Word to Word translation between Myanmar (Burmese) and other Ethnic languages

May Myat Myat Khaing, Ye Kyaw Thu,
Thazin Myint Oo, Thet Thet Zin, Nang Aeindray Kyaw,
Hay Man Htun, Thida San, Zun Hlaing Moe, Hnin Aye Thant

***Abstract*— In Myanmar, there are still many challenges for NLP Researchs and Developements. Bilingual lexicon between Myanmar (Burmese) and other Ethnic Languages is one of many challenges. This paper persent word-to-word translation between Myanmar (Burmese) and Ethnic languages extracted from sentence-level parallel corpora. This is the first time that a bilingual or cross-linigual lexicon has been developed based on between Myanmar and other Ethnic languages. The languages pairs between Myanmar (Burmese) and Ethnic languages are constructed for the bilingual or cross-linigual lexicon based on 12 Myanmar (Burmese) and Ethnic Languages. To obtain this dataset, we use a count-based bilingual lexicon extraction model based on the observation that not only source and target words but also source words themselves can be highly correlated [1]. According to the Out-Of-Vocabulary (OOV) rate, the evaluation of word-to-word translation between Myanmar (Burmese) and Ethnic languages is acceptable level for all language pairs. Then we examined the human evaluation for some ethinc language pairs in this paper. Some ethnic language pairs are a level of satisfaction among the ethnic languages where human evaluation was conducted.**

***Index Terms*—Bilingual lexicons, word-to-word Translation, Myanmar (Burmese) and Ethnic languages.**

I. Introduction

THERE are eight main ethnic groups in Myanmar. Some ethinic groups have their own languages. As Myanmar ethnic languages are low resource languages, there are still very challenges in Myanmar NLP Researchs and Developements. The main challenge is the Myanmar Ethnic data sets are rarely for publicly available and thre are still difficulties in the preprocessing steps.

Bilingual lexicons [2] are valuable resources for cross-lingual tasks, including low-resource machine translation [3] [4] and cross-lingual word embeddings [5]. However, it is often difficult to find a large enough set of bilingual lexicons that is freely and readily available across various language pairs [6].

Motivated by the lack of publicly available and high coverage bilingual lexicons for 12 Myanmar (Burmese) and Ethnic Languages, we present word-to-word translation. There are approximately a hundred languages spoken in Myanmar [7]. Among them, word-to-word translation are evaluate for some Ethnic languages based on Myanmar(Burmese) language. The word-to-word translations are constructed based on the Myanmar (my), Biek (bk), Mizo Chin (ch), Jingpho Kachin (kc), Kayah (ky), Mon (mo), western Pwo Kayin (pk), Pa'O (po), Rakhine (rk), Rawang (rw), Shan (sh) and S'gaw Kayin (sk) languages. For these languages, there are 24 language pairs. In this paper, we consider approaches based on (monolingual and cross-lingual) cooccurrence counts: co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE).

The structure of this paper is organized as follows. A brief related work of word-to-word translation systems for different language pairs is presented in the section II. The upcoming section (III) is language introduction and the next section (IV) is a summarization of word-to-word dataset. Section V describes the methodologies used in the word-to-word experiments. The next section VI, we present parallel corpus preparation between Myanmar (Burmese) and other Ethnic languages for word-to-word translation and some word segementation rules for those languages. Section VII presents the used evaluation schemes to determine the performance of word-to-wrod translation. Section VIII discusses the exprimental results accorrding to the evaluation schemes. The summary of word to word translation and futher research are presented in Section IX.

II. Related Work

The paper [8] are extracted automatically Bilingual lexicons for Low resource languages (Spanish-Nahuatl) pairs. Parallel corpora size for these language are available small amount. [8]The proposal is to construct bilingual word vectors from a graph. This graph is generated using translation pairs obtained from an unsupervised word alignment method.

[9]The author state that bilingual dictionary can be build without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way. According to the test results, They can even translate

May Myat Myat Khaing is with the Faculty of Information Science, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar, Corresponding author email: mmyatm@gmail.com

Ye Kyaw Thu is with LST Lab., NECTEC, Thailand and IDRI, CADT, Cambodia, Corresponding author email: yktnlp@gmail.com.

Manuscript received Feb 21, 2022; accepted Aug 30, 2022; revised Oct 11, 2022, published on 31 Oct 2022.

distant language pairs such as English-Russian or English-Chinese.

The research paper [10] listed context-based projection method, the projection of monolingual word embeddings, as well as a method dedicated to identify translations of rare words. The effect of these methods is to experiment the translation of English words on Wikipedia into French Words.

