Using a 5-gram language model is based on the previous work of various machine translation experiments between other ethnic languages (Rakhine, Dawei, Myeik) and Burmese [4]. Minimum error rate training (MERT) [21] was used to tune the decoder parameters. For decoding, Moses decoder (version 2.1.1) [16] has been used. We used the default settings of Moses for all experiments.

F. SMT System Combinations

We made SMT system combinations on PBSMT, HPBSMT, and OSM for Burmese and Pa’O language pairs for better results. In system combination, the translated sentences with the greater RIBES score among the translated sentences from each SMT approach are chosen as the final translated outputs. System combinations are performed in four ways based on three SMT methods. The four ways are the combination of PBSMT and HPBSMT (PBSMT+HPBSMT), the combination of PBSMT and OSM (PBSMT+OSM), the combination of HPBSMT and OSM (HPBSMT+OSM), and the combination of PBSMT, HPBSMT, and OSM (PBSMT+HPBSMT+OSM). The following is a shell program (system-combination.sh) for the combination of two SMT systems.

```
#!/bin/bash

# How to use: ./system-combination.sh <model1-hypothesis> <model2-hypothesis> <reference>
# e.g. ./system-combination.sh ./fm_hyp.txt ./sm_hyp.txt ./ref.txt

fm=$1; sm=$2; ref=$3;
i=0; j=0; k=0;

while read line
do
        fm_arr[$i]="$line";
        i=$((i+1));
done < "$fm"

while read line
do
        sm_arr[$j]="$line";
        j=$((j+1));
done < "$sm"

while read line
do
        ref_arr[$k]="$line";
        k=$((k+1));
done < "$ref"

len=${#fm_arr[@]};

for (( i=0; i<$len; i++ ));
do
 echo "" > fm_hyp.txt;
 echo "" > sm_hyp.txt;
 echo "" > ref.txt;

 echo "${fm_arr[$i]}" > fm_hyp.txt;
 echo "${sm_arr[$i]}" > sm_hyp.txt;
 echo "${ref_arr[$i]}" > ref.txt;

#Evaluation with RIBES scores
fm_rs=`python ./RIBES-1.03.1/RIBES.py -r ref.txt fm_hyp.txt`;
sm_rs=`python ./RIBES-1.03.1/RIBES.py -r ref.txt sm_hyp.txt`;

if [[ "$fm_rs" > "$sm_rs" ]]; then
   echo "${fm_arr[$i]}" >> rs.txt;
else
  echo "${sm_arr[$i]}" >> rs.txt;
fi

done
```

Listing 1: system-combination.sh for combination of two SMT systems

In the combination of all three SMT systems, firstly, the first two SMT systems combination was made and then, we combined the results from the combination of the first

two SMT systems and the third SMT system for Burmese and Pa'O language pairs using the above shell program (see Listing 1).

V. Evaluation

We used three automatic criteria for the evaluation of the machine-translation output. They are the standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [22], the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [23] and Character n-gram F-score (chrF^{++}) [24]. The BLEU score measures the precision of n-gram (overall $n \leq 4$ in our case) concerning a reference translation with a penalty for short translations. RIBES is based on rank correlation coefficients modified with precision and special care is paid to the word order of the translation results. The RIBES score is suitable for distance language pairs such as Burmese and English. chrF^{++} is represented as a very promising evaluation metric for machine translation, especially for morphologically rich target languages. Large BLEU, RIBES, and chrF^{++} scores are better.

VI. Results and Discussion

In this research work, PBSMT, HPBSMT, and OSM were performed with three-word segmentation schemes (word, syllable, and syllable-SentencePiece). The results of PBSMT, HPBSMT, and OSM with word segmentation are low. For example, the top BLEU scores are 7.84 for Burmese to Pa'O and 13.80 for Pa'O to Burmese. This is because the Burmese-Pa'O parallel corpus contains some Pa'O phrases and sentences that require to make word segmentation. In addition, Burmese and Pa'O also have grammatical differences. Therefore, in this paper, we present only the results with syllable segmentation and syllable-SentencePiece segmentation.

