Statistical Machine Translation of Myanmar Dialects

Thazin Myint Oo, UCSY, Myanmar, Ye Kyaw Thu, NECTEC, Thailand,
Khin Mar Soe, UCSY, Myanmar, and Thepchai Supnithi, NECTEC, Thailand

***Abstract*— The goal of this work is to contribute the first evaluation of the quality of machine translation between Standard Myanmar and Other Myanmar Dialectal Languages. Myanmar Dialects present many challenges for machine translation, which is the lack of data resources. To fulfill this requirement, we also developed three Myanmar Dialect corpora based on the Myanmar language of ASEAN MT corpus. They are Myanmar-Rakhine (18K), Myanmar-Myeik (10K) and Myanmar-Dawei (9K) parallel corpora. The 10 folds cross-validation experiments were carried out using three different statistical machine translation approaches: phrase-based, hierarchical phrase-based, and the operation sequence model. In addition, two types of segmentation; word and syllable units were studied. The results show that all three statistical machine translation approaches give higher and comparable BLEU and RIBES scores between Myanmar and three dialects (Rakhine, Dawei and Myeik) in both directions. The OSM approach achieved the highest BLEU and RIBES scores among three approaches for both word and syllable segmentations. Moreover, we found that syllable segmentation is appropriate for translation quality comparing with word level segmentation results.**

***Index Terms*—Statistical Machine Translation, Parallel Corpus Developing, Myanmar (Burmese), Rakhine (Arakanese), Dawei (Tavoyan), Myeik (Beik).**

I. Introduction

MYANMAR language includes a number of mutually intelligible Myanmar dialects, with a largely uniform standard dialect used by most Myanmar standard speakers. Speakers of the standard Myanmar may find the dialects hard to follow. The alternative phonology, morphology, and regional vocabulary cause some problems in communication. Machine Translation has so far neglected the importance of properly handling the spelling, lexical and grammar divergences among language varieties. Our main motivation for this work is to investigate SMT performance for Myanmar (Burmese) and Dialectal language pair including Rakhine (Arakanese), Dawei (Tavoyan) and Myeik (Beik). The state-of-the-art techniques of statistical machine translation (SMT) [1], [2] demonstrate good performance on translation of languages with relatively similar word orders [3]. To date, there have been some studies on the SMT of Myanmar language. Ye Kyaw Thu et al. (2016) [4] presented the first large-scale study of the translation of the Myanmar language. A total of 40 language pairs were used in the study that included languages both similar and fundamentally different from Myanmar. The results show that the hierarchical phrase-based SMT (HPBSMT) [5] approach gave the highest translation quality in terms of both the BLEU [6] and RIBES scores [7]. Win Pa Pa et al (2016) [8] presented the first comparative study of five major machine translation approaches applied to low-resource languages. PBSMT, HPBSMT, tree-to-string (T2S), string-to-tree (S2T) and OSM translation methods to the translation of limited quantities of travel domain data between English and Thai, Laos, Myanmar in both directions. The experimental results indicate that in terms of adequacy (as measured by BLEU score), the PBSMT approach produced the highest quality translations. Here, the annotated tree is used only for English language for S2T and T2S experiments. This is because there is no publicly available tree parser for Lao, Myanmar and Thai languages. According to our knowledge, there is no publicly available tree parser for Myanmar, Rakhine, Dawei and Myeik languages and thus we cannot apply S2T and T2S approaches for Myanmar dialect translations. From their RIBES scores, we noticed that OSM approach achieved best machine translation performance for Myanmar to English translation. Moreover, we learned that OSM approach gave highest translation performance translation between Khmer (the official language of Cambodia) and twenty other languages, in both directions [9]. Based on the experimental results of previous works, in this paper, the machine translation experiments were carried out using PBSMT, HPBSMT and OSM.

II. Related Work

Karima Meftouh et al. built PADIC (Parallel Arabic Dialect Corpus) corpus from scratch, then conducted experiments on cross dialect Arabic machine translation [10]. PADIC is composed of dialects from both the Maghreb and the Middle-East. Some interesting results were achieved even with the limited corpora of 6,400 parallel sentences. Using SMT for dialectal varieties usually suffers from data sparsity, but combining word-level and character-level models can yield good results even with small training data by exploiting the relative proximity

Thazin Myint Oo and Khin Mar Soe are with the NLP Lab., University of Computer Studies Yangon, Myanmar.

Ye Kyaw Thu and Thepchai Supnithi are with National Electronics and Computer Technology Center, Thailand., Contact author e-mail: yktnlp@gmail.com

Manuscript received December 21, 2019; accepted March 6, 2020; revised March 18, 2020; published online April 30, 2020.

between the two varieties [11]. Friedrich Neubarth et al. described a specific problem and its solution, arising with the translation between standard Austrian German and Viennese dialect. They used hybrid approach of rule-based preprocessing and PBSMT for getting better performance. Pierre-Edouard Honnet et al. proposed solutions for the machine translation of a family of dialects, Swiss German, for which parallel corpora are scarce [12]. They presented three strategies for normalizing Swiss German input in order to address the regional and spelling diversity. The results show that character-based neural MT was the most promising one for text normalization and that in combination with PBSMT achieved 36% BLEU score.

III. Dialectal Languages

Dialet refers to a variety of a language that is a characteristics of a particular group of the language's speakers. The dialects or varieties of a particular language are closely related, and despite their differences, are most often largely mutually intelligible, especially if close to one another on the dialect continuum. The term is applied most often to regional speech patterns. Arakanese, Intha and Tavoyan are three regional dialects of Burmese [13]. There are many other regional dialects in Myanmar such as Danu, Taung-yoe, Myeik and Yaw. We studied on three main dialect such as Rakhine (Arakanese), Dawei (Tavoyan) and Myeik (Beik).

A. Rakhine Language

Rakhine (Arakanese) is one of the eight national ethnic groups in the Republic of the Union of Myanmar. The Arakan was officially altered to "Rakhine" in 1989 and is located on a narrow coastal strip on the west of Myanmar, 300 miles long and 50 to 20 miles wide. The total population in all countries is nearly 3 million. The Rakhine language has been studied by researchers. L.F-Taylor's "The Dialects of Burmese" described comparative pronunciation, sentence construction, and grammar usage in Rakhine, Dawei, In-tha, Taung-yoe, Danu, and Yae. Professor Denise Bernot, in "The vowel system of Arakanese and Tavoyan," mainly emphasized the vowels of standard Myanmar and Tavoyan (Dawei) in 1965. In "Three Burmese Dialects" (1969), the linguist John Okell studied the spoken language of Myanmar, Dawei, and In-tha: specifically, usage of grammar and vowel differences [13]. Although the Rakhine language used the script as Arakanese or Rakkhawanna Akkhara before at least the 8th century A.D., the current Rakhine script is nearly the same as the Myanmar script. Generally, the Arakanese language is mutually intelligible with the Myanmar language and has the same word order (namely, subject-object-verb (SOV)). Examples of parallel sentences in Myanmar (my) and Rakhine (rk) are given as follows.

rk: ဒယော တစ် ထည် ဈာလောက်လေး ။
my: လုံချည် တစ် ထည် ဘယ်လောက်လဲ ။
("How much for a longyi?" in English)

rk: ကလေချေ တိ ဘောလုံး ကျောက် နိရေ ။
my: ကောင်လေး တွေ ဘောလုံး ကန် နေတယ် ။
("Boys are playing football" in English)

rk: ဇာ ပြော နိချင့် ယင်းသူရို့ ။
my: သူတို့ ဘာ ပြော နေတာလဲ ။
("What are they talking about" in English)

rk: အဘောင်သျင် ဈီး က သပုံ ဝယ် လာတယ် ။
my: အဘွား ဈေး က ဆပ်ပြာ ဝယ် လာတယ် ။
("The grandmother buys soap from the market" in English)

