Statistical Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)

Thazin Myint Oo[†], Ye Kyaw Thu[∫,ʎ], Khin Mar Soe[†]
[†]Natural Language Processing Lab., *University of Computer Studies, Yangon, Myanmar*
[∫] *Artificial Intelligence Lab., Okayama Prefectural University (OPU), Japan*
[ʎ] *Language and Speech Science Research Lab., Waseda University, Japan*
thazinmyintoo@ucsy.edu.mm, ye@c.oka-pu.ac.jp, khinmarsoe@ucsy.edu.mm

Abstract

This paper contributes the first evaluation of the quality of machine translation between Myanmar (Burmese) and Rakhine (Arakanese). We also developed a Myanmar-Rakhine parallel corpus (around 18K 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). The results show that all three statistical machine translation approaches give higher and comparable BLEU and RIBES scores for both Myanmar to Rakhine and Rakhine to Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three approaches.

1. Introduction

Our main motivation for this research is to investigate SMT performance for Myanmar (Burmese) and Rakhine (Arakanese) language pair. The Rakhine (Arakanese) 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) [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 Rakhine language and thus we cannot apply S2T and T2S approaches for Myanmar-Rakhine 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 [9].

Based on the experimental results of previous works, in this paper, the machine translation experiments 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 [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 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.

3. Rakhine Language

Rakhine (Arakanese) is one of the national eight ethnic groups in the Republic of Myanmar. The Arakan was officially altered to 'Rakhine' in 1989 is located on a narrow coastal strip on the west of Myanmar, 300 miles long and 50 to 20 miles wide. It is separated by central plain by a range of mountain, the "Rakhine Yo-ma," along with the administrative boundary runs today. There are three main dialects corresponding to the five administrative districts of Rakhine division that are Sittwe, Kyauk-phyu, Mrauk-U, Thandwe and Maungdaw. The total population for all countries is nearly about 3,000,000. Rakhine mainly located on Rakhine state and Paletwa township, Chin State in the Republic of Myanmar [13][14].

Although Rakhine language used the script Arakanese or Rakkhawanna Akkhara before at least the 8th century A.D. [15], [16], current Rakhine script is exactly the same with Myanmar script. We would like to present four examples this may be same and different vocabulary between them. Moreover, Arakanese language is generally mutually intelligible with Myanmar language and same word order (i.e. SOV). However, the Arakanese language notably retains on /r/ sound (i.e. "ရ") that has become /j/ sound (i.e. "ယ") in Burmese. And thus " ကြား" ("to hear" in English) and "ကြာ" ("tiger" in English) pronounced differently as "kya" and "kra" in Rakhine language. The followings are some example parallel sentences of Myanmar (my) and Rakhine (rk):

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

my: အဘွား နေ မကောင်းဘူး ။
rk: အဘောင်သျင် နီ မကောင်းပါ ။
("Grandma is not feeling well." in English)

my: ကလေး များ ကစား နေကြတယ် ။
rk: အချေ တိ ကဇော် နီကတ်တေ ။
("Children are playing." in English)

my: ငပလီ ကမ်းခြေ သို့ ကားလမ်း က ဖြစ်ဖြစ် လေကြောင်းလမ်း ကပဲဖြစ်ဖြစ် လာရောက် နိုင်ပါတယ် ။
rk: ငပလီ ကမ်းခြီ ကို ကားလမ်း က ဖြစ်ဖြစ် လီကြောင်းလမ်း ကပဲဖြစ်ဖြစ် လာရောက် နိုင်ပါရေ ။
("Ngapali beach can be reached by land or by air." in English)

In the above examples, the underlined words that have same meaning but have different spellings such as "ဘယ်လောက်လဲ" vs "ဇာလောက်လေး" ("how much?" in English), "အဘွား" vs "အဘောင်သျင်" ("grandmother" in English), "ကလေးများ" vs "အချေတိ" ("children" in English), "ကမ်းခြေ" vs "ကမ်းခြီ" ("beach" in English), "လေကြောင်း" vs "လီကြောင်း" ("air or airline" in English).

Rakhine language is largely monosyllabic and analytic language, with a Subject Object Verb (SOV) word order and uses the Myanmar script. It is considered by some to be a dialect of the Myanmar language, though differs significantly from standard Myanmar language in its vocabulary, including loan words from Bengali, Hindi and English. Comparing with Myanmar language, the speech of the Rakhine language is likely to be closer to the written form. Rakhine language notably retains an /r/ sound that has become /j/ in Myanmar language. Rakhine 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 "ေ" (/i/ sound) in Rakhine language. And thus, for example, the word "dog" in Myanmar language writes "ခွေး"(Khwe) and in Rakhine language writes " ခွီး" (khwii). Similarly, Rakhine pronounce "ဲး" (/e:/) for Myanmar pronunciation of "ိုင်" (/ai/) syllable. And thus, Myanmar word "ပဲဟင်း" (peh-hinn) (pea curry in English) is pronounced "ပေးဟင်း" (pay-hinn) in Rakhine language. Some Pali words are also using in Rakhine language.