The BLEU, RIBES, and chrF^{++} score results for Burmese and Pa'O machine translation experiments for PBSMT, HPBSMT, OSM, and system combinations are shown in Table I, II, III, and IV. Bold numbers indicate the highest scores among three SMT approaches and system combinations. Here, "bm" stands for Burmese and "po" stands for Pa'O respectively.

From the SMT with syllable segmentation results (Table I), OSM is the best BLEU and chrF^{++} score for both Burmese-Pa'O and Pa'O-Burmese. Although the OSM approach achieved the highest score for Burmese-Pa'O, the HPBSMT achieved the highest score for Pa'O-Burmese in terms of RIBES score.

The results of BLEU, RIBES, and chrF^{++} scores of syllable-SentencePiece segmentation between Burmese and Pa'O are shown in Table II. From the results, the OSM method achieved the highest BLEU score for both Burmese-Pa'O and Pa'O-Burmese machine translations. Compared to syllable segmentation results in Table I, the BLEU scores of PBSMT and HPBSMT for Burmese-Pa'O with syllable-SentencePiece were higher, but the BLEU score of OSM with syllable is higher than with syllable-SentencePiece. The BLEU scores for Pa'O-Burmese and the chrF^{++} scores for both Burmese-Pa'O and Pa'O-Burmese of three SMT approaches with syllable-SentencePiece are more effective. As for the RIBES scores, the results with syllable are better than with syllable-SentencePiece in all three SMT approaches.

Table III shows BLEU, RIBES, and chrF^{++} score results for system combinations of three SMT approaches between Burmese and Pa'O with syllable segmentation. From the Table III syllable-based results, PBSMT+HPBSMT+OSM achieved the highest BLEU, RIBES, and chrF^{++} score among system combinations for both Burmese-Pa'O and Pa'O-Burmese machine translations.

Table IV presents BLEU, RIBES, and chrF^{++} score results for system combinations of three SMT approaches between Burmese and Pa'O with syllable-SentencePiece segmentation. According to Table IV results, PBSMT+HPBSMT+OSM also achieved the highest BLEU, RIBES, and chrF^{++} score among system combinations for both Burmese-Pa'O and Pa'O-Burmese machine translations.

Comparison of the results presented in Table III and Table IV, PBSMT+HPBSMT+OSM achieved the highest BLEU (38.46 for Burmese-Pa'O and 50.03 for Pa'O-Burmese), RIBES (0.868276 for Burmese-Pa'O and 0.902760 for Pa'O-Burmese), and chrF^{++} (77.9013 for Burmese-Pa'O and 79.9984 for Pa'O-Burmese) scores.

Finally, from the results shown in Table I, II, III, and IV, it can be seen that the results of system combinations for three SMT approaches have significant improvements than individual approach. Interestingly, the BLEU, RIBES, and chrF^{++} scores of all three methods achieved the highest performance. Our results with the current parallel corpus indicate that Pa'O to Burmese machine translation is better performance than Burmese to Pa'O translation direction. This is because the problem of Out Of Vocabulary (OOV) is less common in the Pa'O to Burmese machine translation than Burmese to Pa'O machine translation.

VII. Error Analysis

Word Error Rate (WER) of all experiments are calculated by using the SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 [25]. The formula for WER can be stated as Equation 3:

$$WER = \frac{(N_i + N_d + N_s) \times 100}{N_d + N_s + N_c} \tag{3}$$

where N_i is the number of insertions; N_d is the number of deletions, N_s is the number of substitutions; N_c is the number of correct words. Note that if the number of insertions is very high, the WER can be greater than 100%. The SCLITE program printout confusion pairs and Levenshtein distance calculations for all hypothesis sentences in detail.

