Statistical Machine Translation between Myanmar (Burmese) and Kayah

Zar Zar Linn [†], Ye Kyaw Thu [λ] and Pushpa B. Patil [‡]

Abstract— **This paper contributes the first evaluation of the quality of Statistical Machine Translation (SMT) between Myanmar (Burmese) and Kayah (Kayah Li) languages. We also developed a Myanmar-Kayah parallel corpus (6,590 sentences) based on the Myanmar language of ASEAN MT corpus. The experiments were carried out using three different statistical machine translation approaches: Phrase-based Statistical Machine Translation (PBSMT), Hierarchical Phrase-based Statistical Machine Translation (HPBSMT), and the Operation Sequence Model (OSM). The results show that HPBSMT approach achieves the highest BLEU score for Myanmar to Kayah translation and Operation Sequence Model approach achieves the highest BLEU score for Kayah to Myanmar translation.**

Index Terms—**Statistical Machine Translation (SMT), Phrase-based SMT, Hierarchical phrase-based SMT, Operation Sequence Model, Myanmar-Kayah SMT, under-resourced SMT.**

I. Introduction

Our main motivation for this research is to investigate SMT performance for Myanmar (Burmese) and Kayah (Kayah Li) language pair. Red Karen or Karenni, known in Burmese as Kayah, is a Karen dialect continuum spoken by over half a million Kayah people (Red Karen) in Myanmar [1]. The state-of-the-art techniques of statistical machine translation (SMT) [2], [3] demonstrate good performance on translation of languages with relatively similar word orders [4]. To date, there have been some studies on the SMT of Myanmar language. Ye Kyaw Thu et al. (2016) [5] 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) [6] approach gave the highest translation quality in terms of both the BLEU [7] and RIBES scores [8]. Win Pa Pa et al (2016)[9] 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 Kayah language and thus we cannot apply S2T and T2S approaches for Myanmar-Kayah 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 [10].

Relating to Myanmar language dialects, Thazin Myint Oo et al. (2018) [11] 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. Thazin Myint Oo et al. (2019) also contributed the first SMT evaluations between Myanmar and Dawei (Tavoyan) language pair. The SMT results with developed 9K Myanmar-Dawei parallel corpus showed that higher BLEU (21.70 for Myanmar-Dawei and 29.56 for Dawei-Myanmar) and RIBES (0.78 for Myanmar-Dawei and 0.82 for Dawei-Myanmar) scores achieved with OSM approach [12]. Based on the experimental results of previous works, in this paper, the machine translation experiments between Myanmar and Kayah language pair were carried out using PBSMT, HPBSMT and OSM.

II. Kayah Language

Kayah or Red Karen, this language is a Central Karenic language. It is also the family of the Tibeto Burman Language. Kayah people from Myanmar speak this language. The Kayah people live mostly in Kayah State, Shan State and along Thailand's north western border. In more recent years, Kayah Li people have been given

[†] with Myanmar Institute of Information Technology (MIIT), Myanmar

[λ] with National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA), Thailand

[‡] with BLDEA CET, India

Corresponding Authors: zzlinn@gmail.com and yktnlp@gmail.com

Manuscript received December 21, 2019; accepted March 6, 2020; revised April 23, 2020; published online April 30, 2020.

the prospect to emigrate all over the world from New Zealand and Australia to Finland, the USA and other countries. In Myanmar Census 2014, total population of Kayah state is over 280,000 people. An estimate of the Kayah speaking population is over 100,000 people although there is no recent census data to draw official figures. The reality is that different speech varieties can be found from village to village while the Kayah language has three dialects, western, eastern and northern. In the Western Kayah national script, there are 24 consonants and 9 vowels (plus 9 breathy vowels), 1 diphthong and 3 tones. Twenty-Four Consonants (ꤊ to ꤡ) of Kayah language and it's corresponding sound according to the phonetic symbols are described in Table I [13]. Nine vowels of Kayah are as shown in Table II. Kayah Li numbers (zero to nine: 10 digits) are shown in Table III. Writing the Kayah numbers are very easy. For example, 1 equals ꤁, 10 equals ꤁꤀, 100 equals ꤁꤀꤀, and etc.