The Rakhine language is a largely monosyllabic and analytic language, with a SOV word order, and it uses the Myanmar script. It is considered by some to be a dialect of the Myanmar language, though it differs significantly from the standard Myanmar language in its vocabulary and includes loan words from Bengali, Hindi, and English. Compared with the Myanmar language, the speech of the Rakhine language is likely to be closer to the written form. The Rakhine language notably retains an /r/ sound that has become /j/ in the Myanmar language. Rakhine speakers pronounce the medial "ျ" as "Yapint" (i.e., /j/ sound) and the medial "ြ" as "Rayit" (i.e., /r/ sound). Moreover, Myanmar vowel "ေ" (/e/ sound) is pronounced as "ိ" (/i/ sound) in Rakhine language. Thus, for example, the word "dog" in the Myanmar language is written as "ခွေး" (Khwe), and in the Rakhine language it is written as "ခွိး" (khwii). Similarly, Rakhine pronounce "ေး"(/e:/) for Myanmar pronunciation of "ဲ"(/ai/) syllable. Thus, Myanmar word "ပဲဟင်း"(peh-hinn) (pea curry in English) is pronounced "ပေးဟင်း" (pay-hinn) in the Rakhine language. Some Pali words are also used in the Rakhine language. For example, the word "guest" of Myanmar monks "အာဂန္တု" (agantu) is used in normal speech of Rakhine and it is similar to the word of normal Myanmar people "ဧည့်သည်" (ei the), "guest," in English. In summary, the most significant differences between the Rakhine and Myanmar languages are in their pronunciation and vocabulary, and there are no grammatical differences.

B. Dawei Language

The Tavoyan or Dawei dialect of Burmese is spoken in Dawei (Tavoy), in the coastal Tanintharyi Region of southern Myanmar (Burma). The large and quite distinct Dawei variety is spoken in and around Dawei (formerly Tavoy) in Tanintharyi (formerly Tenasserim) by about 400,000 people; its sterotyped characteristic is the mesial /I/, found in earlist Bagan inscriptions but by merger there nearly 800 years ago; for further information see Pe Maung Tin (1933) [14] and Okell (1995) [13]. Dawei is a city of south-eastern Myanmar and is the capital of

Tanintharyi Region, formerly known as the Tenasserim is bounded by Mon state to the north, Thailand to the east and south, and the Andaman sea to the west. Dawei language retains /-l-/ medial that has since merged into the /-j-/ medial in standard Burmese and can form the following consonant clusters: /gl-/, /kl-/, /kʰl-/, /bl-/, /pl-/, /pʰl-/, /ml-/, /m̥l-/. Examples include "မ္လေ" (/mlè/ → Standard Burmese /mjè/) for "ground" and "ကျွောင်း" (klɑ́ʊɴ/ → Standard Burmese tʃáʊɴ/) for "school". Also, voicing only with unaspirated consonants, whereas in standard Burmese, voicing can occur with both aspirated and unaspirated consonants. Also, there are many loan words from Malay and Thai not found in Standard Burmese. An example is the word for goat, which is hseit "ဆိတ်" in Standard Burmese but be "ဘဲ" in Dawei language.

In the Dawei dialect, terms of endearment, as well as family terms, are considerably different from Standard Burmese. For instance, the terms for "son" and "daughter" are "ဖစု" (/pʰa̰ ðu/) and "မိစု"(/mḭ ðu/) respectively. Moreover, the honorific "နောင်" (Naung) is used in lieu of "မောင်" (Maung) for young males. Another evidence of "Dawei" is "Dhommarazaka" pogoda inscription of Bagan period. It was inscription of Bagan period. It was inscribed in AD 1196 during the region of Bagan King Narapatisithu (AD 1174-1201) . In this inscription line 6 to 19, when the demarcation of Bagan is mentioned "Taung-Kar-Htawei" (up to Htawei to the south) and "Taninthaye" (Tanintharyi) are including. Therefore, the name of "Dawei" appeared particulary since Bagan period, at the time of the first Myanmar Empire. (Dawei was established at Myanmar year 1116) is actually meant that the present name Dawei appears as the name of the settlers later and the original name of the city is Tharyarwady, which was established at Myanmar year 1116 according to the saying. As "Dawei" nationality deserves as one nationalist in our country. Actually, Dawei region is a place where local people lived since very ancient Stone Age. After that, Stone Age, Bronze Age and Iron Age culture developed. Moreover, as there has sound evidence of Thargara ancient city, comtemporary to Phu Period, the Dawei people, can be assumed that they are one nationality of high culture in Myanmar.

Dawei usage and vocabularies is divided into three main groups. The first one is using Myanmar vocabularies with Dawei speech, the second is the vocabularies same with Myanmar vocabularies and using isolated Dawei words and vocabularies. In Myanmar word "ထို, ဟို", ("here, there") is used "သယ်" ("here") and "ဟောက်" ("there") in Dawei language. For example Dawei word "သယ်မျိုး" is same as "ဒီလို" in Myanmar language and "ဟောက်မျိုး" means "ဟိုလို" in Myanmar language. The question words "နည်း (သနည်း), လဲ (သလဲ)" are used in Myanmar language, similarly "လော, လော်" are used instead of "လား (သလား)" in Dawei language. Moreover, "ဘာလဲ"(what) and "ဘာဖြစ်တာလဲ" ("what happened") is same with "ဖြာနူး" and "ဖြာဖြစ်နူး" in Dawei usage. In negative sense of Myanmar word "ဘူး" is not usually used in Dawei word. The negative Dawei words are "ဟု (ရ)" or "ဟန်း" ("No" in English). Myanmar adverb word "သိပ်, အလွန်, အလွန်အလွန်" (very, extremely) is used as "ရရာ, ရမိရရာ, ဗြင်း". Some more example of Dawei vocabularies are "ဝန်းရှင်း", "ကိုယ်ဝန်ဆောင်" in Myanmar language, ("pregnant" in English), "ကောန်သား", "ကောင်လေး" in Myanmar language, ("boy" in English), "ဝယ်သား", "ကောင်မလေး" in Myanmar language ("girl" in English), "ကပ်" "ပိုက်ဆံ" in Myanmar language, ("money" in English), "ချော့-က်တိုအိုးသီး" "ကျွဲကောသီး" in Myanmar language, ("pomelo" in English) and "သစ်ခတ်ကွား" "ကျားသစ်" in Myanmar language ("leopard" in English). The followings are some example parallel sentences of Myanmar (my) and Dawei (dw):

dw: သယ်ဝယ်သား က လှ ဗြင်း ဟုယ် ။
my: ဒီကောင်မလေး က လှ လွန်း တယ် ။
("The girl is so beautiful" in English)

dw: လတ်ဖတ်ရယ် က ရှိ ဗြင်း ဟုယ် ။
my: လက်ဖက်ရည် က ချို လွန်း တယ် ။
("The tea is so sweet" in English)

dw: ကောန်သား ကျောန်း မှန်းမှန် သွား ဟုယ် ။
my: ကောင်လေး ကျောင်း မှန်မှန် တက် တယ် ။
("The boy goes to school regularly" in English)

C. Myeik Language

Myeik dialect has peculiar characteristics in terms of tonal contours, and voice quality in the tones and vowels. tone of this dialect, which corresponds to the Standard Burmese creaky falling tone, has a rising contour and is pharyngealized [28]. Vowels of the syllables corresponding to Standard Burmese stopped syllables are pronounced with a conspicuous creaky phonation. Tone sandhispeculiar to this dialect are also described in this paper [29]. Dialogues cover as many as possible of the most basic grammatical items of Burmese, translating them into the Myeik dialect can be the basis for future studies of morphosyntactic phenomena of this dialect [30].