TABLE I: BLEU, RIBES and chrF^{++} scores for PBSMT, HPBSMT and OSM using syllable segmentation

	BLEU		**RIBES**		**chrF^{++}**	
	bm-po	**po-bm**	**bm-po**	**po-bm**	**bm-po**	**po-bm**
PBSMT	33.04	41.20	0.822720	0.863816	c6+w2-F2: 74.1484 c6+w2-avgF2: 74.5963	c6+w2-F2: 74.7827 c6+w2-avgF2: 75.5417
HPBSMT	33.23	41.70	0.816880	**0.868130**	c6+w2-F2: 74.1182 c6+w2-avgF2: 74.5990	c6+w2-F2: 75.4671 c6+w2-avgF2: 76.1924
OSM	**35.06**	**43.45**	**0.828840**	0.868051	c6+w2-F2: **74.7709** c6+w2-avgF2: 75.2073	c6+w2-F2: **75.8797** c6+w2-avgF2: 76.6384

TABLE II: BLEU, RIBES and chrF^{++} scores for PBSMT, HPBSMT and OSM using syllable-SentencePiece segmentation

	BLEU		**RIBES**		**chrF^{++}**	
	bm-po	**po-bm**	**bm-po**	**po-bm**	**bm-po**	**po-bm**
PBSMT	34.46	43.58	0.816151	0.857324	c6+w2-F2: 76.2035 c6+w2-avgF2: 76.5766	c6+w2-F2: 77.7346 c6+w2-avgF2: 78.4389
HPBSMT	33.29	44.73	0.811699	**0.866729**	c6+w2-F2: 76.6195 c6+w2-avgF2: 77.2141	c6+w2-F2: 78.4436 c6+w2-avgF2: 79.2002
OSM	**34.48**	**45.86**	**0.821221**	0.865667	c6+w2-F2: **76.9888** c6+w2-avgF2: 77.4713	c6+w2-F2: **78.4552** c6+w2-avgF2: 79.1778

TABLE III: BLEU, RIBES and chrF^{++} scores for system combination using syllable segmentation

	BLEU		**RIBES**		**chrF^{++}**	
	bm-po	**po-bm**	**bm-po**	**po-bm**	**bm-po**	**po-bm**
PBSMT+HPBSMT	35.55	44.83	0.853441	0.888488	c6+w2-F2: 74.9682 c6+w2-avgF2: 75.5420	c6+w2-F2: 76.2639 c6+w2-avgF2: 77.1372
PBSMT+OSM	37.00	45.56	0.851019	0.881752	c6+w2-F2: 75.4682 c6+w2-avgF2: 75.9877	c6+w2-F2: 76.4553 c6+w2-avgF2: 77.2992
HPBSMT+OSM	37.54	46.77	0.860096	0.891996	c6+w2-F2: 75.4529 c6+w2-avgF2: 76.0592	c6+w2-F2: 77.0910 c6+w2-avgF2: 77.9927
PBSMT+HPBSMT+OSM	**38.45**	**47.57**	**0.868276**	**0.896817**	c6+w2-F2: **75.8660** c6+w2-avgF2: 76.4931	c6+w2-F2: **77.2953** c6+w2-avgF2: 78.2121

TABLE IV: BLEU, RIBES and chrF^{++} scores for system combination using syllable-SentencePiece segmentation

	BLEU		**RIBES**		**chrF^{++}**	
	bm-po	**po-bm**	**bm-po**	**po-bm**	**bm-po**	**po-bm**
PBSMT+HPBSMT	36.78	47.40	0.845609	0.891085	c6+w2-F2: 77.2030 c6+w2-avgF2: 77.8438	c6+w2-F2: 79.1240 c6+w2-avgF2: 79.9693
PBSMT+OSM	36.92	47.60	0.840138	0.881706	c6+w2-F2: 77.2778 c6+w2-avgF2: 77.7723	c6+w2-F2: 79.0864 c6+w2-avgF2: 79.9170
HPBSMT+OSM	37.16	49.29	0.849800	0.896932	c6+w2-F2: 77.6732 c6+w2-avgF2: 78.3184	c6+w2-F2: 79.7598 c6+w2-avgF2: 80.6516
PBSMT+HPBSMT+OSM	**38.46**	**50.03**	**0.857091**	**0.902760**	c6+w2-F2: **77.9013** c6+w2-avgF2: 78.5511	c6+w2-F2: **79.9984** c6+w2-avgF2: 80.9252