The main purpose of [11] paper is to construct word-aligned parallel corpus to be able in Myanmar-English machine translation. The first three IBM models and EM algorithm are proposed for translation between words in one language and in other language.

In [1], present cross-lingual word translations extracted from sentence-level parallel corpora. The paper provides top-k word translations in 3,564 (directed) language pairs across 62 languages in [12]. The author consider approaches based on (monolingual and cross-lingual) cooccurrence counts: co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE) to build biligual or cross-lingual word translation.

Based on the experimental results of previous works, in this paper, the word-to-word translation for Low resource languages (Burmese and Other Ethnic languages) experiments were carried out based on 12 languages.

III. Languages Introduction

Figure 1 shows the different languages used by different ethnic groups using different colors. For example, for the Burmans, orange color is used; Green color is used for Shan and red color is used for Karen and so on. Some languages have more than one dialect language. However, the image does not keep going to color for the dialect languages. You can get an overview of the ethnic groups in Myanmar and the languages they use by looking at the Figure 1. Each ethnic group has its own language, so better communication is needed for every topic. That's why we proposed this paper.

Although the Burmese language is used as an official language, every Ethnic groups also have their own language and language structure. Similarly, some ethnic groups which have the dialect language have own dialect language structures. There are also ethnic languages with more than one dialect language (For example Burmese, Karen, Kachin and Chin languages). Some ethnic languages are derived from the Burmese script as their own scripts. Some ethnic languages are derived from the Latin script. Word order of some Ethinc languages are not same order.

The alphabet of ethnic languages used in the proposed system; Grammar, Phonology, Vocabulary, Orthography, History, Writing system and Religion are described in the following paragraph. You can also estimate the number of language users and where the language is used. You can also learn the dialect languages of these languages.

You can learn the basics of the Burmese language at [21]. Sentence structure of Burmese are as follow :

English(eg): She knew me.

Burmese(my): သူမ က ကျွန်တော့် ကို သိ ခဲ့တယ် ။

The word order of the Burmese language is generally sub+obj+verbs. [22] explains Myeik language. It is one dialect language of Burmese. you will find the basic nature of the Chin language and sentence structures at [23], [24]. Chin is also a language with many dialect languages. In [25], present the kachin (Jingpho) language.

English(eg): He is a boy.

Kachin(kc): shi gaw la sha langai re le .

Kayah language is one of the proposed ethnic languages in this paper. The nature of language structure can study in the [26], [27]. [28] explain the concept of mon language. The sentence sturcture of mon language is:

English(eg): Will you not bring it for me?

Mon(mo): ဟွံ ကေတ်နင် သွက် အဲ ရပ္ပဟ်ဟာ ။

The Karen are the third largest ethnic group in Myanmar. It is also an ethnic group with many local languages. S'gaw Karen (sk) is one of the most widely spoken Karen languages. You can learn the basic details here [29]. Another local Karen proposed in this paper is Western Pwo Karen (pk). [30] Here is a basic guide to the language of Western Po Karen. The following example shows the sentence structure of S'gaw Karen (sk) and Western Po Karen (pk): English(eg): She will not ask me.

S'gaw Karen(sk): တမ့္ အဝဲ က သံကွ့ ယၤ ဘဉ်.

English(eg): Is it possible to try to forget someone you really love?

Western Pwo Karen(pk): ဆ၊ကျ့ၤစၢံၤ ဘနှ့ၣ်ဒုၤ ပှ လဂ၊ လ၊ ပ အဲ၊ဘ၊ တရျိရျိ နီ၊ မ က နှ့ဉ့၊ ဧ၊ ?

sk and pk languages script are derived from Burmese script. But word orders are not the same with Burmese. The word order of sk and pk languages are generally sub+verb+obj order and same with english language. [17], [31] explain the Pa'O language. It is also drived from the Buremese script. Arakanese (also known as Rakhine) is a language closely related to Burmese. It is used Burmese script. Word order is same with the Burmese. The basic natures of Rakhine language are explain in [32] and [33]. Rawang is spoken in Putao District, northern Kachin State. It is a Kachin ethnic group. This paper is also proposed the Rawang language. The basic element of the Rawang language can see at [34]. The sentence structure of language is as follow:

English(eg): This is Secretary Shwe.

Rawang(rw): YÀGØNØ BVNLĪÀNGKÀNG YWĒ ÍÈ.

Finally, the ethnic language we want to propose is Shan. The Shan are also Myanmar's second largest ethnic group. The basice concetp of Shan language are explain in [18] and [35].

