Statistical Machine Translation between Myanmar (Burmese) and Dawei (Tavoyan)

Thazin Myint Oo
UCSY, Myanmar
thazinmyintoo@ucsy.edu.mm

Ye Kyaw Thu
NECTEC, Thailand
ka2pluskha2@gmail.com

Khin Mar Soe
UCSY, Myanmar
khinmarsoe@ucsy.edu.mm

Thepchai Supnithi
NECTEC, Thailand
thepchai.supnithi@nectec.or.th

Abstract

This paper contributes the first evaluation of the quality of statistical machine translation (SMT) between Myanmar (Burmese) and Dawei (Tavoyan). We also developed a Myanmar-Dawei parallel corpus (around 9K sentences) based on the Myanmar language of ASEAN MT corpus. 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 (OSM). In addition, two types of segmentation were studied: word and syllable segmentation. The results show that all three statistical machine translation approaches give comparable BLEU and RIBES scores for both Myanmar to Dawei and Dawei to Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three SMT approaches for both word and syllable segmentation.

1 Introduction

Our main motivation for this research is to investigate SMT performance for Myanmar (Burmese) and Dawei (Tavoyan) language pair. The Dawei (Tavoyan) language is closely related to Myanmar (Burmese) language and it is often considered as dialect of Myanmar language. The state-of-the-art techniques of statistical machine translation (SMT) (Koehn et al., 2003). This demonstrate good performance on translation of languages with relatively similar word orders (Koehn, 2005). To date, there have been some studies on the SMT of Myanmar language. (Thu et al., 2016) 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) (Chiang, 2007) approach gave the highest translation quality in terms of both the BLEU (Papineni et al., 2002) and RIBES scores (Isozaki et al., 2010). Win Pa Pa et al (2016) (Pa et al., 2016) presented the first comparative study of five major machine translation approaches applied to low-resource languages. Phrase-based statistical machine translation (PBSMT), HPBSMT, tree-to-string (T2S), string-to-tree (S2T) and operation sequence model (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 both Dawei and Myanmar languages and thus we cannot apply S2T and T2S approaches for Myanmar-Dawei language pair. 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 (Thu et al., 2015). Relating to Myanmar langauge dialects, Thazin Myint Oo et al. (2018) (Oo et al., 2018) contributed the first PBSMT, HPBSMT and OSM machine translation evaluations between Myanmar and Rakhine. The experiment was used the 18K Myanmar-Rakhine parallel corpus that constructed to analyze the behavior of a dialectal Myanmar-Rakhine machine translation. The results showed that higher BLEU (57.88 for Myanmar-Rakhine and 60.86 for Rakhine-Myanmar) and RIBES (0.9085 for Myanmar-Rakhine and 0.9239 for Rakhine-Myanmar) scores can be achieved for Rakhine-Myanmar language pair even with the limited data. Based on the experimental results of previous works, in this paper, the machine translation experiments between Myanmar and Dawei were carried out using PBSMT, HPBSMT and OSM.

2 Related Work

Karima Meftouh et al. built PADIC (Parallel Arabic Dialect Corpus) corpus from scratch, then conducted experiments on cross dialect Arabic machine translation (Meftouh et al., 2015) 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 between the two varieties (Neubarth et al., 2016). 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 (Honnet et al., 2018). 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.

3 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 or Tavoyan 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) and Okell (1995)(OKELL, 1995). 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. Tavoyan retains /-l-/ medial that has since merged into the /-j-/ medial in standard Burmese and can form the following consonant clusters: /gl-/, /kl-/, /k^{h}l-/, /bl-/, /pl-/, /p^{h}l-/, /ml-/, /m̥l-/. Examples include “မ္လေ” (/mlè/ → Standard Burmese /mjè/) for “ground” and “က္လောင်း” (kláʊɴ/ → Standard Burmese tʃáʊɴ/) for “school”. [4] 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 Tavoyan. In the Tavoyan 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” (/p^{h}a̰ òu/) and “မိစု”(/mḭ òu/) respectively. Moreover, the honorific “နှောင်” (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(Tavoyan) 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” is 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 “ဟှ (ရ)” or “ဟန်း” (“No” in English). Myanmar adverb word “သိပ်, အလွန်, အလွန့်အလွန်”(very, extremely) is used as “ရရာ, ရမိရရာ, ပြင်း”. 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 “သစ်ခတ်က္လား” (“ကျားသစ်” in Myanmar language, “leopard” in English). The followings are some example parallel sentences of Myanmar (my) and Dawei(dw):