For example, the word "guest" of Myanmar monks "အာဂန္တု" (Agantu) is used in normal speech of Rakhine and it will be equal to the word of normal Myanmar people guest, "ဧည့်သည်" (Ai thay). In Summary, there are no grammatical differences and the most significant differences between Rakhine and Myanmar languages are in their pronunciations and their vocabularies.

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 [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 $\hat{\mathbf{e}}$ that maximizes the translation probability $\mathbf{P}(\mathbf{e}|\mathbf{f})$ given the source sentences; mathematically. Here, the source language is French and the target language is an English. The translation of a French sentence **f** into an English sentence **e** is modeled as equation 1.

$$\hat{e} = \boldsymbol{argmax_e} \boldsymbol{P}(\boldsymbol{e}|\boldsymbol{f}) \quad (1)$$

Applying the Bayes' rule, we can factorized the $\mathbf{P}(\mathbf{e}|\mathbf{f})$ into three parts.

$$\boldsymbol{P}(\boldsymbol{e}|\boldsymbol{f}) = \frac{\boldsymbol{P}(\boldsymbol{e})}{\boldsymbol{P}(\boldsymbol{f})} \boldsymbol{P}(\boldsymbol{f}|\boldsymbol{e}) \quad (2)$$

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

$$\boldsymbol{argmax_e P}(\boldsymbol{e}|\boldsymbol{f}) = \boldsymbol{argmax_e P}(\boldsymbol{f}|\boldsymbol{e})\boldsymbol{P}(\boldsymbol{e}) \quad (3)$$

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

4.2. 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 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 [18]. An example of hierarchical phrase-based grammar rules between Myanmar and Rakhine from a HPBSMT model is as follows:

[X][X] ကနေ [X] ||| [X][X] ကနိန် [X]
[X][X] ကနေ [X] ||| [X][X] ကနီ [X]
[X][X] ကနေ [X] ||| [X][X] ကနေ [X]
[X][X] ကနေ [X] ||| [X][X] ကပင်ဆိုကေ [X]
[X][X] ကနေ [X] ||| [X][X] ဂနီ [X]

Here, the Myanmar word "ကနေ" means "from" in English .

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 [19][20]. 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, ဒီမှာ)

Operation 4: Jump Back (1)
Operation 5: Generate (sit, ထိုင်)

5. Experiments

5.1. Corpus Statistics

We used 18,373 Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [21], 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 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.

5.2. Word Segmentation

In both Myanmar and Rakhine 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 Rakhine sentences. We defined Rakhine ‘‘word’’ to be meaningful units and affix, root word and suffixe(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. Similar to Myanmar language, Rakhine plural nouns are identified by following particle. We also put a space between 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 used after number or text number. For example, a Rakhine word “အကြွေစ့နှစ်ခတ်” (‘‘two coins’’ in English) is segmented as “အကြွေစ့ နှစ် ခတ်”. In our manual word segmentation rules, compound nouns are considered as one word and thus, a Rakhine compound word “ဖေ့သာ + အိတ်” (‘‘money’’ + ‘‘bag’’ in English) is written as one word “ဖေ့သာအိတ်” (‘‘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-conditioner in our room, two are out of order.’’

Unsegmented sentence:
အကျွန်ရို့အခန်းထဲမှာဟိရေလီအီးစက်လေးလုံးမှာနှစ်လုံးပျက်နီရေ။

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

In this example, “လီအီးစက်” (‘‘air-conditioner’’ in English) is a compound word of “လီအီး” (‘‘cold air’’ in English) and “စက်” (‘‘machine’’ in English). Two Rakhine words, text number “လေး” and a particle “လုံး” (‘‘four machine’’ in English) are segmented as two words. A root word “ပျက်” and the suffix “နီရေ” are also segmented as two words “ပျက် နီရေ” (‘‘out of order’’ in English).

5.3. Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit [22] 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++ [23]. The alignment was symmetrize by grow-diag-final and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [24]. We use KenLM [25] for training the 5-gram language model with modified Kneser-Ney discounting [26]. Minimum error rate training (MERT) [27] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [22]. 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) [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 $n \leq 4$ in our case) with respect to a reference translation with a penalty for short translations [6]. 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, "rk" stands for Rakhine, "src" stands for source language and "tgt" stands for target language respectively.