IV. The Word-to-Word Dataset

We focus on the language pairs between Myanmar (Burmese) and Ethinc languages. Although these languages are spoken by Myanmar and other ethnic groups, data sets for that language pairs are scarce. So the propose system started introducing word-to-word translation

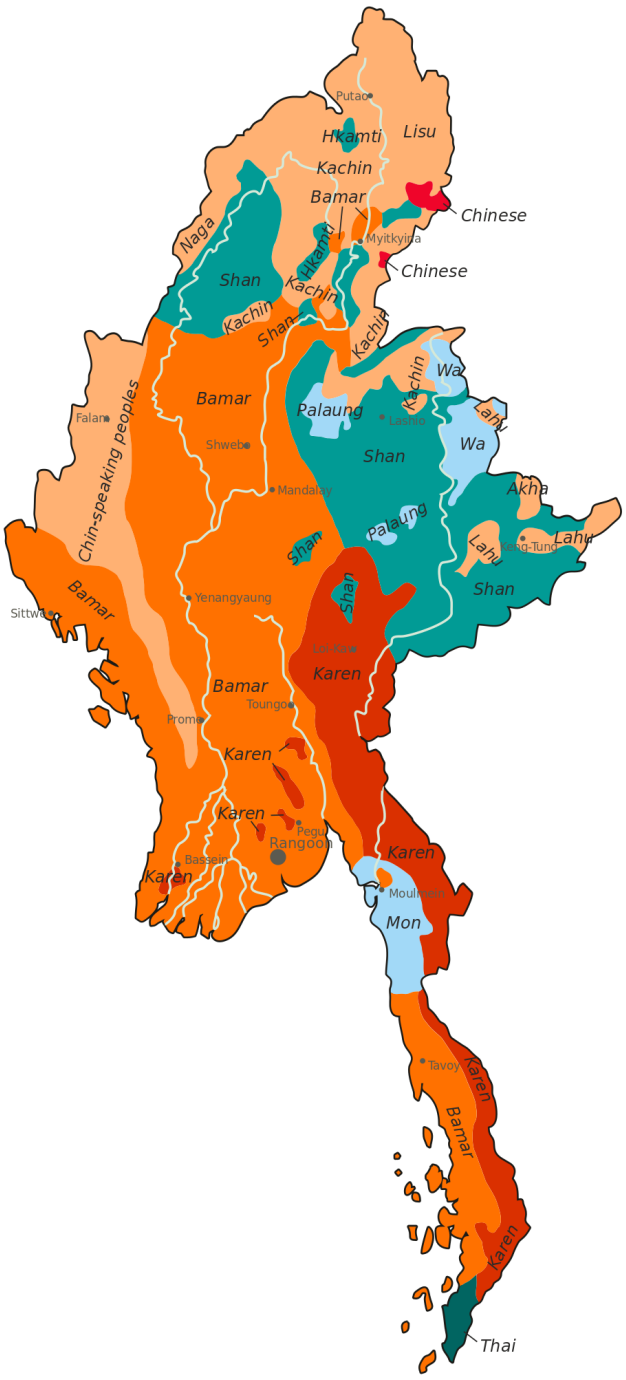


Fig. 1: Ethnolinguistic_map_of_Burma_1972_en

between Myanmar (Burmese) and some ethnic languages. We support 24 language pairs from 12 languages for word-to-word translation. The lexicon size of the entire dataset is 15743. Average translation accuracy is 91. But we need to compute for more accurate. An overview of the entire dataset is shown in Table I.

V. Methodology

Bilingual lexicon extraction (BLE) is a task in natural language processing. The goal of (BLE) is to search word-level correspondences from a (parallel) corpus. There are many difference approaches for (BLE).

[1]Among them, we focus on simple approaches that can work well with various sizes of parallel corpora. In particular, we avoid methods that require high-resource parallel corpora (e.g., neural machine translation) or external corpora (e.g., unsupervised or semi- supervised cross-lingual word embeddings). Also, since bilingual word-to-word mappings are hardly one-to-one [2]; [13]; [6], we consider methods that yield relevance scores between every source-target word pair, such that we can extract not just one but the top-5 correspondences. For these reasons, we consider approaches based on (monolingual and cross-lingual) co-occurrence counts: co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE).