Kayah is a tonal language. Two or three of Kayah words may be spelled exactly the same but these words have different meanings based on whether the syllable has a low tone (꤬), mid tone (꤭), and high tone (꤫). Tones only occur on the vowel and are marked directly below it. The three main tone symbols of Kayah are as shown in Table IV.

Word order of Kayah sentence is Subject-Verb-Object (SVO). English word order is the same SVO order. Grammatical order of Kayah language is like English. The followings are the three example Kayah sentences (the first line is written in Kayah, the second and third lines are equivalent meaning in Myanmar and in English languages):

ꤞꤢꤌꤢꤧꤢ꤬ꤦ ꤙꤛꤢ꤬ꤗꤢꤧ ꤗꤢ꤬ ꤊꤢ꤬ꤟꤢ꤬ꤊꤛꤢ꤭ ꤙꤛꤢ꤬ꤗꤢ꤬ꤧꤢꤧ꤬꤯
လွိုင်ကော်မြို့သည် ကယားပြည်နယ်၏ မြို့တော် ဖြစ်သည်။
Loikaw is the capital city of Kayah state.

ꤜꤛꤢꤧ꤬ꤦꤧꤢ꤬ꤊꤧ ꤕꤟꤢꤧ꤬ꤦꤧꤛꤢꤧꤊꤛꤢ꤬ ꤗꤢ꤬ ꤢꤢ꤬ ꤧꤞ ꤊꤢ꤬ꤟꤢ꤬ꤊꤛꤢ꤭꤯
လောပိတ ရေတံခွန်သည် ကယားပြည်နယ်တွင် တည်ရှိသည်။
Law Pi Ta waterfall is situated in Kayah State.

ꤌꤢꤧ꤬ꤊꤢꤢ꤬ ꤕꤟꤢꤧ꤬ꤥꤢꤧ꤬ ꤗꤢ꤬ ꤢꤢ꤬ ꤖꤢꤧ꤬ ꤔꤢꤧ ꤧꤢ꤬ꤗꤛꤢꤧꤊꤢꤢ꤬ ꤊꤜꤢꤧ꤯
ငွေတောင် ဆည် သည် ဒီးမော့ဆို ဈေး အနီး တွင် တည် ရှိ သည် ။
Ngwe Taung Dam is located near Demoso Market.

III. Methodology

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

A. Phrase-Based Statistical Machine Translation (PB-SMT)

A PBSMT translation model is based on phrasal units [2]. 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 [14].

This is another example: $\sum_i^j$.

The phrase translation model is based on noisy channel model. To find best translation $\hat{e}$ 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 into an English sentence is modeled as equation 1.

$$\hat{e} = arg \max_e P(e|f) \tag{1}$$

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

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

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

$$arg \max_e P(e|f) = arg \max_e P(f|e)P(e) \tag{3}$$

We note that denominator $P(f)$ 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 PBSMT paradigm.

IV. Hierarchical Phrase-Based Statistical Machine Translation

The hierarchical phrase-based SMT approach is a model based on synchronous context-free grammar [6]. 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 [15]. An example of hierarchical phrase-based grammar rules

TABLE I
KAYAH CONSONANT

Consonant					
ꤊ k	ꤋ kh	ꤌ g	ꤍ ng	ꤎ s	ꤏ sh
ꤐ zh	ꤑ ny	ꤒ t	ꤓ ht	ꤔ n	ꤕ p
ꤖ ph	ꤗ m	ꤘ d	ꤙ b	ꤚ r	ꤛ y
ꤜ l	ꤝ w	ꤞ th	ꤟ h	ꤠ v	ꤡ c