dw: သယ်ဝယ်သား က လှ ပြင်း ဟှယ် ။
my: ဒီကောင်မလေး က လှ လွန်း တယ် ။
(“The girl is so beautiful” in English)

dw: လတ်ဖတ်ရယ် က ရှို ပြင်း ဟှယ် ။
my: လက်ဖက်ရည် က ချို လွန်း တယ် ။
(“The tea is so sweet” in English)

dw: ကောန်သား က္လောန်း မှန်းမှန် သွား ဟှယ် ။
my: ကောင်လေး ကျောင်း မှန်မှန် တက် တယ် ။
(“The boy goes to school regularly” in English)

4 Methodology

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

4.1 Phrase-Based Statistical Machine Translation

A PBSMT translation model is based on phrasal units (Koehn et al., 2003). 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 (Specia, 2011). 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.

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

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

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

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.

[X] [X] ကို တောပန် [X] ||| [X] [X] ကို တောင်းပန် [X]
[X] [X] ကို တွန်းပန် [X] ||| [X] [X] မ တောင်းပန် [X]
[X] [X] ကို တွန်းပန် ဟှလား [X] ||| [X] [X] မ တောင်းပန် ဘူးလား [X]
[X] [X] ကို တွိ [X] ||| [X] [X] ကို တွေ့ [X]
[X] [X] ကို တွိ ဟှို့ [X] ||| [X] [X] ကို တွေ့ ဖို့ [X]

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

4.2 Hierarchical Phrase-Based Statistical Machine Translation

The hierarchical phrase-based SMT approach is a model based on synchronous context-free grammar (Specia, 2011). The model is able to be learned from a corpus of unannotated parallel text. The advantage this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word re-ordering process. The re-ordering is represented explicitly rather than encoded into a lexicalized re-ordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to language pairs that require long-distance re-ordering during the translation process (Braune et al., 2012). Some examples of hierarchical phrase based grammar between Dawei and Myanmar phrases are shown in Figure 1.

4.3 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 (Durrani et al., 2011) (Durrani et al., 2015). 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)

5 Experiment

5.1 Corpus Statistics

We used 9,000 Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus (Prachya and Thepchai, 2013), 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). Manual Translation into Rakhine Language was done by native Rakhine students from two Myanmar universities and the translated corpus was checked by the editor of Rakhine newspaper. Word segmentation for Rakhine was done manually. 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.

5.2 Word Segmentation

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 is already segmented, we have to consider some rules for manual word segmentation of

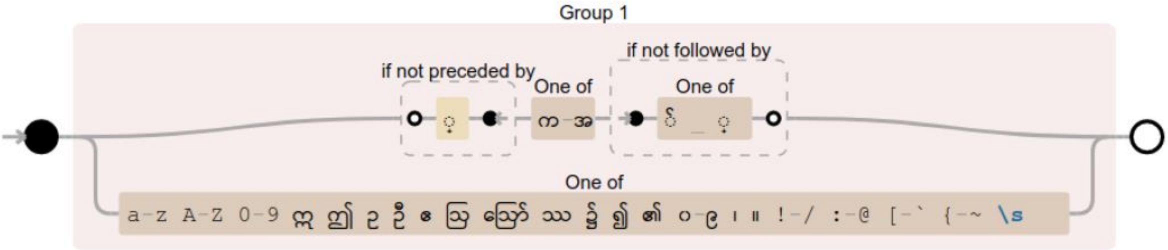

Figure 2: Visualizaiton of syllable breaking with regular expression for Myanmar language

Dawei sentences. We defined Dawei "word" to be meaningful units and affix, root word and suffixe(s) are separated such as "စား ဟုယ်", "စားပီးဟုယ်", "စား ဖို့ဟုယ်". 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 "ဖွန်သားဒေ" (shrimp) is segmented as two words "ဖွန်သား" 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 "ရှိုးခိုသီးတစ်လုံး" ("papaya" in English) is segmented as "ရှိုးခိုသီး 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), "ပြင်း" ("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: ဖွန်သားဒေရရာရှားဟုယ်၊ငါးဖောင်းသားဘဲ့ဝယ်လာရဟုယ်။