1) Co-occurrences

The simplest baseline for our goal is to compute the co-occurrences between each source word x and target word y. For each source word x, we can score any target word y based on the conditional probability $p(y|x) \propto p(x, y)$:

$$p(y|x) = \frac{p(x,y)}{p(x)} \approx \frac{\#(x,y)}{\#(x)} \propto \#(x,y) \tag{1}$$

where $\#(\cdot)$ denotes the number of co-occurrence counts of the word or word pair across the parallel corpus. The top-k translations of source word x can be computed as the top-k target words with respect to their co-occurrence counts with x.

2) Pointwise Mutual Information

Another simple baseline is pointwise mutual information(PMI), which further accounts for the monolingual frequency of a candidate target word y:

$$\begin{aligned} p(x,y) &= log\frac{p(x,y)}{p(x)p(y)} \\ &\approx log\frac{\#(x,y)}{\#(x)\#(y)} \propto log\#(x,y) - log\#(y) \end{aligned} \tag{2}$$

Compared to the co-occurrence model in (1), PMI can help prevent stop words from obtaining high scores. The use of PMI has been connected to the skip-gram with negative sampling (SGNS) [14] model of word2vec [15]. PMI can also be interpreted as a conditional version of TF-IDF [2].

3) Controlled Predictive Effects

While conditional probability and PMI are proportional to cross-lingual co-occurrence counts, they can fail to distinguish exactly which source word in the sentence is the most predictive of the corresponding target word in the translated sentence. To deal with this issue, we add a correction term that averages the probability of seeing y given a confounder x' in the source language, i.e. $p(y|x')$. This probability is then weighted by the probability of actually seeing that confounder, i.e. $p(x'|x)$. We call the resulting approach as the method of controlled predictive

TABLE I: An overviews of word-to-word datasets

Languages	12
Language pairs	24
Avg. Lexicon Size	15743
Avg. Translations	91

effects (CPE).
Formally, we define the corrected CPE score as follows:

$$\begin{aligned} CPE(y|x) &= p(y|x) - \sum_{x' \in X} p(y|x')p(x'|x) \\ &= \sum_{x' \in X} CPE_{y|x}(x')p(x'|x) \end{aligned} \tag{3}$$

where X is the source vocabulary and

$$CPE_{y|x}(x')$$

denotes the CPE term of any other source word x' when predicting y from x. Formally, this term is defined as 4

$$CPE_{y|x} = p(y|x, x') - p(y|x') \tag{4}$$

All language pairs,lexicon sizes, and number of sentences are provide in Table II. The dictionary size of the three methods is the alike. The figure 2 shown the work flow diagram for word-to-word translation.

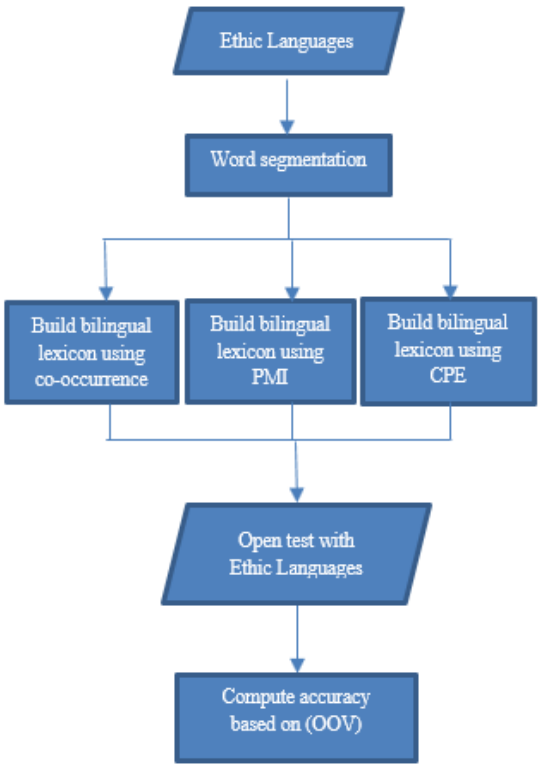


Fig. 2: work flow diagram

VI. Experiments Setup

1) Data Statistics

We have 240500 sentences translated from Myanmar (Burmese) to other ethnic languages. Most of these Burmese sentences are from the ASEAN-MT Parallel Corpus (without name entity tags) [16], 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), and special needs (emergency and health). Translation between Myanmar (Burmese) and other Ethnic languages are done by native speakers as manual translation and those sentences are checked by other native speakers. We used those parallel corpus for word-to-word translation between them. Our proposed system apply 100 sentences for every languages as open test sentences and the rest of sentences from parallel corpus used as training data.

