Statistical Machine Translation between Myanmar and Myeik

Thazin Myint Oo [1+], Ye Kyaw Thu [2], Khin Mar Soe [1] and Thepchai Supnithi [2]

[1] University of Computer Studies Yangon, Myanmar

[2] National Electronics and Computer Technology Center Thailand

Abstract. This paper contributes the first evaluation of the quality of machine translation between Myanmar and Myeik (also known as Beik) . We also developed a Myanmar-Myeik parallel corpus (around 10K sentences) based on the Myanmar language of ASEAN MT corpus. In addition, two types of segmentation were studied word and syllable segmentation. 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 Myeik and Myeik to Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three approaches. We also found that syllable segmentation is appropriate for translation quality comparing with word level segmentation results.

Keywords: Statistical machine translation, Myanmar language (Burmese), Myeik dialect, Machine translation for dialects, Parallel corpus developing

1. Introduction

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

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 [13]. 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.

Myeik dialect has peculiar characteristics in terms of tonal contours, and voice quality in the tones and vowels. The tone of this dialect, which corresponds to the Standard Burmese creaky falling tone, has a rising contour and is pharyngealized [14]. Vowels of the syllables corresponding to Standard Burmese stopped syllables are pronounced with a conspicuous creaky phonation. Previous studies have paid little attention to these facts. Tones and his peculiar to this dialect are also described in this paper [15]. 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 [16].

There are some examples of myeik and Myanmar.

bk : မင်း ငါ့ ကို ကြေးပြား ပေး ဝို့ မေ့ နေရယ်လား။
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).

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 e that maximizes the translation probability P(e|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.

$$e = \text{argmax}_e P(e|f) \quad (1)$$

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

$$\text{argmax e } P(e|f) = \text{argmax e } P(f|e)\ P(e) \quad (2)$$

We note that denominator P(e|f) can be dropped because for all translations the probability of the source sentence remains the same. The P(e|f) variable can be viewed as the bilingual dictionary with probabilities attached to each entry to the dictionary (phrase table). The P(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 un-annotated 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 Myeik from a HPBSMT model is as follows:

ငါ တွေးစာ တအား [X] ||| ကျွန်တော့် အတွေးနဲ့ [X]
ငါ တွေးစာ တအား [X][X] [X] ||| ကျွန်တော့် အတွေးနဲ့
ငါ တွေးစာ တအား တူ [X] ||| ကျွန်တော့် အတွေးနဲ့ တူ [X]
ငါ တွေးစာ တအား တူ ရယ် [X] ||| ကျွန်တော့် အတွေးနဲ့ တူ တယ် [X]

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 10K 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 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.

5.2. Word Segmentation

In both Myanmar and Myeik 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 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 suffix "ရယ်". 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). Myeik 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:
ဘာဖြစ်ရိသားကင်းငယ်တွေကိုရိုက်နေရယ်။

Segmented sentence:
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**"** ("are beating" 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:

Syllable := 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). In our experiments, we used RE based Myanmar syllable segmentation tool named “sylbreak”. The following is an example of syllable segmentation for a Myeik sentence in our corpus and the meaning is “Why are you beaten the children?”

Unsegmented Myeik sentence:

bk: ဘာဖြစ်ရိသားကင်းငယ်တွေကိုရိုက်နေရယ်။

Syllable Segmented Myeik sentence:

bk: ဘာ ဖြစ် ရိ သား ကင်း ငယ် တွေ ကို ရိုက် နေ ရယ် ။

5.4. 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++ [22]. The alignment was symmetrize by grow-diag-final and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [23]. We use KenLM [24] for training the 5-gram language model with modified Kneser-Ney discounting [25]. Minimum error rate training (MERT) [26] 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.

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, “bk” stands for Myeik, “src” stands for source language and “tgt” stands for target language respectively.

Table 1: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using word segmentation (Evaluation with syllable unit)

src-tgt	PBSMT	HPBSMT	OSM
bk-my	44.12 (0.87488)	44.07 (0.87513)	**44.33 (0.87531)**
my-bk	33.25 (0.84045)	33.33 (0.83882)	**33.41 (0.83991)**

To compare with syllable results, the translation results were decomposed into their constituent syllables to ensure that the results are cross-comparable. From the results, OSM method achieved the highest BLEU

and RIBES score for both Myanmar-Myeik and Myeik-Myanmar machine translations. Interestingly, 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 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 (see Table 2). Our investigation clearly show that getting the higher scores with syllable segmentation for bi-directional Myanmar to Myeik machine translation.