TABLE II
KAYAH VOWEL

Vowel								
ꤢ a	ꤣ ơ	ꤤ i	ꤥ ô	ꤢꤦ ư	ꤢꤧ e	ꤢꤨ u	ꤢꤩ ê	ꤢꤪ o

TABLE III
KAYAH NUMBER

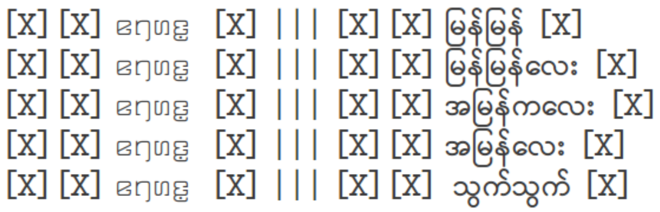
Number									
꤀ 0	꤁ 1	꤂ 2	꤃ 3	꤄ 4	꤅ 5	꤆ 6	꤇ 7	꤈ 8	꤉ 9

TABLE IV
KAYAH TONE MARKERS

Tone Markers		
꤫ high	꤬ low	꤭ medial

TABLE V
SYLLABLE TYPE

Syllable Type	Kayah Li	English Meaning
V	ꤢ꤬	he/she/it
V	ꤥ꤭	drink
CV	ꤒꤣ꤬	one
CV	ꤙꤤ	when
CCV	ꤕꤜꤢꤨ	punch
CCV	ꤊꤚꤢꤨ꤬	firewood
CCCV	ꤊꤜꤛꤢ	road
CCCV	ꤕꤚꤛꤢ꤭	quickly



```
[X] [X] ꤕꤚꤛꤢ꤭ [X] ||| [X] [X] မြန်မြန် [X]
[X] [X] ꤕꤚꤛꤢ꤭ [X] ||| [X] [X] မြန်မြန်လေး [X]
[X] [X] ꤕꤚꤛꤢ꤭ [X] ||| [X] [X] အမြန်ကလေး [X]
[X] [X] ꤕꤚꤛꤢ꤭ [X] ||| [X] [X] အမြန်လေး [X]
[X] [X] ꤕꤚꤛꤢ꤭ [X] ||| [X] [X] သွက်သွက် [X]
```

between Kayah and Myanmar languages from a HPBSMT model is as follows:

Here, the Kayah word "ꤕꤚꤛꤢ꤭" means "quickly" in English.

V. 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 [16], [17]. 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 (generation of a sequence of source and target words), insert gap (insertion of gaps as explicit target positions of reordering operations), jump (forward and backward jump which perform the actual reordering) [26]. The probability of a sequence of operations is defined according to an n-gram model, i.e., the probability of an operation depends on the n-1 preceding operations. Let $O = o_1, \cdots, o_n$ be a sequence of operations as hypothesized by the translator to generate a word-aligned bilingual sentence pair $< F; E; A >$; the model is then defined as:

$$P_{osm}(F, E, A) = P(O_1^J) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, \cdots, o_{j-1}) \tag{4}$$

The following shows an example translation process of English sentence “He does NLP research” into Myanmar language “သူ NLP သုတေသန လုပ်တယ်” with the OSM.

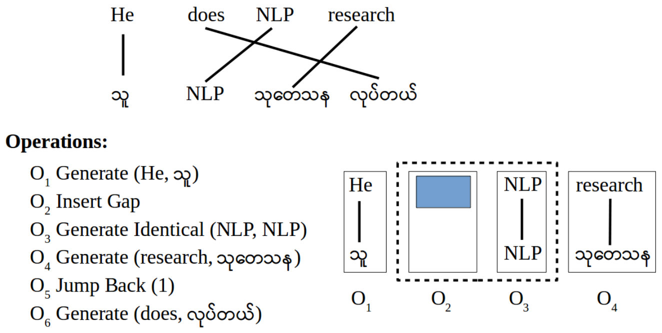


Fig. 1. An example of operation sequence translation

The example shown in figure 1 is deterministically converted to the following operation sequence:

Generate (He, သူ) – Insert Gap – Generate Identical (NLP, NLP), Generate (research, သုတေသန), Jump Back (1), Generate (does, လုပ်တယ်)

VI. 5. Experiments

A. Corpus Statistics

We used 6,590 Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [18], 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 Kayah Language was done by native Kayah students from University of Computer Studies, Loikaw. Kayah language use space between words and there are exactly 42,955 words in total. We used 6,300 sentences for training, 190 sentences for development or tuning process and 100 sentences for evaluation respectively.

B. Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit [19] for training the PBSMT, HPBSMT and OSM statistical machine translation systems. The word segmented source language was aligned with the word segmented target language using GIZA++ [20]. The alignment was symmetrize by grow-diag-final and heuristic [2]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [21]. We use KenLM [22] for training the 5-gram language model with modified Kneser-Ney discounting [24]. Minimum error rate training (MERT) [23] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [19]. We used default settings of Moses for all experiments.

VII. 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) [7] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [8]. 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 [7]. 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.

VIII. Results and Discussion

A. Machine Translation Performance

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Table VI. Bold numbers indicate the highest scores among three SMT approaches. The RIBES scores are inside the round brackets. Here, “my” stands for Myanmar, “ky” stands for Kayah, “src” stands for source language and “tgt” stands for target language respectively.

TABLE VI
BLEU scores for PBSMT, HPBSMT and OSM

src-tgt	PBSMT	HPBSMT	OSM
my-ky	20.58 (0.56)	**22.87** **(0.59)**	20.97 (0.56)
ky-my	13.79 **(0.66)**	12.15 (0.62)	**15.19** (0.62)

From the results, HPBSMT method achieved the highest BLEU and RIBES score for Myanmar-Kayah translation. Although PBSMT achieved the highest RIBES score, OSM achieved the highest BLEU score for Kayah to Myanmar translation. Interestingly, BLEU score results with current parallel corpus indicate that Myanmar to Kayah translation is better performance (around 7 BLEU score) than Kayah to Myanmar translation direction. However, Kayah to Myanmar translation is better performance in terms of RIBES score results (around 7 RIBES score). Based on the highest RIBES score "0.66" for Kayah to Myanmar translation, not only OSM but also PBSMT approach might be applicable to real world translation system.

B. Error Analysis

We also used the SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 [25] 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):

$$WER = \frac{(N_i + N_d + N_s) \times 100}{N_d + N_s + N_c} \tag{5}$$

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.

The WER % of PBSMT, HPBSMT and OSM for Myanmar to Kayah and Kayah to Myanmar translations with 100 open-test sentences are as shown in Table VII. Bold numbers indicate the lowest WER among three SMT approaches.

TABLE VII
AVERAGE WER% FOR PBSMT, HPBSMT AND OSM (LOWER IS BETTER)

src-tgt	**PBSMT**	**HPBSMT**	**OSM**
my-ky	83.9	**82.4**	85.0
ky-my	73.6	73.5	**73.2**

From the Table VII, we found that the lowest WER% for Myanmar to Kayah translation is 82.4% with HPBSMT approach and the lowest WER% for Kayah to Myanmar translation is 73.2% with OSM. These results are inversely proportional to the BLEU scores. 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 translation mistake patterns of Kayah to Myanmar OSM translation are "subject missing", "Post-positional Marker missing" as follows:

### **Subject Missing**
Scores: (#C #S #D #I) 6 0 2 0
REF: ခင်ဗျား ကို ဘာ ကူညီ ပေး ရ မလဲ ။
HYP: ******************** ********* ဘာ ကူညီ ပေး ရ မလဲ ။
Eval: D D

### **Post-positional Marker Missing**
Scores: (#C #S #D #I) 4 1 1 0
REF: ကျွန်တော့် ကို ကူညီ နိုင် မလား ။
HYP: **************************** ကျွန်တော် ကူညီ နိုင် မလား ။
Eval: D S

We also found that although some hypothesis sentences are correct and exactly the same meaning with reference, insertion operation of WER calculation happen by containing a polite form Myanmar word such as "ပါ":