The Myeik dialect is a dialect of Burmese that is spoken in Myeik (Beik), a town situated in the southern part of Tanintharyi Division (around 12Â°25'N, 98Â°37'E), Republic of the Union of Myanmar. Myeik dialect is one of the southernmost dialects of Burmese and can be regarded as the southernmost distribution of the Tibeto-Burman languages. Myeik was formerly called Mergui in English. Standard Burmese pronunciation of the name of the town Beik and the Myeik dialect calls the town Beik. This article presents basic material on the Myeik dialect of Burmese, covering sounds, conversational texts, and basic vocabulary.

[X] [X] ကို တောပန် [X] ||| [X] [X] ကို တောင်းပန် [X]
[X] [X] ကို တွန်းပန် [X] ||| [X] [X] မ တောင်းပန် [X]
[X] [X] ကို တွန်းပန် ဟ့လား [X] ||| [X] [X] မ တောင်းပန် ဘူးလား [X]
[X] [X] ကို တို့ [X] ||| [X] [X] ကို တွေ့ [X]
[X] [X] ကို တို့ ဟို့ [X] ||| [X] [X] ကို တွေ့ ဖို့ [X]

Fig. 1: Some examples of hierarchical phrase-based grammar between Dawei and Myanmar phrases

bk : မင်း ငါ့ ကို ကြေးပြား ပေး ဝို့ မေ့ နေရယ် လား ။
my: မင်း ငါ့ ကို ပိုက်ဆံ ပေး ဖို့ မေ့ နေပြီလား ။
(“Do you forget paying money to me.” in English)

bk : ငါ မောလင်း နိုင်ငံခြား သော မယ်။
my : ကျွန်တော် မနက်ဖြန် နိုင်ငံခြား သွား မယ် ။
(“I will go foreign tomorrow .” in English)

bk : ကျွန်တော် ဒယ် ဝို လာ ရဇာ ပျော် ရယ် ။
my :ကျွန်တော် ဒီ လာ ရတာ ပျော် တယ် ။
(“I am happy to come here.” in English)

In the above examples, the underlined words that have same meaning but have different spellings such as “ကြေးပြား” vs “ပိုက်ဆံ” (“money”) in English), “မောလင်း” vs “မနက်ဖြန်” (“tomorrow” in English), “ဒယ်” vs “ဒီ” (“this” in English).

IV. Methodology

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

A. Phrase-Based Statistical Machine Translation

A PBSMT translation model is based on phrasal units [1]. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We can describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table [17]. The phrase translation model is based on noisy channel model. To find best translation $\hat{e}$ that maximizes the translation probability $\mathbf{P}(f)$ given the source sentences; mathematically. Here, the source language is French and the target language is an English. The translation of a French sentence into an English sentence is modeled as equation 1.

$$\hat{e} = argmax_e \mathbf{P}(e|f) \tag{1}$$

Applying the Bayes' rule, we can factorized the into three parts (see equation 2).

$$P(e|f) = \frac{\mathbf{P}(e)}{\mathbf{P}(f)} \mathbf{P}(f|e) \tag{2}$$

The final mathematical formulation of phrase-based model is as equation 3:

$$argmax_e \mathbf{P}(e|f) = argmax_e \mathbf{P}(f|e)\mathbf{P}(e) \tag{3}$$

We note that denominator $\mathbf{P}(f)$ can be dropped because for all translations the probability of the source sentence remains the same. The $\mathbf{P}(e|f)$ variable can be viewed as the bilingual dictionary with probabilities attached to each entry to the dictionary (phrase table). The $\mathbf{P}(e)$ variable governs the grammaticality of the translation and we model it using n-gram language model under the PBMT paradigm.

B. Hierarchical Phrase-Based Statistical Machine Translation

The hierarchical phrase-based SMT approach is a model based on synchronous context-free grammar [5]. The model is able to be learned from a corpus of unannotated parallel text. The advantage of this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word reordering process. The reordering is represented explicitly rather than encoded into a lexicalized reordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to language pairs that require long-distance reordering during the translation process [16]. Some examples of hierarchical phrase based grammar between Dawei and Myanmar phrases are shown in Figure 1.

C. Operation Sequence Model

The operation sequence model that can combines the benefits of two state-of-the-art SMT frameworks named n-gram-based SMT and phrase-based SMT. This model simultaneously generate source and target units and does not have spurious ambiguity that is based on minimal translation units [17] [18]. It is a bilingual language model that also integrates reordering information. OSM motivates better reordering mechanism that uniformly handles local and non-local reordering and strong coupling of lexical generation and reordering. It means that OSM can handle both short and long distance reordering. The operation types are such as generate, insert gap, jump back and jump forward which perform the actual reordering. The following shows an example translation process of English sentence “Please sit here”

into Myanmar language with the OSM.

Source: Please sit here
Target: ကျေးဇူးပြုပြီး ဒီမှာ ထိုင်

Operation 1: Generate (Please, ကျေးဇူးပြုပြီး)
Operation 2: Insert Gap
Operation 3: Generate (here, ကျေးဇူးပြုပြီး ဒီမှာ)
Operation 4: Jump Back (1)
Operation 5: Generate (sit, ကျေးဇူးပြုပြီး ဒီမှာ ထိုင်)

V. Experiments

A. Corpus Statistics

We used Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [19], 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), special needs (emergency and health). In Rakhine-Myanmar parallel corpus, we used 18,373 Myanmar sentences. Word segmentation for Rakhine was done manually and there are exactly 123,018 words in total. We held 10-fold cross-validation experiments and used 14,023 to 14,078 sentences for training, 2,475 to 2,485 sentences for development and 1,810 to 1,875 sentences for evaluation respectively.

In Dawei-Myanmar corpus, using 9,000 Myanmar sentences We held 10-fold cross-validation experiments and used 6,883 to 6,893 sentences for training, 1,212 to 1,217 sentences for development and 890 to 922 sentences for evaluation respectively.

Myeik-Myanmar parallel corpus have 10K sentences in total. Manual Translation into Myeik Language was done by native Myeik students from Computer University (Myeik). Word segmentation for Myeik was done manually and there are exactly 68,035 words in total. We held 10-fold cross-validation experiments and used 7,867 to 7,893 sentences for training, 1,389 to 1,393 sentences for development and 1,014 to 1,044 sentences for evaluation respectively.

B. Word Segmentation

1) Word Segmentation For Rakhine Language

In both Myanmar and Rakhine texts, spaces are used to separate the phrases for easier reading. The spaces are not strictly necessary and are rarely used in short sentences. There are no clear rules for using spaces. Thus, spaces may (or may not) be inserted between words, phrases, and even between root words and their affixes. Although Myanmar sentences of ASEAN-MT corpus [19] are already segmented, we have to consider some rules for manual word segmentation of Rakhine sentences. We defined Rakhine "word" to be a meaningful unit. Affix, root word, and suffix (s) are separated such as "စား ဗျာယ်", "စား ပီးဗျာယ်", "စား ဖို့ဗျာယ်". Here, "စား" ("eat" in English) is a root word and the others are suffixes for past and future tenses. As Myanmar language, Rakhine plural nouns are identified by the following particle. We added a space between the noun and the following particle: for example a Rakhine word "ကလိန့်မေချေ တိ" (ladies) is segmented as two words "ကလိန့်မေချေ" and the particle "တိ". In Rakhine grammar, particles describe the type of noun and are used after a number or text number. For example, a Rakhine word "အကြိုစေ့နှစ်ခတ်" ("two coins" in English) is segmented as "အကြိုစေ့ နှစ် ခတ်". In our manual word segmentation rules, compound nouns are considered as one word. Thus, a Rakhine compound word "ဖေ့သာ" + "အိတ်" ("money" + "bag" in English) is written as one word ဓဖေ့သာအိတ်'' ("wallet" in English). Rakhine adverb words such as "အဂယောင့်" ("really" in English), "အမြန်" ("quickly" in English) are also considered as one word. The following is an example of word segmentation for a Rakhine sentence in our corpus, and the meaning is "Among the four air conditioners in our room, two are out of order."