As we expected, generally, machine translation performance of all three SMT approaches between Myanmar and Myeik languages achieved comparable scores for both BLEU and RIBES. The reason is that as we mentioned in Section 3, the two languages, Myanmar and Myeik are very close languages.

Table 2: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using syllable segmentation

src-tgt	PBSMT	HPBSMT	OSM
bk-my	70.017 (0.95728)	69.894 (0.95656)	**70.545** **(0.95793)**
my-bk	54.606 (0.92213)	54.404 (0.92194)	**55.112** **(0.92315)**

8. Error Analysis

The top 10 confusion pairs of OSM model for for Myeik-Myanmar machine translation with word segmentation are shown in table 3.

Table 3: Top 10 confusion pairs of OSM model for Myeik-Myanmar machine translation with word segmentation

Freq	Confusion Pair (REF→HYP)
45	ို ==> ကို
35	မင်း ==> နှင်
23	ကို ==> ို
15	သူ ==> ဒယ်ကောင်မငယ်
14	မင့် ==> နှင်
12	ငါ ==> ကျွန်တော်
12	နှင် ==> ခင်ဗျား
12	လဲ ==> ရို
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 of Myanmar-Myeik machine translation for each category:

Male-Female Vocabulary Error

SOURCE: သူမ က သူ့ ကို အပြစ်တင် တယ် □
Scores: (#C #S #D #I) 3 3 0 1
REF: ****** ဒယ်ကောင်မငယ် ဟ သူ့ို 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: ********** ဒယ်စာ၀ို ဒယ်ကောင်မငယ် 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: ဒယ်ကောင်မငယ် 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: သူကိုသူ အားမပေး ချင်ဘူး ဟုတ်လား □
Scores: (#C #S #D #I) 2 3 1 1
REF: ********** သူ၀ိုသူ အားမပေး ရမော် ဟုတ် ဝယ်မှန်း □
HYP: သူကို သူ အားမပေး ********** ချင်ရမော် ဟုတ်ဝယ်လား □
Eval: I S D S S

SOURCE: အတ္ထုပတ္တိ တွေ ဘယ်နားမှာ တွေ့ နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ ပါလား □
Scores: (#C #S #D #I) 3 5 0 1
REF: အတ္ထုပတ္တိ ဒေ ********** ဘယ်နားမှာ တွေ့နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ နိုင်လား □
HYP: အတ္ထုပ္ပတ္တိ 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 အဲဇာ၀ို ချီးကျူးချင်ချီးကျူး မချီးကျူး ချင်နေ □
HYP: မင့် အဲဇာ၀ို 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 ၀ို 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

"SOURCE" is the test sentence of Myanmar language, "Scores" are operation scores of the Edit Distance , "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. Myeik sentence), "HYP" for hypothesis (i.e. Myeik sentence) and "Eval" is the ordered sequence of edit operations. We found that translation error of male to female vocabulary and vice versa happen between Myeik-Myanmar 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 languages. 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).

9. Conclusion

This paper contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Myeik and Myeik to Myanmar. We used the 10K Myanmar-Myeik parallel corpus that we constructed to analyze the behavior of a dialectal Myanmar-Myeik machine translation with word and syllable segmentation units. The result get better translation result in syllable translation unit than word level. We showed that higher BLEU and RIBES scores can be achieved for Myeik-Myanmar language pair even with the limited data. This paper also present detail analysis on confusion pairs of machine translation between Myanmar-Myeik and Myeik-Myanmar. In the future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as Yaw and Danu.

10. Acknowledgements

We would like to express our gratitude to 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.

11.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] Wikipedia of Myeik:

https://en.wikipedia.org/wiki/Myeik_dialect

https://en.wikipedia.org/wiki/Myeik,_Myanmar

[14] 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.

[15] John Okell , ”Three Burmese Dialects”, 1981, London Oxford University press, Univeristy of London.

[16] Khin Pale 1974. “A study of Myeik daily vocabulary” , B.A.term paper, Mawlamyaing University, Myanmar

[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] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.