TABLE V: *WER*% for PBSMT, HPBSMT, OSM and system combination using syllable segmentation with nearly 1,800 sentences test data (lower is better)

src-tgt	**PBSMT**	**HPBSMT**	**OSM**	**PBSMT+ HPBSMT**	**PBSMT +OSM**	**HPBSMT +OSM**	**PBSMT+ HPBSMT +OSM**
bm-po	51.0%	51.7%	49.8%	47.3%	46.8%	46.1%	**44.6%**
po-bm	44.8%	43.9%	43.0%	40.1%	40.5%	38.7%	**37.7%**

TABLE VI: *WER*% for PBSMT, HPBSMT, OSM and system combination using syllable-SentencePiece segmentation with nearly 1,800 sentences test data (lower is better)

src-tgt	**PBSMT**	**HPBSMT**	**OSM**	**PBSMT+ HPBSMT**	**PBSMT +OSM**	**HPBSMT +OSM**	**PBSMT+ HPBSMT +OSM**
bm-po	54.3%	56.1%	53.8%	51.2%	51.2%	50.5%	**49.1%**
po-bm	47.4%	45.6%	44.4%	41.5%	42.3%	39.6%	**38.6%**

TABLE VII: The top 10 confusion pairs of OSM model for Burmese-Pa'O machine translation with word segmentation

Freq	Confusion Pair (REF ==> HYP)
26	တဝ်းဟောင်း ==> ဟောင်း
23	နေ့ဟောင်း ==> ဟောင်း
16	တမွေးတဝ်းဟောင်း ==> မွေးတဝ်းဟောင်း
15	ဒျား ==> လွၚ်မ်
13	နော ==> ကရိ
12	ကို ==> နဝ့်
12	ဒျား ==> ဟောင်း
12	ယို ==> ယိုနဝ့်
10	ဗားတဝ်း ==> တဝ်း
10	ဝွဲ့နဝ့် ==> ဝွဲ့

For example, scoring I, D and S for the translated Pa'O sentence "ဝွဲ့မူး အီးကူးဒီး ပါ့မဲ့ဉ် ဟောင်း" ("Who will she help?" in English, "သူမ ဘယ်သူ့ ကို ကူညီ မလဲ" in Burmese) compare to a reference sentence, the output of the SCLITE program is as follows:

Scores: (#C #S #D #I) 3 1 1 0
REF: ဝွဲ့မူး အီး ကူးဒီး ပါ့မဲ့ဉ် ဟောင်း
HYP: ဝွဲ့မူး ********* အီးကူးဒီး ပါ့မဲ့ဉ် ဟောင်း
Eval: D S

In this case, one deletion (အီး => ***) and one substitution (ကူးဒီး => အီးကူးဒီး) happened, that is $S = 1$, $D = 1$, $I = 0$, $C = 3$ and thus WER is equal to 40%.

The WER% of PBSMT, HPBSMT, OSM, and three SMT models combination with syllable and syllable-SentencePiece based between Burmese and Pa'O machine translation with around 1,800 test sentences are as shown in Table V and Table VI. Bold numbers indicate the lowest WER among three SMT approaches and system combinations of these three SMT approaches. Here, "src" stands for source language and "tgt" stands for target language respectively.

From Table V, we found that the lowest WER% are 44.6% for Burmese to Pa'O and 37.7% for Pa'O to Burmese machine translations using syllable segmentation with the combination of all three SMT approaches. In Table VI, we also found that the lowest WER% are 49.1% for Burmese to Pa'O and 38.6% for Pa'O to Burmese machine translations using syllable-SentencePiece segmentation with the combination of all three SMT approaches. These results are inversely proportional to the BLEU scores.

Some confusion pairs relating to word segmentation errors were found after we analyzed confusion pairs of each model in detail. The top 10 confusion pairs of OSM model for Burmese-Pa'O machine translation with word segmentation are shown in Table VII. Here, the confusion pairs of "တဝ်းဟောင်း ==> ဟောင်း", "နေ့ဟောင်း ==> ဟောင်း", "တမွေးတဝ်းဟောင်း ==> မွေးတဝ်းဟောင်း", "ယို ==> ယိုနဝ့်", "ဗားတဝ်း ==> တဝ်း" and "ဝွဲ့နဝ့် ==> ဝွဲ့" are happened because of word segmentation errors. These kinds of confusion pairs can be reduced by cleaning of current word segmentation of our parallel corpus.