2) Word Segmentation

The Writing style of the Burmese language and some ethnic languages is written without spaces between words. But spaces is used between phrases to make it easier to read. In short sentences, spaces are not usually used. There are no clear rules about where to apply space.
In this experiment, the sentences from the ASEAN-MT Parallel Corpus were word segmented. The rules of myanmar word segmentation and some ethnic word segmentation are not similar. [17], [18] Therefore, a few rules are used for word segmentation of soem ethnic languages. For example, some rules of Pa'O languages and Shan languages are as follow:
∗ space is put between the noun and the plural form of the noun for Pa'O
∗ nouns are written similarly to English plural words for Shan
∗ compound nouns are considered as one word for Pa'O and Shan
The detail rules of word segmentation of Some ethin languane are as followd:
In [22] explain the word segmentation rule for Myeik Lanugae and the rules of Rakhine language are shown at [33]. The segmentation rule of chin language is explain in [24]. Word segmentation is done manually for all ethnic languages .

3) The word-to-word Python Interface

[1]As part of releasing the dataset and making it easily accessible and reproducible, we also introduce the word-to-word Python package. The open-source package provides an easy-to-use interface for both downloading and accessing bilingual lexicons and building a custom bilingual lexicon on other language pairs for which there is a parallel corpus. Source code is available on PyPi as

TABLE II: summary statistics for bilingual lexicons

Sources Languages	**Target Languages**	**Lexicon Sizes**	**Number of Sentences**
my	bk	9.4k	10722
my	ch	14.4k	14983
my	kc	10.7k	38173
my	Ky	9.4k	10231
my	mo	9.8k	8973
my	pk	16.2k	19139
my	po	16.2k	18354
my	rk	16.5k	18373
my	rw	3.5k	5376
my	sh	15k	16533
my	sk	20.9k	68571
bk	my	9.6k	10722
ch	my	5.7k	14983
kc	my	4.6k	38173
ky	my	11.3k	10231
mo	my	9.6k	8973
pk	my	19.7k	19139
po	my	26.7k	18354
rk	my	17.9k	18373
sh	my	24k	16533
sk	my	81.6k	68571
bk	rk	9.6	10722
rk	bk	9.9k	10722
kc	rw	2.2k	10000

TABLE III: Human Evaluation between Myanmar(Burmese) and Some Ethnic languages

Sources Languages	**Target Languages**	**co-occurrences**	**PMI**	**CPE**
my	bk	2.5%	7.2%	23.3%
bk	my	2%	6%	18%
my	po	13%	4.8%	7.42%
po	my	14.9%	1.53%	13.5%
my	rk	32.4%	11.2%	42.1%
rk	my	11.6%	17%	39%
my	sk	27.9%	2.4%	23.4%
sk	my	21.2%	10%	22.5%
rk	bk	3%	8.2%	22.1%
bk	rk	8%	4%	19%
my	pk	3%	8.2%	22.1%
pk	my	8%	4%	19%

https://pypi.org/project/word2word/.
You can access all the experiments we have done as knowledge sharing at
https://github.com/ye-kyaw-thu/error-overflow/blob/master/word2word_translation-exp-log.md
https://github.com/ye-kyaw-thu/error-overflow/blob/master/word-to-word-translation-exp-for-ethnic-languages.md

VII. Evaluation

The proposed paper used three criteria for the evaluation of the translation output. They are the standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU), Out Of Vocabulary (OOV) and Human evaluation. The word-to-word translations level are determined based on the Out Of Vocabulary (OOV) rate. When the input includes words which were not present in a system's dictionary or database during its preparation [19], these words are call out of vocabulary(OOV). The word-to-word translations level are determined based on the OOV rate. BLUE score is the standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [20]. The BLEU score measures the precision of n-gram concerning to a reference translation with a penalty for short translations. Although our experiment is a word-to-word translation, we use the blue score to determine the evaluation of sentence level translation between Burmese and other Ethnic languages. For some ethinc language pairs, the human evaluation was checked to provide more accurate results.