Word Segmented Dawei sentence:
dw: ဖွန်သား ဒေ ရရာ ရှား ဟုယ် ၊ ငါးဖောင်းသား ဘဲ့ ဝယ် လာရဟုယ် ။

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 "လာရဟုယ်" are also segmented as two words "ဝယ် လာရဟုယ်" ("bought" in English)

5.3 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:

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

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-kyawthu/sylbreak). The visualization of the syllable breaking based on the RE for Myanmar language is as shown in Figure 2. In our experiments, we used RE based Myanmar syllable segmentation tool named "sylbreak". The following is an example of syllable segmentation for a Dawei sentence in our corpus and the meaning is "You are cute."

Unsegmented Dawei sentence:
dw: နန်ရှစ်ဇရာကွန်းဇမား။

Syllable segmented Dawei sentence:
dw: နန် ရှစ် ဇ ရာ ကွန်း ဇ မား ။

5.4 Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit (Koehn et al., 2007) 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++

src-tgt	PBSMT	HPBSMT	OSM
dw-my	29.143 (0.82286)	29.09 (0.82203)	**29.563 (0.82369)**
my-dw	21.575(0.62624)	21.697 (0.78651)	**21.701 (0.78667)**

Table 1: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using word segmentation

src-tgt	PBSMT	HPBSMT	OSM
dw-my	60.788 (0.94613)	60.472 (0.94476)	**63.221 (0.94825)**
my-dw	44.8 (0.91601)	45.441 (0.91496)	**45.584 (0.91550)**

Table 2: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using Syllable Segmentation

(Och and Ney, 2000). The alignment was symmetrize by grow-diag-final and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option (Tillmann, 2004). We use KenLM (Heafield, 2011) for training the 5-gram language model with modified Kneser-Ney discounting (Chen and Goodman, 1996). Minimum error rate training (MERT) (Och, 2003) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1). We used default settings of Moses for all experiments.

6 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) (Papineni et al., 2002) and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) (Isozaki et al., 2010). The BLEU score measures the precision of n-gram (over all n $\leq$ 4 in our case) with respect to a reference translation with a penalty for short translations (Papineni et al., 2002). 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.

7 Results and Discussion

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Table 1. Bold numbers indicate the highest scores among three SMT approaches. The RIBES scores are inside the round brackets. Here, "my" stands for Myanmar, "dw" stands for Dawei, "src" stands for source language and "tgt" stands for target language respectively. 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 1. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Dawei and Dawei-Myanmar machine translations. Interestingly, the BLEU and RIBES score of all three methods are comparable performance. Our results with current parallel corpus indicate that Dawei to Myanmar machine translation is better performance (around 8 BLEU and 0.03 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 2. Our results with syllable segmentation also indicate that Dawei to Myanmar machine translation is better performance (around 17 BLEU and 0.03 RIBES score higher) than Myanmar to Dawei translation direction. As we expected, generally, machine translation performance of all three SMT approaches between Myanmar and Dawei languages with limited parallel corpus achieved suitable scores for both BLEU and RIBES. The reason is that as we mentioned in Section 3, the two languages, Myanmar and Dawei are close languages. We assume that long distance reordering is relatively rare and only local reordering is enough for the Myanmar-Dawei language pair. We can expect that we can increase these scores higher than current results by increasing the corpus size in the near future.

8 Error Analysis

We also used the SCLITE (score speech recognition system output) program from the NIST scor-

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	သူ့ ==> သူမ

Table 3: The top 15 confusion pairs of OSM model for Dawei-Myanmar machine translation with word segmentation

ing toolkit SCTK version 2.4.10 for making dynamic programming based alignments between reference and hypothesis strings for detail analysis on translation errors. From our studies, the top 15 confusion matrix for Dawei-Myanmar OSM machine translation (with word segmentation) can be seen in Table 3. 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 ဟှို ဖြင် 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: **သူ့ကိုယ်သူ သိ ဟှယ် ။**
Scores: (#C #S #D #I) 3 1 0 0
REF: သူမကိုယ်သူမ သိ ပါတယ် ။
HYP: သူ့ကိုယ်သူ သိ ပါတယ် ။
Eval: S