VIII. Related Work

There are some researches for the SMT of Burmese. Ye Kyaw Thu et al. (2016) [26] presented the first large-scale study of the translation of the Burmese. In the study, 40 kinds of language pairs were used and the languages consist of not only similarities but also fundamentally differences from Burmese. The results showed that the HPBSMT approach gave the highest translation quality in terms of both the BLEU and RIBES scores. Moreover, we learned that the OSM approach gave the highest translation performance translation between Khmer (the official language of Cambodia) and twenty other languages, in both directions [27].

Relating to Myanmar (Burmese) language dialects, Thazin Myint Oo et al. (2018) [28] contributed the first PBSMT, HPBSMT, and OSM machine translation evaluations between Burmese and Rakhine. The experiment was used the 18K Burmese-Rakhine parallel corpus that constructed to analyze the behavior of a dialectal Burmese-Rakhine machine translation. The results showed that higher BLEU (57.88 for Burmese-Rakhine and 60.86 for Rakhine-Burmese) and RIBES (0.9085 for Burmese-Rakhine and 0.9239 for Rakhine-Burmese) scores can be achieved for the Rakhine-Burmese language pair even with the limited data. Thazin Myint Oo et al. (2019) also contributed the first SMT evaluations between the Burmese and Dawei (Tavoyan) language pair. The SMT results with developed 9K Burmese-Dawei parallel corpus showed that higher BLEU (21.70 for Burmese-Dawei and 29.56 for Dawei-Burmese) and RIBES (0.78 for Burmese-Dawei and 0.82 for Dawei-Burmese) scores were achieved with the OSM approach [29]. Thazin Myint Oo et al. (2020) [30] further distributed the first SMT evaluations between Myanmar (Burmese) and Myeik (Beik). The results with 10K Burmese-Myeik parallel corpus demonstrated that the OSM method showed the highest BLEU (44.33 for Burmese-Myeik and 33.41 for Myeik-Burmese) and RIBES (0.87531 for Burmese-Myeik and 0.83991 for Myeik-Burmese) scores for both Burmese-Myeik and Myeik-Burmese machine translations in evaluations with syllable unit.

In terms of word segmentation, SMT between Burmese and Pa'O results are slightly lower than SMT between Burmese and other ethnic languages mentioned earlier. The reason is that some Burmese words can translate more than one Pa'O word. Therefore, there are many Pa'O words in the Burmese-Pa'O parallel corpus that have the same meaning and different words. Burmese and Pa'O have different grammatical orders. Although the Pa'O script uses the Burmese alphabet, the writing structure is quite different.

IX. Conclusion

In this paper, we contribute the very first PBSMT, HPBSMT, and OSM machine translation evaluations from Burmese to Pa'O and Pa'O to Burmese. We used the 18K Burmese-Pa'O parallel corpus that we constructed to analyze the behavior of Burmese-Pa'O machine translation with three segmentation units (word, syllable, and syllable-SentencePiece). The result gets better translation result in syllable-SentencePiece translation unit than word and syllable level. Moreover, the system combination results of all three statistical translation models with syllable-SentencePiece segmentation unit are the highest among the experimental results. From the results, OSM methods achieved better scores for machine translation between Burmese and Pa'O. We showed that higher BLEU, RIBES, and chrF^{++} scores can be achieved for the Pa'O-Burmese language pair even with the limited data. This paper also presents a detailed analysis of confusion pairs of machine translation between Burmese-Pa'O and Pa'O-Burmese. In the future, corpus extension will also be implemented, and other machine translation approaches, such as neural machine translation are being considered.