VIII. Results and Discussion

We consider lexicons with three approaches: co-occurrences, PMI (Pointwise Mutual Information) and CPE (Controlled Predictived Effects), taking into account both monolingual and crosslingual or bilingual co-occurrences. Building word-to-word lexicon with low-resourced data helps to some extent for word-to-word level translation. The lexicon building speed is also fast. There were two types of evaluation: word-to-word translation, which counts how many OOVs there are (that is, counting how many OOVs can be translated) and the word-to-word translation for each word is constructed as a sentence (ie build as a sentence) and the BLEU score Calculating.
First, We compare the rate of OOV between Myanmar(Burmese) and other Ethnic languages based on three approches co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE). The detail of OOV rate between Myanmar(Burmese) and other Ethnic languages are display in Table IV. The OOV rates of all approaches are lowest in Myanmar and S'gaw Karen, Myanmar and Mon, Kachin and Myanmar, Myanmar and Kayah languages pairs. Especially in Myanmar and S'gaw Kayin, the OOV rate is 0. Po'O and Myanmar language pair have highest OOV rate and other languages pair have medium level OOV rate between lowest and higrest OOV rate. The OOV rates of all language pair are favorable in all approaches.
Then, we compaer Blue sorce between Myanmar(Burmese) and other Ethnic languages based on three approches co-occurrences, pointwise mutual information (PMI), and co-occurrences with controlled predictive effects (CPE). The BLUE scores of sentence level are not good level. In a language pair like Myanmar-Rakhine, the BLEU score is 14, 15, but this very similar dialect is very effective if you train SMTs at the same parallel corpus size. Some parallel-data is not available and can not be used, especially at the sentence level, depending on the corpus constructed. This is not yet possible, especially in sentence-level translations, depending on a lexicon built with little parallel-data.

1) Human Evalutaion

In this section, we provide the human evaluation results for some Ethnic languages. We investigate the human evalution for the output results from the word level translation. Table III are the details of the Ethinc language pairs and results of human evaluation.

IX. Conclusion

This paper contribute the very first word-to-word translation between Burmese and other Ethnic languages in both dierctions accorrding to the approches:co-occurrences, PMI (Pointwise Mutual Information) and CPE (Controlled Predictived Effects). The capabilities of word-to-word translation are at an acceptable level in all three approaches. The ability of sentences level translation are poor quality or a low standard.
The current word-to-word evaluation has some points to make. OOV% is calculated, but the actual translation is not properly analyzed. The problem is that we do not have a word-to-word dictionary ready to use. For human evaluation, research funds are needed to seek the help of native speakers (eg, those who understand Kachin, those who understand Shan, those who understand Khyan, etc.) for each language series. It also takes time. So, another extension of the experiment is the phrase-level alignment experiment, which translates between Burmese and Ethnic languages. If we could increase the size of the current parallel-corpus, Well done for drafting a bigger lexicon that is actually more useful. So,data extension is require.

Acknowledgment

This paper uses most of the Ethnic languages. So much help is needed from the relevant native speakers. Special thanks to Hnin Ye Aye for Chin (Mizo). Thanks to Htay Mon (Sakhangyi Middle School), Hsu Pan Oo (UCS Thaton), Pale Phyu (UCS Thaton), Thazin Oo (UCS Thaton), Thida Oo (UCS Thaton) speakers and other speakers for helping to build the Mon ethnic language corpus. Thank you for U Oo Hla Kyaw (Editor of the Rakhine Newspaper), Mg Than Htun Soe (CU Sittwe), Mg Htet Myart Kyaw (CU Sittwe) and Ma Oo Moe Wai (CU Sittwe) hard work on this Rakhine language corpus. Thanks again, we couldn't have pulled Kachin and Rawarn languages

TABLE IV: The OOV rate between Myanmar(Burmese) and other Ethnic languages words lexicons

Sources Languages	Target Languages	co-occurrences	PMI	CPE
my	bk	6.92%	6.92%	6.92%
bk	my	9.71%	9.71%	9.71%
my	ch	7.53%	7.53%	7.53%
ch	my	15.21%	15.21%	15.21%
my	kc	2.12%	2.12%	2.12%
kc	my	0.31%	0.31%	0.31%
my	ky	9.96%	9.96%	9.96%
ky	my	12.29%	12.29%	12.29%
my	mo	4.40%	4.40%	4.40%
mo	my	8.69%	8.69%	8.69%
my	pk	11.11%	11.11%	11.11%
pk	my	24.06%	24.06%	24.06%
my	po	12.65%	12.65%	12.65%
po	my	61.57%	61.57%	61.57%
my	rk	10.26%	10.26%	10.26%
rk	my	12.05%	12.05%	12.05%
my	sh	7.10%	7.10%	7.10%
sh	my	11.25%	11.25%	11.25%
my	sk	0.00%	0.00%	0.00%
sk	my	9.52%	9.52%	9.52%
rk	bk	8.73%	8.73%	8.73%
bk	rk	9.71%	9.71%	9.71%
my	rw	3.71%	3.71%	3.71%
kc	rw	15.90%	15.90%	15.90%