Unsegmented sentence:
အကျွန်ရို့အခန်းထဲမှာဟိရေလီအီးစက်လေးလုံးမှာနှစ်လုံးပျက်နိရေ ။

Segmented sentence:
အကျွန်ရို့ အခန်း ထဲမှာ ဟိ ရေ လီအီးစက် လေး လုံး မှာ နှစ် လုံး ပျက် နိရေ ။

2) Word Segmentation For Dawei Language

In both Myanmar and Dawei text, spaces are used for separating phrases for easier reading. It is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces, and thus spaces may (or may not) be inserted between words, phrases, and even between a root words and their affixes. Although Myanmar sentences of ASEAN-MT corpus [19] is already segmented, we have to consider some rules for manual word segmentation of Dawei sentences.

We defined Dawei "word" to be meaningful units and affix, root word and suffixe(s) are separated such as "စား ဟှယ်", "စားပီးဟှယ်", "စား ဖို့ဟှယ်''. Here, "စား" ("eat"in English) is a root word and the others are suffixes for past and future tenses. Similar to Myanmar language, Dawei plural nouns are identified by following particle. We also put a space between noun and the following particle, for example a Dawei word "ဇွန်သားဒေ" (shrimps) is segmented as two words "ဇွန်သား" and the particle "ဒေ". In Dawei grammar, particles describe the type of noun, and used after number or text number. For example, a Dawei word "ရီးခိုသီးတစ်လုံး" ("papaya" in English) is segmented as "ရီးခိုသီး တစ် လုံး". In our manual word segmentation rules, compound nouns are considered as one word and thus, a Dawei compound word "ကပ်" + "အိတ်" ("money" + "bag" in English) is written as one word "ကပ်အိတ်" ("wallet" in English). Dawei adverb words such as "ရရာ, ရမိရရာ" ("very" in English), "ဗြင်း" ("extremely" in English) are

also considered as one word. The following is an example of word segmentation for a Dawei sentence in our corpus and the meaning is "Shrimps are very rare and bought fishes."

Unsegmented Dawei sentence:
dw: ဖွန်သားဒေရရာရှားဟှယ်၊ငါးဗောင်းသားဘဲ့ဝယ်လာရဟှယ်။

Word Segmented Dawei sentence:
dw: ဖွန်သား ဒေ ရရာ ရှား ဟှယ် ၊ ငါးဗောင်းသား ဘဲ့ ဝယ် လာရဟှယ် ။

In this example, "ဖွန်သားဒေ" (shrimps) is segmented as two words "ဖွန်သား" and the particle "ဒေ". Dawei adverb words such as "ရရာ" ("rare" in English) is also considered as one word and a root word "ဝယ်" and the suffix "လာရဟှယ်" are also segmented as two words "ဝယ် လာရဟှယ်" ("bought" in English).

3) Word Segmentation For Myeik Language

To consider some rules for manual word segmentation of Myeik sentences. We defined Myeik "word" to be meaningful units and affix, root word and suffixe(s) are separated such as "စား ရယ်". Here, "စား" ("eat" in English) is a root word and suffixes for past. Similar to Myanmar language, Myeik plural nouns are identified by following particle. We also put a space between noun and the following particle, for example a Myeik word "သားကင်းငယ်တွေ" (children) is segmented as two words "သားကင်းငယ်" and the particle "တွေ". In our manual word segmentation rules, compound nouns are considered as one word and thus, a compound word "ကြေးပြား" + "အိတ်" ("money" + "bag" in English) is written as one word "ကြေးပြားအိတ်" ("wallet" in English). Rakhine adverb words such as "အား" ("very" in English) also considered as one word. The following is an example of word segmentation for a sentence in our corpus and the meaning is "why are you beaten the children."

Unsegmented sentence:
dw:ဘာဖြစ်ရိသားကင်းငယ်တွေကိုရိုက်နေရယ်။

Segmented sentence:
dw: ဘာဖြစ်ရိ သားကင်းငယ် တွေ ကို ရိုက် နေရယ်။

In this example, "သားကင်းငယ်တွေ " ("children" in English) is a compound word of "သားကင်းငယ်" ("child" in English) and a particle "တွေ" are segmented as two words. A root word "ရိုက်" and the suffix "နေရယ်" are also segmented as two words "ရိုက် နေရယ်" ("out of order" in English.

C. Syllable Segmentation

Generally, Myanmar words are composed of multiple syllables, and most of the syllables are composed of more than one character. Syllables are composed of Myanmar words. If we only focus on consonant-based syllables, the structure of the syllable can be described with Backus normal form (BNF) as follows:

$$\text{Syllable} := \text{CMV[CK][D]}$$

Here, C stands for consonants, M for medials, V for vowel, K for vowel killer character, and D for diacritic characters. Myanmar syllable segmentation can be done with a rule-based approach, finite state automation (FSA) or regular expressions (RE) (https://github.com/ye-kyaw-thu/sylbreak).

D. Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit [2] for training the PBSMT, HPBSMT and OSM statistical machine translation systems. The word segmented source language was aligned with the word segmented target language using GIZA++ [20]. The alignment was symmetrize by grow-diag-final and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [21]. We use KenLM [22] for training the 5-gram language model with modified Kneser-Ney discounting [31]. Minimum error rate training (MERT) [20] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [2]. We used default settings of Moses for all experiments.

VI. Evaluation

We used two automatic criteria for the evaluation of the machine translation output. One was the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [6] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [7]. The BLEU score measures the precision of n-gram (over all n 4 in our case) with respect to a reference translation with a penalty for short translations. Intuitively, the BLEU score measures the adequacy of the translation and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES score is suitable for distance language pairs such as Myanmar and English. Large RIBES scores are better.

VII. Results and Discussion

The average BLEU and RIBES score results for Rakhine-Myanmar bi-directional machine translation experiments with two types of segmentation for PBSMT, HPBSMT and OSM are shown in Table I and Table II. Bold numbers indicate the highest scores among three SMT approaches. The RIBES scores are inside the round brackets. Here, "my" stands for Myanmar, "rk" stands for Rakhine, "src" stands for source language and "tgt" stands for target language respectively. The average BLEU and RIBES scores for Dawei-Myanmar bi-directional word and syllable segmentation unit is shown in Table III and

Table IV. Here, “dw” stands for Dawei language. The Myeik-Myanmar bi-directional machine translation unit is demonstrated on Table V and Table VI. At these tables “bk” stands for Beik or Myeik language and the average BLEU and RIBES scores are also indicated.

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM between Myanmar and Rakhine languages using word segmentation evaluation with syllable units are shown in Table I. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Rakhine and Rakhine-Myanmar machine translations. Interestingly, the BLEU and RIBES score of all three methods are comparable performance. Our results with current parallel corpus indicate that Rakhine to Myanmar machine translation is better performance (around 3 BLEU and 0.02 RIBES scores higher) than Myanmar to Rakhine translation direction. The BLEU and RIBES score results for syllable segmentation for machine translation experiments with PBSMT, HPBSMT and OSM between Myanmar and Rakhine Languages are shown in Table II. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Rakhine and Rakhine-Myanmar machine translations. Interestingly, the BLEU and RIBES score of all three methods are comparable performance. Our results with current parallel corpus indicate that Rakhine to Myanmar machine translation is better performance (around 2 BLEU and 0.001 RIBES scores higher) than Myanmar to Rakhine translation direction.