References

[1] https://en.wikipedia.org/wiki/Languages _of _Myanmar

[2] https://themimu.info/sites/themimu.info/files/documents/Report _Teaching _Ethnic _Minority _Languages _In _Government _Schools1.pdf

[3] Ye Kyaw Thu, Manar Hti Seng, Thazin Myint Oo, Dee Wom, Hpau Myang Thint Nu, Seng Mai, Thepchai Supnithi and Khin Mar Soe, "Statistical Machine Translation between Kachin and Rawang", In Proceedings of the 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2019), Oct 30 to Nov 1, 2019, Chiang Mai, Thailand, pp. 329-334

[4] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, Thepchai Supnithi, "Statistical Machine Translation of Myanmar Dialects", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 14-26.

[5] Zar Zar Linn, Ye Kyaw Thu, Pushpa B. Patil, "Statistical Machine Translation between Myanmar (Burmese) and Kayah Languages", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 62-68.

[6] Hnin Yi Aye, Yuzana Win, Ye Kyaw Thu, "Statistical Machine Translation between Myanmar (Burmese) and Chin (Mizo) Language", In Proceedings of The 23rd Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (Oriental COCOSDA 2020), Nov 5-7 2020, Yangon, Myanmar, pp. 211-216.

[7] Nang Aeindray Kyaw, Ye Kyaw Thu, Hlaing Myat Nwe, Phyu Phyu Tar, Nandar Win Min, Thepchai Supnithi, "A Study of Three Statistical Machine Translation Methods for Myanmar (Burmese) and Shan (Tai Long) Language Pair", In Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020), Nov 18 to Nov 20, 2020, Bangkok, Thailand, pp. 218-223

[8] https://en.wikipedia.org/wiki/Pa'O _language

[9] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation." in Proc. of HTL-NAACL, 2003, pp. 48–54.

[10] Chiang, D., "Hierarchical phrase-based translation", Computational Linguistics 33(2), 2007, pp. 201-228.

[11] Braune, Fabienne and Gojun, Anita and Fraser, Alexander, "Long-distance reordering during the search for hierarchical phrase-based SMT", in Proc. of the 16th Annual Conference of the European Association for Machine Translation, 2012, Trento, Italy, pp. 177-184.

[12] http://www.statmt.org/moses/manual/manual.pdf

[13] Nadir Durrani, Helmut Schmid, Alexander M. Fraser, Philipp Koehn and Hinrich Schutze, "The Operation Sequence Model - Combining N-Gram-Based and Phrase-Based Statistical Machine Translation", Computational Linguistics, Volume 41, No. 2, 2015, pp. 185-214.

[14] Prachya, Boonkwan and Thepchai, Supnithi, "Technical Report for The Network-based ASEAN Language Translation Public Service Project", Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013

[15] Taku Kudo and John Richardson, "Sentence-Piece: A simple and language-independent sub-word tokenizer and detokenizer for Neural Text Processing", EMNLP2018.

[16] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[17] Och Franz Josef and Ney Hermann, "Improved Statistical Alignment Models", in Proc. of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 2000, pp. 440-447.

[18] Tillmann Christoph, "A Unigram Orientation Model for Statistical Machine Translation", in Proc. Of HLT-NAACL 2004: Short Papers, Stroudsburg, PA, USA, 2004, pp. 101-104.

[19] Heafield, Kenneth, "KenLM: Faster and Smaller Language Model Queries", in Proc. of the Sixth Workshop on Statistical Machine Translation, WMT 11, Edinburgh, Scotland, 2011, pp. 187-197.

[20] Chen Stanley F and Goodman Joshua, "An empirical study of smoothing techniques for language modeling", in Proc. of the 34th annual meeting on Association for Computational Linguistics, 1996, pp. 310-318.

[21] Och Franz J., "Minimum error rate training in statistical machine translation", in Proc. of the 41 st Annual Meeting n Association for Computational Linguistics – Volume 1, Association for Computer Linguistics, Sapporo, Japan, July 2003, pp.160-167.

[22] Papineni, K., Roukos, S., Ward, T., Zhu, W., "BLEU: a Method for Automatic Evaluation of Machine Translation", IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center, 2001

[23] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H, "Automatic evaluation of translation quality for distant language pairs", in Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944-952.