without Manar Hti Seng (UCS Banmaw), Dee Wom (UCS Banmaw), Hpau Myang Thint Nu (UCS Banmaw), Seng Mai (UCS Banmaw). Thank you, Zar Zar Linn (MIIT), Dr. Myar Doo Myar Reh, Mr. Khu Phya Reh (Tutor, History Department, Loikaw University), Mrs. Cember Paw (Kayah Freelance Translator), Mr.Khu Syah Reh (Kayah Bible Translation Team, Kayah National Literacy and Culture Committee) amazing! for Kayhan language. i wish to thank you to the Sai Saw Han (UCSY), Naw Si Ge Pho (Phyu Aung Dwe Bible School), Saw Kaung San (Dawei Inn Village, Bago Township), Saw Sai Mon (Tutor, English Center, Yangon), Saw Edison Thein (Yangon University of Foreign Studies, France) for all the support to build the S'gaw Kayin corupus. Special thanks to Mahn Samson or Mahn Doe (Myaungmya) for Pwo Kayin language corpus. I appreciated to U Sai Kham Sint (Shan freelance translator and editor), Sai Kwang Kham, Sai Nyan Lin, Nang Mwe Nom and Nang Vo Hein (Shan translation team, UTYCC) for Shan language. Thanks also to the native speakers for translating and verifying other ethnic languages from Myanmar to build coupus. Thanks also to National Electronics and Computer Technology Center (NECTEC), Thailand for sharing the ASEAN-MT corpus.

References

[1] Yo Joong Choe, Kyubyong Park, Dongwoo Kim, "word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs", 2019, Computation and Language (cs.CL).

[2] Fung, P. , "A statistical view on bilingual lexicon ex- traction: from parallel corpora to non-parallel corpora", 1998, Machine Translation and the Information Soup, pages 1– 17.

[3] Ramesh, S. H. and Sankaranarayanan, K. P. , " Neu- ral machine translation for low resource languages using bilingual lexicon induced from comparable corpora", 2018, In Proceedings of the 2018 Conference of the North Ameri- can Chapter of the Association for Computational Lin- guistics: Student Research Workshop, pages 112–119, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.

[4] Gū, J., Shavarani, H. S., and Sarkar, A. , "Pointer- based fusion of bilingual lexicons into neural machine translation", 2019 arXiv preprint arXiv:1909.07907.

[5] Ruder, S., Vulić, I., and Søgaard, A. , " A sur- vey of cross-lingual embedding models ", 2017, arXiv preprint arXiv:1706.04902.

[6] Levy, O., Søgaard, A., and Goldberg, Y. , " A strong baseline for learning cross-lingual word embed- dings from sentence alignments", 2017, In Proceedings of the 15th Conference of the European Chapter of the Asso- ciation for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 765–774.

[7] https://en.wikipedia.org/wiki/Languages_of_Myanmar

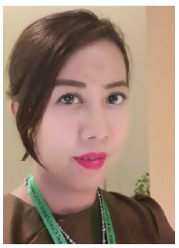
[8] Ximena Gutierrez-Vasques, Victor Mijangos., "Low-resource bilingual lexicon extraction using graph based word embeddings", 2017, Computation and Language (cs.CL)

[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou. (2008). " Word Translation Without Parallel Data". conference paper at ICLR 2018.