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM using word level segmentation between Myanmar and Dawei languages are shown in Table III. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Dawei and Dawei-Myanmar machine translations. Our results with current parallel corpus indicate that Dawei to Myanmar machine translation is better performance (around 9 BLEU and 0.02 RIBES scores higher) than Myanmar to Dawei translation direction. The results of BLEU and RIBES scores of syllable segmentaion between Myanmar and Dawei languages are shown in Table IV. Our results with syllable segmentation also indicate that Dawei to Myanmar machine translation is better performance (around 18 BLEU and 0.03 RIBES score higher) than Myanmar to Dawei translation direction. Our investigatation clearly show that getting the higher scores with syllable segmentation for bi-directional Myanmar to Dawei machine translation. The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM between Myanmar and Myeik languages are shown in Table V. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Myeik and Myeik-Myanmar machine translations. The BLEU and RIBES score of all three methods are comparable performance. Our results with current parallel corpus indicate that Myeik to Myanmar machine translation is better performance (around 10 BLEU and 0.04 RIBES scores higher) than Myanmar to Myeik translation direction. Our results with syllable segmentation shown in Table VI also indicate that Myeik to Myanmar machine translation is better performance (around 15 BLEU and 0.03 RIBES score higher) than Myanmar to Myeik translation direction. Our investigation clearly show that getting the higher scores with syllable segmentation for bi-directional Myanmar to Myeik machine translation.

VIII. Error Analysis

We also used the SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK (Speech Recognition Scoring Toolkit) version 2.4.10 [26] for making dynamic programming based alignments between reference and hypothesis strings for detail analysis on translation errors in terms of WER (word error rate). The SCLITE scoring method for calculating the erroneous words in WER: first make an alignment of the hypothesis (the translated sentences) and the reference and then perform a global minimization of the Levenshtein distance function which weights the cost of correct words, insertions (I), deletions (D), substitutions (S) and the number of words in the reference (N). The formula for WER can be stated as equation 4:

$$WER = \frac{(N_i + N_d + N_s) \times 100}{N_d + N_s + N_c} \tag{4}$$

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 details.

A. Error Analysis for Rakhine Language

We studied on detailed error analysis on calculation Word Error Rate (WER) for Rakhine Language. For example, scoring I, D and S for the translated Rakhine sentence “ဇာ အိမ်မှာ မင်းနီလေး ။ ” (“Which house do you live in?”) in English, “ဘယ် အိမ် မှာ မင်း နေ သလဲ ” in Myanmar language) compare to a reference sentence, the output of the SCLITE program is as follows:

Scores: (#C #S #D #I) 2 1 0 1
REF : *** ဇာအိမ်မှာ မင်းနီလေး ။
HYP : ဇာ အိမ်မှာ မင်းနီလေး ။
Eval : I S

In this case, one insertion (*** => ဇာ) and one substitution (ဇာအိမ်မှာ => အိမ်မှာ) happened , that is $S = 1$, $D = 0$, $I = 1$, $C = 1$, $N = 2$ and thus WER is equal to 66.67%.

TABLE I: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM for Rakhine and Myanmar Translation using word Segmentation (Evaluation with Syllable Unit)

src-tgt	PBSMT	HPBSMT	OSM
my-rk	57.68 (0.9077)	57.70 (0.9073)	**57.88 (0.9085)**
rk-my	60.58 (0.9233)	60.42 (0.9230)	**60.86 (0.9239)**

TABLE II: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM for Rakhine and Myanmar Translation using syllable Segmentation

src-tgt	PBSMT	HPBSMT	OSM
my-rk	83.39 (0.9778)	83.17 (0.9778)	**83.83 (0.9784)**
rk-my	84.27 (0.9784)	84.06 (0.9779)	**85.18 (0.9798)**

TABLE III: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM for Dawei and Myanmar Translation using word Segmentation (Evaluation with Syllable Unit)

src-tgt	PBSMT	HPBSMT	OSM
my-dw	39.46 (0.8894)	39.22 (0.8870)	**39.77 (0.8938)**
dw-my	47.49 (0.9181)	47.80 (0.9179)	**48.15 (0.9187)**

TABLE IV: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM for Dawei and Myanmar Translation using syllable Segmentation

src-tgt	PBSMT	HPBSMT	OSM
my-dw	44.80 (0.9160)	45.44 (0.9149)	**45.58 (0.9155)**
dw-my	60.78 (0.9461)	60.47 (0.9447)	**63.22 (0.9482)**

TABLE V: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM For Myeik and Myanmar Translation using word Segmentation (Evaluation with Syllable Unit)

src-tgt	PBSMT	HPBSMT	OSM
my-bk	33.25 (0.8403)	33.33 (0.8388)	**33.41 (0.8399)**
bk-my	44.12 (0.8749)	44.07 (0.8751)	**44.33 (0.8753)**

TABLE VI: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM For Myeik and Myanmar Translation using syllable Segmentation

src-tgt	PBSMT	HPBSMT	OSM
my-bk	54.60 (0.9221)	54.40 (0.9220)	**55.11 (0.9232)**
bk-my	70.02 (0.9573)	69.89 (0.9566)	**70.55 (0.9579)**

Scores: (#C #S #D #I) 3 2 0 1
REF : မင်း*******အစ်မ အိမ်ထောင် ဟိလား ။
HYP : မင်း အစ်မ အိမ်ထောင် ဟိလား ။
Eval: I S S

In this case, one insertion (***=>အစ်မ), two substitution (အိမ်ထောင်ဟိ =>အိမ်ထောင်) and (လား =>ဟိလား) happened, that is $S = 2$, $D = 0$, $I = 1$, $C = 2$ and thus WER is equal to 60%. The WER% of PBSMT, HPBSMT and OSM for Myanmar to Rakhine and Rakhine to Myanmar translations with around 1,800 test sentences (one-tenth of 18,373 total sentences) are as shown in Table VII. From the Table VI, we found that WER% for all three approaches are very closed to each other. OSM achieved the lowest WER% and on the other hand, HPSMT method is highest WER%. However, WER calculation does not consider the contextual and syntactic roles of a word. For this reason, we made manual analysis on error types of each SMT model. We found that some extra words are containing in the translated outputs of all three SMT approaches especially for Myanmar to Rakhine machine translation. For example, translated output containing one extra word "က" for Myanmar to Rakhine translation for the source sentence "နောက် တချက်ချေမာပင် မျောက်တိ က အလားတူ လိုက်လုပ် ကတ်ရေ ။ " ("The next moment, the monkeys were doing the same." in English). However, Rakhine to Myanmar translation, all three models rarely gave that kind of error. See following example, source, reference, and hypothesis of three models in detail:

SOURCE:
နောက် တချက်ချေမာပင် မျောက်တိ က အလားတူ လိုက်လုပ် ကတ်ရေ ။
REF:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ လိုက်လုပ် ကတ်ရေ ။
HYP of PBSMT:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ **က** လိုက်လုပ် ကတ်ရေ ။

Here, the word highlighted with bold color is the extra Rakhine word "Ka."

After we made analysis of confusion pairs of each model in details, we found that some of the confusion pairs are relating to word segmentation and typing errors (refer Table VIII). Here, the confusion pairs of "ပါ။" ==> "။", "ငါ" ==> "ကျွန်တော်", "ကို "==> "ယင်းချင့်ကို", "ရို့" ==> "သူရို့", "အကျွန်" ==> "ကျွန်တော်" and "လား" ==> "ပါလား" are happened because of words segmentation error of Myanmar sign section "။". The confusion pair "နန့် "==> "နန့်" is occurred because of different typing order. Although they look the same, the typing order of the reference "နန့်" is "နှ, န, ် , ့" (correct order) and the that of hypothesis "နန့်" is "န, န, ့ , ်". These kind of confusion pairs can be reduced by cleaning of current word segmentation and typing errors of our parallel corpus.