Table 1. Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM

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)

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM between Myanmar and Rakhine languages are shown in Table 1. 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.

As we expected, generally, machine translation performance of all three SMT approaches between Myanmar and Rakhine languages achieved higher scores for both BLEU and RIBES. The reason is that as we mentioned in Section 3, the two languages, Myanmar and Rakhine are very close languages. We assume that long distance reordering is relatively rare and only local reordering is enough for the Myanmar-Rakhine language pair.

Table 2. Average BLEU and RIBES scores for PBSMT and OSM with reordering and without reordering

src-trg	PBSMT	OSM
my-rk (without)	57.70 (0.9078)	57.89 (0.9086)
my-rk (reordering)	57.69 (0.9077)	57.89 (0.9086)
rk-my (without)	60.56 (0.9232)	60.86 (0.9239)
rk-my (reordering)	60.57 (0.9232)	60.86 (0.9239)

To see clearly on the requirement of reordering process between Myanmar and Rakhine language pair, we held two more 10-folds cross-validation experiments (for PBSMT and OSM) with and without reordering. The experimental results are shown in Table 2. Looking at the results of Table 2, SMT running of my-rk language pair with reordering and without reordering for both PBSMT and OSM gave approximately the same results. From the results, without reordering or zero distortion limit is not affect the performance of machine translation between Myanmar and Rakhine language pair.

8. Error Analysis

We also used the SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 [28] for making dynamic programming based alignments between reference and hypothesis strings and calculation of Word Error Rate (WER). 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 (2):

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

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. 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%.

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%.

Table 3. Average WER% for PBSMT, HPBSMT and OSM with nearly 1,800 sentences test data (lower is better)

src-tgt	PBSMT	HPBSMT	OSM
my-rk	25.89%	25.94%	25.78%
rk-my	22.46%	22.53%	22.26%

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 3.

From the Table 3, 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 ။

REFERENCE:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ လိုက်လုပ် ကတ်ရေ ။

HYP of PBSMT:
နောက် ခဏချင်းမှာပဲ မျောက်တိ က အလားတူ က လိုက်လုပ် ကတ်ရေ ။
Here, the word highlighted with red color is the extra Rakhine word "Ka".

Table 4. The top 10 confusion pairs of PBSMT model for Myanmar- Rakhine

Freq	Confusion Pair (REF→HYP)
15	ပါ ။ ==> ။
13	ငါ ==> ကျွန်တော်
12	ရှိ ==> သူရှိ
12	အကျွန် ==> ကျွန်တော်
10	ကို ==> ယင်းချင့်ကို

10	လား ==> ပါလား
9	နန့် ==> နန့်
9	လိမ့်မေ ==> လိမ့်မယ်
9	လေး။ ==> ။
8	ကတ်တေ ==> ကတ်ရေ

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.. 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 "နှန့် ==> နှန့်" is occurred because of different typing order. Although they look the same, the typing order of the reference "နှန့်" is "န, ှ and ့" (correct order) and the that of hypothesis "နှန့်" is "န, ှ, ့ and ်". These kind of confusion pairs can be reduced by cleaning of current word segmentation and typing errors of our parallel corpus.

9. Conclusion

This paper contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Rakhine and Rakhine to Myanmar. We used the 18K Myanmar-Rakhine parallel corpus that we constructed to analyze the behavior of a dialectal Myanmar-Rakhine machine translation. We showed that higher BLEU and RIBES scores can be achieved for Rakhine-Myanmar language pair even with the limited data. This paper also present detail analysis on confusion pairs of machine translation between Myanmar-Rakhine and Rakhine-Myanmar. In the future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as Yaw and Dawei.

Acknowledgement

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.

References

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

[2]. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation." in Proc. of ACL, 2007, pp. 177–180.

[3]. P. Koehn, "Europarl: A parallel corpus for statistical machine translation." in Proc. of MT 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 Proc. of SNLP2016, February 10-12, 2016.

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

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

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

[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", 1981, London Oxford University press, Univeristy of London.

[14]. Wikipedia of Arakanese language: https://en.wikipedia.org/wiki/Rakhine_State https://en.wikipedia.org/wiki/Arakanese_language

[15]. E. Forchhammar, Papers on Subjects Relating to the Archaeology of Burma including Arakan, Ran-goon, 1891.

[16]. U San Shwe Bu, "A Brief Note on the Old Capitals of Arakan", in Wai Tun(Ed), Golden Tit-Bit of Arkan,1966.

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

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

[19]. 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.

[20]. 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.

[21]. 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

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

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

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

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

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

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

[28]. (NIST) The National Institute of Stan- dards and Technology. Speech recog- nition scoring toolkit (sctk), version:2.4.10,2015.