### **Polite-form Myanmar Word Containing**
Scores: (#C #S #D #I) 6 0 0 1
REF: နှစ် ယောက် ခန်း ရှိ ****** သလား ။
HYP: နှစ် ယောက် ခန်း ရှိ ပါ သလား ။
Eval: I

Some translation mistakes of Myanmar to Kayah occurred because of manual word segmentation error in references of Kayah language as follows:

### **Word Segmentation Error in Reference**
Scores: (#C #S #D #I) 2 1 0 1
REF: ꤒꤟꤢ꤬ ꤛꤢ꤬ ****** ꤞꤨꤞꤨꤜꤢꤦꤞꤨ꤯
HYP: ꤒꤟꤢ꤬ ꤛꤢ꤬ ꤞꤨꤞꤨꤜꤢꤦꤞꤨ ꤯
Eval: I S

In the above example, the last word of the reference sentence contained Kayah li punctuation sign Shya "꤯". This punctuation sign should be segmented as one word.

We also found some translated errors are caused by encoding or ASCII based font of Kayah language. The reason is that in our Myanmar-Kayah parallel corpus, Kayah text are written by ASCII encoding based Kayah font named "Karenni Font" and it is a non-Unicode font. And thus all Kayah text are stored as English characters such as a sentence "ꤟꤢ꤬ ꤔꤢꤦ꤬ ꤔꤛꤢꤙꤢ꤬ ꤢꤔꤛꤢ ꤗꤢꤘꤤ꤯" (in English "I can bring it by myself") is stored as "h zKh sky[Fg fgsky lrG /". Here, small and capital letters have different Kayah character mapping and some of the typing mistakes of translators such as "lrG" and "LRG" might occur alignment error. The following are one example of translation error based on encoding:

### Encoding or ASCII Font of Kayah based Error
Scores: (#C #S #D #I) 2 3 0 1
REF: [illegible] [illegible] [illegible] [illegible] *** [illegible]।
HYP: [illegible] BD[illegible]OD[illegible]A[illegible]VD[illegible]D[illegible]P[illegible] B[illegible] [illegible] L[illegible] ।
Eval: S S I S

Some of the errors of machine translation from Kayah to Myanmar are occurred by out-of-vocabulary (OOV) words as follows:

Scores: (#C #S #D #I) 2 2 0 1
REF: [illegible] [illegible] [illegible] **** [illegible]?
HYP: တားပွဲ [illegible] [illegible] [illegible] ?
Eval: S I S

After we made analysis of confusion pairs of each model in details, we found that some of the confusion pairs are relating to word segmentation and typing errors (refer Table VIII and Table IX.

TABLE VIII
The top 10 confusion pairs of HPBSMT model for Myanmar-Kayah

Frequency	Confusion Pair (REF==>HYP)
5	[illegible]।==> ।
5	[illegible]? ==> ?
4	[illegible] ==>
4	[illegible]। ==> ।
3	[illegible]। ==> ।
2	([illegible])[illegible] ==> ।
2	[illegible]। ==> ।
2	[illegible] ==> ကြက်ဥ
2	[illegible]। ==> ।
2	[illegible]। ==> ।

TABLE IX
The top 10 confusion pairs of OSM model for Kayah-Myanmar

Frequency	Confusion Pair (REF==>HYP)
3	ခွက် ==> [illegible]
3	ပါ ==> မယ်
3	မယ် ==> [illegible]।
2	က ==> တဲ့
2	ကြွရောက် ==> က
2	ထပ် ==> [illegible]
2	နေ ==> အတွက်
2	ပါ ==> တယ်
2	ဘူး ==> [illegible]।
2	။ ==> ([illegible])[illegible]

IX. Conclusion

This paper contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Kayah and Kayah to Myanmar. We used our developing 6,590 sentences Myanmar-Kayah parallel corpus that we constructed to see the machine translation performance between Myanmar language and one of it's ethnic languages Kayah. We believe the parallel corpus extension will improve the translation performance between Myanmar and Kayah language pair. In the future we plan to develop a Kayah-ASCII to Kayah-Unicode font converter and hold the new machine translation experiments with Kayah-Unicode sentences.