B. Error Analysis for Dawei Language

From our studies, the top 15 confusion matrix for Dawei-Myanmar OSM machine translation (with word segmentation) can be seen in Table IX.

We also made manual error analysis on translated outputs of the best OSM model, and we found that dominant errors are different in sentence level. We will introduce four frequent error patterns and they are "Male-Female Vocabulary Error", "Paraphrasing Error", "Word Segmentation Error" and "Negative Error". The followings are some example translation mistakes for each category:

Male-Female Vocabulary Error

SOURCE: **သူ နန့် ဟို့ ဗြင် လား ။**
Scores: (#C #S #D #I) 3 2 0 1
REF: ****** သူမ မင်းကို မြင် သလား ။
HYP: သူ မင်း ကို မြင် သလား ။
Eval: I S S

SOURCE: **သူကိုယ်သူ သိ ဟုယ် ။**
Scores: (#C #S #D #I) 3 1 0 0
REF: သူမကိုယ်သူမ သိ ပါတယ် ။
HYP: သူကိုယ်သူ သိ ပါတယ် ။
Eval: S

Paraphrasing Error

SOURCE: **ဍား ဟုးဟို့ အီ လေ ။**
Scores: (#C #S #D #I) 4 1 0 0
REF: ဍားရမ်း ထားတဲ့ အိမ် တွေ ။
HYP: ဍား ထားတဲ့ အိမ် တွေ ။
Eval: S

SOURCE: **လူတိုင်း သတ္တိ ရှိ ကေ့ဟုယ် ။**
Scores: (#C #S #D #I) 4 1 0 0
REF: လူတိုင်း သတ္တိ ရှိ ကြပါတယ် ။
HYP: လူတိုင်း သတ္တိ ရှိ ကြတယ် ။
Eval: S

SOURCE: **ကျွန်တော် အိ ရှင်နေဟုယ် ။**
Scores: (#C #S #D #I) 3 1 0 2
REF: ကျွန်တော် အိပ် ****** ချင်နေတယ် ။
HYP: ကျွန်တော် အိပ် ဖို့ ဆန္ဒရှိ တယ် ။
Eval: I I S

SOURCE: **သူဟု ရတိုင်း လု မား ။**
Scores: (#C #S #D #I) 3 2 0 0
REF: သူက အရမ်း လှ တာပဲ ။
HYP: သူက သိပ် လှ ရော ။
Eval: S S

Word Segmentation Error

SOURCE: **အဲဝယ်ဟုး ကားမွန်း ဟို့မူဝလား ။**
Scores: (#C #S #D #I) 4 1 1 0
REF: သူမ ကား မောင်း မှာ မဟုတ်ဘူးလား ။
HYP: သူမ ********* ကားမောင်း မှာ မဟုတ်ဘူးလား ။
Eval: D S

SOURCE: **အယ်မို့ဇာ ပိုဆိုး လာဟုယ် ။**
Scores: (#C #S #D #I) 3 1 1 0
REF: အဲဒါ ပို ဆိုး လာတယ် ။
HYP: အဲဒါ ********* ပိုဆိုး လာတယ် ။
Eval: D S

Negative Error

SOURCE: **ဖြေ ပေး ဟို့ ရှစ် နေလား ။**
Scores: (#C #S #D #I) 5 1 0 1
REF: အဖြေ *** ပေး ဖို့ ရှက် နေသလား ။
HYP: အဖြေ မ ပေး ဖို့ ရှက် နေတာလား ။
Eval: I S

SOURCE: **ဝယ်ရား နူတ်ဆက် သွား ဟု ။**
Scores: (#C #S #D #I) 5 0 1 0
REF: သူမ နူတ်ဆက် မ သွား ဘူး ။
HYP: သူမ နူတ်ဆက် *** သွား ဘူး ။
Eval: D

Where "SOURCE" is the test sentence of Dawei language, "Scores" are operation scores of the Edit Distance [27], "C" is the number of correct words, "S" is the number of substitutions, "D" is the number of deletions, "I" is the number of insertions, "REF" for reference (i.e. Myanmar sentence), "HYP" for hypothesis and "Eval" is the ordered sequence of edit operations.

We found that translation error of male to female vocabulary and vice versa happen between Dawei-Myanmar translation such as "သူမ" ("she" in English) to "သူ" ("he" in English), "သူမကိုယ်သူမ" ("herself" in English) to "သူကိုယ်သူ" ("himself" in English). The second category, paraphrasing errors are really interesting and it is also proved that two language are similar. In our paraphrasing error examples, the meanings of all reference and hypothesis pairs are the same. Some errors are just the difference

TABLE VII: Average WER% for PBSMT, HPBSMT and OSM with word segmentation (about 1,875 sentences test data for Myanmar-Rakhine, about 922 sentences test data for Myanmar-Dawei, and 1,044 sentences test data for Myanmar-Myeik), lower WER is better

src-tgt	PBSMT	HPBSMT	OSM
my-rk	25.89%	25.94%	**25.78%**
my-dw	51.90%	**51.70%**	**51.70%**
my-bk	56.98%	56.70%	**51.18%**

TABLE VIII: The top 10 confusion pairs of PBSMT model for Myanmar-Rakhine machine translation with word segmentation

Freq	Reference ==> Hypothesis
15	ပါ။ ==> ။
13	ငါ ==> ကျွန်တော်
12	ရှိ့ ==> သူရှိ့
12	အကျွန် ==> ကျွန်တော်
10	ကို ==> ယင်းချင့်ကို
10	လား==> ပါလား
9	နှန့် ==> နှန့်
9	လိမ့်မေ ==> လိမ့်မယ်
9	လေး။ ==> ။
8	ကတ်တေ ==> ကတ်ရေ

TABLE IX: The top 15 confusion pairs of OSM model for Dawei-Myanmar machine translation with word segmentation

Freq	Reference ==> Hypothesis
16	သူမ ==> သူ
14	ခင်ဗျား ==> မင်း
9	ပါတယ် ==> တယ်
8	ပါဘူး ==> ဘူး
7	သလဲ ==> တယ်
5	ဘာတွေ ==> ဘာ
5	မင်းကို ==> ကို
5	မလား ==> မှာလား
5	လား ==> သလား
5	အဲ့ဒါကို ==> ကို
4	ခဲ့ဘူး ==> ဘူး
4	ဘူးလား ==> ရှိလား
4	မင်းရဲ့ ==> မင်း
4	လဲ ==> သလဲ
4	သူ ==> သူမ

between the formal (polite form) and informal written form such as "ကြပါတယ်" (polite form of ending phrase "ကြတယ်" in Myanmar conversation) and "ကြတယ်". One of the possible reasons for the word segmentation errors is inconsistent word segmentation of human translators such as "ကားမောင်း" and "ကား မောင်း" ("drive a car" in English). We also found that one more frequent translation errors between Dawei-Myanmar and Myanmar-Dawei machine translation is changing into negative form (e.g. "အဖြေပေး" ("to answer" in English) and "အဖြေမပေး" ("no answer" in English).

C. Error Analysis for Myeik Language

From our studies, the top 12 confusion matrix for Myanmar-Myeik OSM machine translation (with word segmentation) can be seen in Table X.