Paraphrasing Error

SOURCE:**ငှား ဟှားဟှို အီ လေ ။**
Scores: (#C #S #D #I) 4 1 0 0
REF: ငှားရမ်း ထားတဲ့ အိမ် တွေ ။
HYP: ငှား ထားတဲ့ အိမ် တွေ ။
Eval: S

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

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

SOURCE: **သူဟှ ရတိုင်း လှ 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: **အဲဝယ်ဟှား ကားမွန်း ဟှို့မှုဝလား ။**
Scores: (#C #S #D #I) 4 1 1 0
REF: သူမ ကား မောင်း မှာ မဟုတ်ဘူးလား ။
HYP: သူမ ********* ကားမောင်း မှာ မဟုတ်ဘူးလား ။
Eval: D S
SOURCE: **အယ်မိုဇာ ပိုဆိုး လာဟှယ် ။**
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 ဟှို့ 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 နှုတ်ဆက် သွား ဟှ ။**
Scores: (#C #S #D #I) 5 0 1 0
REF: သူမ နှုတ်ဆက် မ သွား ဘူး ။
HYP: သူမ နှုတ်ဆက် *** သွား ဘူး ။
Eval: D

Where “SOURCE” is the test sentence of Dawei language, “Scores” are operation scores of the Edit Distance (Miller et al., 2009), “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 “သူ့ကိုယ်သူ” (“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 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).

9 Conclusion

This paper contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Dawei and Dawei to Myanmar. We used the 9K Myanmar-Dawei parallel corpus that we constructed to analyze the behavior of a dialectal Myanmar-Dawei machine translation. We also investigated two types of segmentation schemes (word segmentation and syllable segmentation). We showed that well-grounded BLEU and RIBES scores can be achieved for Dawei-Myanmar language pair even with the limited data. In the near future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as Myeik (Beik).

Acknowledgment

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.

References

Fabienne Braune, Anita Ramm, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based smt. In Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT), pages 177--184.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, ACL '96, pages 310--318, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics, 33(2):201--228.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1045--1054, Portland, Oregon, USA. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, and Hinrich Schütze. 2015. The operation sequence Model---Combining n-gram-based and phrase-based statistical machine translation. Computational Linguistics, 41(2):157--186.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 187--197, Edinburgh, Scotland. Association for Computational Linguistics.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing swiss German. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 944--952, Cambridge, MA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In The tenth Machine Translation Summit, pages 79--86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pages 177--180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 48--54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pages 26--34, Shanghai, China.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press.

Friedrich Neubarth, Barry Haddow, Adolfo Huerta, and Harald Trost. 2016. A hybrid approach to statistical machine translation between standard and dialectal varieties. volume 9561, pages 341--353.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 160--167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00, pages 440--447, Stroudsburg, PA, USA. Association for Computational Linguistics.

John OKELL. 1995. Three burmese dialects. Papers in Southeast Asian Linguistics No.13, Studies in Burmese Languages, 13:1--138.

Thazin Myint Oo, Ye Kyaw Thu, and Khin Mar Soe. 2018. Statistical machine translation between myanmar (burmese) and rakhine (arakanese). In Proceedings of ICCA2018, pages 304--311.

Win Pa Pa, Ye Kyaw Thu, Andrew M. Finch, and Eiichiro Sumita. 2016. A study of statistical machine translation methods for under resourced languages. In SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced languages, 9-12 May 2016, Yogyakarta, Indonesia, pages 250--257.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311--318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Boonkwan Prachya and Supnithi Thepchai. 2013. 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.

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

Ye Kyaw Thu, Vichet Chea, Andrew M. Finch, Masao Utiyama, and Eiichiro Sumita. 2015. A large-scale study of statistical machine translation methods for khmer language. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015.

Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita. 2016. A large scale study of statistical machine translation methods for myanmar language. In Proceedings of SNLP2016.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In Proceedings of HLT-NAACL 2004: Short Papers, pages 101--104, Boston, Massachusetts, USA. Association for Computational Linguistics.