[24] M. Popović, "chrF^{++}: words helping character n-grams," Asso. for Comp. Linguistics, Proc. of the Sec. Conf. on Mac. Translation, pp. 612–618, September 2017

[25] (NIST) The National Institute of Standards and Technology. Speech recognition scoring toolkit (sctk), version: 2.4.10, 2015.

[26] Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, "A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language", in Proc. Of SNLP2016, February 10-12, 2016.

[27] Ye Kyaw Thu, Vichet Chea, Andrew Finch, Masao Utiyama and Eiichiro Sumita, "A Large-scale Study of Statistical Machine Translation Methods for Khmer Language", 29th Pacific Asia Conference on Language, Information and Computation, October 30 - November 1, 2015, Shanghai, China, pp. 259-269.

[28] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, "Statistical Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)", In Proceedings of ICCA2018, February 22-23, 2018, Yangon, Myanmar, pp. 304-311

[29] Thazin Myint OO, Ye Kyaw Thu, Khin Mar Soe and Thepchai Supnithi, "Statistical Machine Translation between Myanmar (Burmese) and Dawei (Tavoyan)", The First Workshop on NLP Solutions for Under-Resourced Languages (NSURL 2019), 11-13 September 2019, Trento, Italy

[30] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe and Thepchai Supnithi, "Statistical Machine Translation between Myanmar and Myeik", In Proceedings of the 12th International Conference on Future Computer and Communication (ICFCC 2020), Feb 26-28, 2020, Yangon, Myanmar, pp. 36-45

Hay Man Htun is a candidate of the M.E (Information Science & Technology) degree program at the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. She is also a member of the NLP Research Lab., UTYCC. She holds the degree of Bachelor of Engineering (Information Science & Technology) from the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. Her current master thesis research focuses on machine translation between Burmese and Pa'O language pair. She is strongly interested in the areas of Natural Language Processing (NLP) such as machine translation, Speech Processing, Image Processing, Machine Learning, and Deep Learning.

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 Head of NLP Research Lab., University of Technology Yatanarpon Cyber City (UTYCC), Pyin Oo Lwin, Myanmar. He is also a founder of Language Understanding Lab., Myanmar, and a Visiting Researcher of Language. He is actively co-supervising/supervising undergrad, masters', and doctoral students of several universities including KMITL, SIIT, UCSM, UCSY, UTYCC, and YTU.

Hlaing Myat Nwe is a PhD candidate of Sirindhorn International Institute of Technology, Thammasat University, Thailand. A native of Myanmar, she holds a master degree of Information Science and Technology, and a bachelor degree of Information Science and Technology from the University of Technology (Yatanarpon Cyber City), Myanmar. Her research interests include human-computer interaction, natural language processing, audio signal, and image processing. She has been working to find efficient and user-friendly text input interfaces and video translation system for Myanmar Sign Language and Myanmar SignWriting. She is also a supervising lab member from NLP-Lab, UTYCC.

May Thu Win received her B.C.Tech, B.C.Tech(Hons), M.C.Tech, and Ph.D (IT) from University of Computer Studies (Magway), University of Computer Studies, Mandalay (UCSM), respectively. She worked as a tutor at the University of Computer Studies (Yangon) in 2007 and Computer University (Kalay) in 2010, University of Technology (Yadanapon Cyber City) in 2011 in Myanmar, respectively. She is also very interested in both teaching and research. Therefore, she has worked as a researcher at the University of Miyazaki, Japan in 2016. Moreover, she published local and international papers in the IT fields. She also served as a supervisor for master students and Ph.D Candidates in Myanmar. Currently, she is a lecturer at the University of Technology (Yatanarpon Cyber City) in Myanmar.

Naw Naw is not only a technological teacher but also a researcher in IT and higher educational fields. There are many local and international journals she published in those fields. She collaborates with regional and international corporations to improve higher education programs in Myanmar. She also developed many e-learning contents and distributed for Myanmar students. She is also a work-based learning coordinator who is responsible for linking teachers' classroom trainings, students' skills, and industrial needs.