TABLE X: The top 15 confusion pairs of OSM model for Myanmar-Myeik machine translation with word segmentation

Freq	Reference ==> Hypothesis
45	ိ ==> ကို
35	မင်း ==> နင်
23	ကို ==> ိ
15	သူ ==> ဒယ်ကောင်မယ်
7	သလဲ ==> တယ်
14	မင့် ==> နင်
12	ငါ ==> ကျွန်တော်
12	နင် ==> ခင်ဗျား
12	လဲ ==> ရဲ
5	အဲ့ဒါကို ==> ကို
11	သွား ==> သော
8	ဝ ==> ရ

We also made manual error analysis on translated outputs of the best OSM model, and we found that dominant errors are different in sentence level. We will introduce four frequent error patterns and they are "Male-Female Vocabulary Error", "Paraphrasing Error", "Word Segmentation Error" and "Negative Error". The followings are some example translation mistakes for each category:

Male-Female Vocabulary Error

SOURCE: **သူမ က သူ့ ကို အပြစ်တင် တယ် ။**
Scores: (#C #S #D #I) 3 3 0 1
REF: ****** ဒယ်ကောင်မယ် ဟ သူ့ဝို့ အပြစ်တင် ရယ် ။
HYP: သူ က သူ့ ကို အပြစ်တင် ရယ် ။
Eval: I S S S

SOURCE: **အဲ့ဒါ ကို သူမ မှတ်မထား ဘူးလား ။**
Scores: (#C #S #D #I) 3 2 0 1
REF: ********** ဒယ်စာဝို့ ဒယ်ကောင်မယ် မှတ်မထား ရလား ။
HYP: ဒယ်ဇာ ကို သူ မှတ်မထား ရလား ။
Eval: I S S

SOURCE: **သူမ အရမ်း စိတ်အားထက်သန် နေတယ် ။**
Scores: (#C #S #D #I) 2 3 0 0
REF: သူလေ တအား စိတ်အားထက်သန် နေရယ် ။
HYP: ဒယ်ကောင်မယ် အလွန် စိတ်အားထက်သန် ရယ် ။
Eval: S S
S

Paraphrasing Error###

SOURCE: **အကြင်နာ ရော ရှိ ရဲ့လား ။**
Scores: (#C #S #D #I) 3 2 0 0
REF: အကြင်နာ ကော ရှိ ရယ်ပဲ့လား ။
HYP: အကြင်နာ ရော ရှိ ပဲ့လား ။
Eval: S S

SOURCE: **သူ့ကိုသူ အားမပေး ချင်ဘူး ဟုတ်လား ။**
Scores: (#C #S #D #I) 2 3 1 1
REF: ********** သူဝိုသူ အားမပေး ရမော် ဟုတ် ဝယ်မှန်း ။
HYP: သူ့ကို သူ အားမပေး ********** ချင်ရမော် ဟုတ်ဝယ်လား ။
Eval: I S D S S

SOURCE: **အထ္ထုပတ္တိ တွေ ဘယ်နားမှာ တွေ့ နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ ပါလား ။**
Scores: (#C #S #D #I) 3 5 0 1
REF: အထ္ထုပတ္တိ ဒေ ********** ဘယ်နားမှာ တွေ့နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ နိုင်လား ။
HYP: အထ္ထုပ္ပတ္တိ ဒေ ဘယ်မှာ တွေ့နိုင် ရယ် ကျေးဇူးပြုပြီး ပြော ပြ ။
Eval: S I S S S S

Word Segmentation Error

SOURCE: **ခင်ဗျား အဲ့ဒါ ကို ချီးကျူးချင်ချီးကျူး မချီးကျူး ချင်နေ ။**
Scores: (#C #S #D #I) 3 3 0 0
REF: မင့် အဲဇာဝို ချီးကျူးချင်ချီးကျူး မချီးကျူး ချင်နေ ။
HYP: မင့် အဲဇာဝို ချီးကျူး ချင်ချီးမွမ်း မချီးမွမ်းချင်နေ ။
Eval: S S

SOURCE: **သူမ ကို တသက်လုံး ခွဲ သွား မှာ မ ဟုတ် ဘူး ။**
Scores: (#C #S #D #I) 4 3 3 0
REF: ဒယ်ကောင်မငယ် ဝို တသက်လုံး ခွဲ သော မှာ မ ဟုတ် ဝ ။
HYP: ဒယ်ကောင်မငယ် ကို တသက်လုံး ခွဲ ********** ********** *** သွားမှာ ဟုတ်ဝ ။
Eval: S D D D S
S

Negative Error###

SOURCE: **သူမ ငို မှာ မ ဟုတ် ဘူး ။**
Scores: (#C #S #D #I) 2 2 3 0
REF: သူ ငို မှာ မ ဟုတ် ဝ ။
HYP: ဒယ်ကောင်မငယ် ငို ********** *** ********** ဟုတ်ဝ ။
Eval: S D D D S

SOURCE: **သူမ စကား မ ပြော ဘူး ။**
Scores: (#C #S #D #I) 2 2 2 1
REF: ********** ဘယ်ဒယ်ကောင်မငယ် စကား မ ပြော ဘူး ။
HYP: အပြင်မှာ ဘယ်ဒယ်ကောင်မငယ် ********** *** စကားပြော ဟုတ်ဝ ။
Eval: I D D S S

SOURCE: **ခင်ဗျား အတင်းဝင် ရမယ် မ ဟုတ် လား ။**
Scores: (#C #S #D #I) 4 1 0 2
REF: ခင်ဗျား အတင်းဝင် ရမယ် *** ********** ဟုတ်ဝ ။
HYP: ခင်ဗျား အတင်းဝင် ရမယ် မ ဟုတ် ဝ ။
Eval: I I S

We found that translation error of male to female vocabulary and vice versa happen between Myanmar-Myeik translation such as "ဒယ်ကောင်မငယ်"("she" in English) to "သူ"("he" in English). The second category, paraphrasing errors are really interesting and it is also proved that two language are similar. In our paraphrasing error examples, the meanings of all reference and hypothesis pairs are the same. Some errors are just the difference between the formal (polite form) and informal written form such as "ရှိရယ်ပဲ့လား"(polite form of ending phrase "ရှိပဲ့လား" in Myeik conversation) and "ရှိလား". One of the possible reasons for the word segmentation errors is inconsistent word segmentation of human translators such as "ချီးကျူးချင်ချီးကျူး" and "ချီးကျူး ချင်ချီးကျူး"("admirably" in English). We also found that one more frequent translation errors between Myeik-Myanmar and Myanmar-Myeik machine translation is changing into negative form (e.g. "စကားပြော"("to speak" in English) and "စကားမပြော"("no speaking" in English).

IX. Conclusion

This work contributes the first Statistical Myanmar Dialect Machine Translation Systems. We used the 18K Myanmar-Rakhine parallel corpus, 9K Myanmar-Dawei parallel corpus and 10K Myanmar-Myeik parallel corpus that we constructed to analyze the language similarity and machine translation performance by applying three existing SMT techniques between standard Myanmar and Myanmar dialacts. We proved that higher BLEU and RIBES scores can be achieved for Rakhine-Myanmar, Dawei-Myanmar and Myeik-Myanmar language pairs even with the limited parallel data. We also found that syllable segmentation provide better machine translation performance than word segmentation unit. The experimental results show that Operational Sequence Model (OSM) is the best model for machine translation between Myanmar langauge and it's dialects. We also present detail analysis on confusion pairs of our current machine translation systems for Myanmar dialects. In the near future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as PaOh and Danu.