Acknowledgment

Firstly, we heartily thank of gratitude to ASEAN-India Research Training (AI-RTF), Department of Science and Technology (DST) in India and Federation of Indian Chambers of Commerce and Industry (FICCI) for providing fund for this research work. Our sincere thanks go to Dr. Myar Doo Myar Reh for linking the Kayah language experts. We are very grateful to Mr. Khu Phya Reh, Tutor, History Department, Loikaw University for his patience, translation and necessary instruction of Kayah Language during the period of the paper. We would like to express our special thanks to Mrs. Cember Paw, Kayah Freelance Translator for big efforts of Myanmar to Kayah language translation. Finally, we would like to thank Mr. Khu Syah Reh, Kayah Bible Translation Team, Kayah National Literacy and Culture Committee, for his very useful translation of Myanmar to Kayah language.

References

[1] Wikipedia of Red Karen Language: https://en.wikipedia.org/wiki/Red_Karen_language
[2] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation", in Proc. of HTL-NAACL, 2003, pp. 48–54.
[3] 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, andE. Herbst, "Moses: Open source toolkit for statistical machine translation", in Proc. of ACL, 2007, pp. 177–180.
[4] P. Koehn, "Europarl: A parallel corpus for statistical machine translation", in Proc. of MT summit, 2005, pp. 79–86.
[5] 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.
[6] Chiang, D., "Hierarchical phrase-based translation", Computational Linguistics 33(2), 2007, pp. 201-228.
[7] 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
[8] 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.
[9] 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.
[10] 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.

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

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

[13] Wikipedia of Kayah Li Alphabet: https://en.wikipedia.org/wiki/Kayah_Li_alphabet

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

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

[26] Moses, Statistical Machine Translation System, User Manual and Code Guide, by Philipp Koehn, University of Edinburgh: http://www.statmt.org/moses/manual/manual.pdf

Zar Zar Linn is a Professor of Faculty of Computer Science in Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar. Previously she was a faculty in Universiy of Computer Studies, Mandalay, Universiy of Computer Studies (Myeik), Mandalay, Loikaw and Maw La Myaing from 2001 to 2018. She is a fellow of ASEAN-INDIA Research Training Fellowship (AIRTF), 2018-2019. She was doing research (Machine Translation between Myanmar and Kayah Languages) in India. Her research interests are Part of Speech Tagging, Machine Translation, Natural Language Understanding and Processing, and Data Analysis.

Ye Kyaw Thu is a Visiting Professor of Language & Semantic Technology Research Team (LST), Artificial Intelligence Research Unit (AINRU), National Electronic & Computer Technology Center (NECTEC), Thailand and Head of NLP Research Lab., University of Technology Yatanarpon Cyber City (UTYCC), Pyin Oo Lwin, Myanmar. He is also a founder of Language Understanding Lab., Myanmar and a Visiting Researcher of Language and Speech Science Research Lab., Waseda University, Japan. He is actively co-supervising/supervising undergrad, masters' and doctoral students of several universities including MTU, UCSM, UCSY, UTYCC and YTU.

Pushpa B. Patil received B.E. and M.Tech. from the Department of Computer Science and Engineering, Karnataka University Dharawad (KUD), Dharawad, Visveshwaraya Technological University(VTU), Belagavi, Karnataka, India in the years 1996, 2006 respectively, from 1997-2000, she was worked as lecture in Computer Science Department at MBEs Engineering College, Ambajogai Maharashtra, India. In 2000, she joined as lecturer in Department of Computer Science at BLDEAs V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur, Karnataka, where she is presently holding position of Professor and Head. In 2014, she completed her Ph.D. thesis entitled "Content Based Image Retrieval with Relevance Feedback at research centre SGGS IOT, Nanded affiliated to the university SRTM University, Nanded. Her research interests include image processing, pattern recognition, relevance feedback in content based image retrieva, Natural language processing. She published 24 papers in international journal and conferences.