Acknowledgment

We would like to thank the following people for their time, effort for translation of Myanmar language to three dialect languages and all the help they gave for our long-term project (2017-2020):

A. for Myanmar-Rakhine parallel corpus development

We would like to express our gratitude to U Oo Hla Kyaw (Ba Gyi Kyaw), Editor of the Rakhine Newspaper for valuable advices. We also thank Mg Than Htun Soe (Computer University Sittwe Students' Union), Mg Htet Myart Kyaw (Computer University Sittwe Students' Union) and Ma Oo Moe Wai (Computer University Sittwe) for their translation of Myanmar language corpus into Rakhine and answering our various questions. Last but not least, we would like to thank U Zaw Tun (Prorector, University of Computer Studies Sittwe) for all the help and support during our stay at University of Computer Studies Sittwe.

B. for Myanmar-Dawei parallel corpus development

We would like to thank U Aung Myo (Leading Charge, Dawei Ethnic Organizing Committee, DEOC) for his advice especially on writing system of Dawei language with Myanmar characters. We are very greatful to Daw Thiri Hlaing (Lecturer, University of Computer Studies Dawei) for her leading the Myanmar-Dawei Translation Team. We would like to thank all students of Myanmar-Dawei translation team namely, Aung Myat Shein, Aung Paing, Aye Thiri Htun, Aye Thiri Mon, Htet Soe San, Ming Maung Hein, Nay Lin Htet, Thuzar Win Htet, Win Theingi Kyaw, Zin Bo Hein and Zin Wai for translation between Myanmar and Dawei sentences. Last but not least, we would like to thank Daw Khin Aye Than (Prorector, University of Computer Studies Dawei) for all the help and support during our stay at University of Computer Studies Dawei.

C. for Myanmar-Myeik parallel corpus development

We would like to thank all students of Myanmar-Myeik translation team namely, Aung Win Htut, Aung Thurin Tun, Nandar Win, Myat Hein Tun, Aye Thet Moe, Yadanar Moe, Paing Paing Tun, Shwe Yi Oo, Ei Ei Hiwe, Hnin Sett Pwint Paing, Zaw Zaw Aung, Zin Pwint Htwe and Hnin Wutyi Oo for translation between Myanmar and Myeik sentences. Last but not least, we would like to thank Daw Thandar Win (Prorector, University of Computer Studies Myeik) for all the help and support during our stay at University of Computer Studies Myeik.

References

[1] Koehn, Philipp and Och, Franz Josef and Marcu, Daniel, "Statistical phrase-based translation," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 2003, pp. 48–54.

[2] Koehn, Philipp and Hoang, Hieu and Birch, Alexandra and Callison-Burch, Chris and Federico, Marcello and Bertoldi, Nicola and Cowan, Brooke and Shen, Wade and Moran, Christine and Zens, Richard and Dyer, Chris and Bojar, Ondřej and Constantin, Alexandra and Herbst, Evan, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 2007, pp. 177–180.

[3] Koehn, Philipp, "Europarl: A parallel corpus for statistical machine translation," Conference Proceedings: the tenth Machine Translation Summit, 2005, pp. 79–86.

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

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

[6] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Philadelphia, Pennsylvania, 2002, pp. 311–318

[7] Isozaki, Hideki and Hirao, Tsutomu and Duh, Kevin and Sudoh, Katsuhito and Tsukada, Hajime, "Automatic evaluation of translation quality for distant language pairs," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 944-952.

[8] Win Pa Pa,Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages," 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop), 09-12 May, 2016, Yogyakarta, Indonesia, Procedia Computer Science, Volume 81, 2016,pp. 250–257.

[9] 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.

[10] Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas and Kamel Smaili, "Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus," oin Proc. of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015, pp. 26-34.

[11] Neubarth Friedrich, Haddow Barry, Huerta Adolfo Hernandez and Trost Harald, "A Hybrid Approach to Statistical Machine Translation Between Standard and Dialectal Varieties," Human Language Technology, Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznan, Poland, December 7-9, 2013, Revised Selected Papers, pp .341–353.

[12] Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat and Michael Baeriswyl, "Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German," CoRR journal, volume (abs/1710.11035), 2017.

[13] John Okell, "Three Burmese Dialects," In David Bradley (ed.), Papers in Southeast Asian Linguistics No. 13: Studies in Burmese Languages, 1995, pp. 1–138

[14] Pe Maung Tin, "The dialect of Tavoy", Journal of the Burma Research Society 23, 1933, pp. 31-46.

[15] Lucia Specia,, "Tutorial, Fundamental and New Approaches to Statistical Machine Translation," International Conference Recent Advances in Natural Language Processing, 2011

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

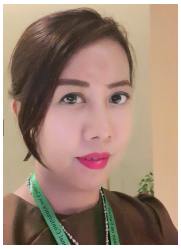
[17] Durrani, Nadir and Schmid, Helmut and Fraser, Alexander, "A Joint Sequence Translation Model with Integrated Reordering," in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, Portland, Oregon, pp. 1045-1054.

[18] 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.

[19] 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.

[20] Och Franz Josef and Ney Hermann, "Improved Statistical Alignment Models," in Proceedings of the 38th Annual Meeting on

Association for Computational Linguistics, Hong Kong, China, 2000, pp. 440-447.
[21] Tillmann Christoph, "A Unigram Orientation Model for Statistical Machine Translation," in Proceedings of HLT-NAACL 2004: Short Papers, Stroudsburg, PA, USA, 2004, pp. 101-104.
[22] Heafield, Kenneth, "KenLM: Faster and Smaller Language Model Queries," in Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT '11, Edinburgh, Scotland, 2011, pp. 187-197.
[23] Chen Stanley F and Goodman Joshua, "An empirical study of smoothing techniques for language modeling," in Proceedings of the 34th annual meeting on Association for Computational Linguistics, 1996, pp. 310-318.
[24] Och Franz J., "Minimum error rate training in statistical machine translation," in Proceedings of the 41st Annual Meeting n Association for Computational Linguistics – Volume 1,Association for Computer Linguistics, Sapporo, Japan, July, 2003, pp.160-167.
[25] 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
[26] (NIST) The National Institute of Standards and Technology, Speech recognition scoring toolkit (SCTK), version: 2.4.10, 2015
[27] Miller, Frederic P. and Vandome, Agnes F. and McBrewster, John, Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), "String Metric, Damerau Levenshtein Distance, Spell Checker, Hamming Distance", ISBN: 6130216904, 9786130216900, Alpha Press, 2009
[28] Armstrong, Lilias E. and Pe Maung Tin, A Burmese Phonetic Reader. London: University of London Press, 1925
[29] Bradley, David. 1982. Register in Burmese. (In) D. Bradley (ed.) Papers in South-East Asian Linguistics No. 8: Tonation. Pacific Linguistics Series "No. 62, pp. 117-132
[30] Khin Pale, A study of Myeik daily vocabulary, B.A. term paper, Mawlamyaing University, Myanmar, 1974
[31] Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the International Conference on Spoken Language Processing, volume 2, pages 901–904, Denver



Thazin Myint Oo is an Associate Professor of Faculty of Computer Science, University of Computer Studies Yangon (UCSY), Myanmar and also a Lab member of Natural Language Processing Lab, UCSY. She is currently pursuing her Ph.D. studies in Machine Translation of Myanmar Dialects.

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 and Speech Science Research Lab., Waseda University, Japan. He is actively co-supervising/supervising undergrad, masters' and doctoral students of several universities including MTU, UCSM, UCSY, UTYCC and YTU.

Khin Mar Soe is currently working as a Professor of Natural Language Processing Lab and Faculty of Computer Science, Unversity of Computer Studies Yangon (UCSY), Myanmar. She is also a head of Research and Development. Her research interest in Artificial intelligence, Natural Language Processing and Machine Translation.

Thepchai Supnithi received the B.S. degree in Mathematics from Chulalongkorn University in 1992. He received the M.S. and Ph.D. degrees in Engineering from the Osaka University in 1997 and 2001 , respectively. He is currently head of language and semantic research team artificial intelligence research unit, NECTEC, Thailand.