[10] JAKUBINA, Laurent and LANGLAIS, Philippe, " A Comparison of Methods for Identifying the Translation of Words in a Comparable Corpus: Recipes and Limits ", Comp. y Sist. [online]. 2016, vol.20, n.3, pp.449-458. ISSN 2007-9737.
[11] K. Nwet, K. Soe, N. Thein., "Developing Word-aligned Myanmar-English Parallel Corpus based on the IBM Models", (2013), Computer Science International Journal of Computer Applications.
[12] Lison, P., Tiedemann, J., and Kouylekov, M, "Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora", 2018, In Proceed- ings of the Eleventh International Conference on Lan- guage Resources and Evaluation (LREC 2018).
[13] Somers, H., "Bilingual parallel corpora and lan- guage engineering", 2001, In Proc. Anglo-Indian Workshop" Language Engineering for South-Asian languages.
[14] Levy, O. and Goldberg, Y., " Neural word embedding as implicit matrix factorization", 2014, In Advances in Neural Information Processing Systems, pages 2177–2185.
[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., " Distributed representations of words and phrases and their compositionality", 2013b, In Advances in neural information processing systems, pages 3111– 3119.
[16] B. Prachya and S. Thepchai, "Technical report for the network-based ASEAN language translation public service project," in Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013.
[17] Hay Man Htun, Ye Kyaw Thu, Hlaing Myat Nwe, May Thu Win, Naw Naw, "Statistical Machine Translation System Combinations on Phrase -based, Hierarchical Phrase-based and Operation Sequence Model for Burmese and Pa'O Language Pair", Journal of Intelligent Informatics and Smart Technology, October Issue, 2021, pp. 1-9.
[18] Nang Aeindray Kyaw, Ye Kyaw Thu, Hlaing Myat Nwe, Phyu Phyu Tar, Nandar Win Min, Thepchai Supnithi, "A Study of Three Statistical Machine Translation Methods for Myanmar (Burmese) and Shan (Tai Long) Language Pair", In Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020), Nov 18 to Nov 20, 2020, Bangkok, Thailand, pp. 218-223.
[19] https://en.wikipedia.org/wiki/N-gram#Examples
[20] 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 sium on Artificial Intelligence and Natural Language Processing Center, 2001
[21] https://en.wikipedia.org/wiki/Burmese_language
[22] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe and Thepchai Supnithi, "Statistical Machine Translation between Myanmar and Myeik", In Proceedings of the 12th International Conference on Future Computer and Communication (ICFCC 2020), Feb 26-28, 2020, Yangon, Myanmar, pp. 36-45
[23] https://en.wikipedia.org/wiki/Hakha_Chin_language
[24] Hnin Yi Aye, Yuzana Win, Ye Kyaw Thu, "Statistical Machine Translation between Myanmar (Burmese) and Chin (Mizo) Language", In Proceedings of The 23rd Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (Oriental COCOSDA 2020), Nov 5-7 2020, Yangon, Myanmar, pp. 211-216.
[25] https://en.wikipedia.org/wiki/Jingpho_language
[26] Zar Zar Linn, Ye Kyaw Thu, Pushpa B. Patil, "Statistical Machine Translation between Myanmar (Burmese) and Kayah Languages", Journal of Intelligent Informatics and Smart Technology, April 1st Issue, 2020, pp. 62-68. (submitted December 21, 2019; accepted March 6, 2020; revised April 23, 2020; published online April 30, 2020)
[27] https://en.wikipedia.org/wiki/Red_Karen_language
[28] https://en.wikipedia.org/wiki/Mon_language
[29] https://en.wikipedia.org/wiki/S'gaw_Karen_language
[30] https://en.wikipedia.org/wiki/Western_Pwo_language
[31] https://en.wikipedia.org/wiki/Pa'O_language
[32] https://en.wikipedia.org/wiki/Arakanese_language
[33] 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.
[34] https://en.wikipedia.org/wiki/Rawang_language
[35] https://en.wikipedia.org/wiki/Shan_language

May Myat Myat Khaing is a Lab member of Language Understanding Lab., Myanmar. She is interested in the research area of natural language processing (NLP) such as machine translation, speech processing and Data Science.

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 Affiliate Professor at Cambodia Academy of Digital Technology (CADT), Cambodia. He is also a founder of Language Understanding Lab., Myanmar. His research lie in the fields of artificial intelligence (AI), natural language processing (NLP) and human-computer interaction (HCI). He is actively co-supervising/supervising undergrad, masters' and doctoral students of several universities including Assumption University (AU), Kasetsart University (KU), King Mongkut's Institute of Technology Ladkrabang (KMITL) and Sirindhorn International Institute of Technology (SIIT).



Thazin Myint Oo received the B.C.Sc. (Hons) and M.C.Sc degrees from the University of Computer Studies Yangon, in 2005 and 2008, respectively. She is now a Ph.D. candidate of Assumption University, Thailand and also a Lab member of Natural Language Understanding, Myanmar. Her research interests include machine translation and natural language processing.

Thet Thet Zin received the M.C.Sc and Ph.D(IT) from University of Computer Studies, Yangon (UCSY) Myanmar in 2007 and 2012 respectively. Her current research interest is Natural Language Processing, Machine Learning and Computer Vision.

Nang Aeindray Kyaw is a master candidate at TAIST-Tokyo Tech, Artificial Intelligence and Internet of Things (AIoT), Thailand. She holds a degree of Bachelor's of Engineering (Information Science and Technology) from the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. She is strongly interested in Natural Language Processing (NLP).

Hay Man Htun is a candidate of the M.E in Artificial Intelligence and Internet of Things (AIoT) at Kasetsart University (KU), Bangkok, Thailand. She holds the degree of Bachelor of Engineering (Information Science & Technology) from the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. She is also a member of the NLP Research Lab.,UTYCC. Her current master thesis research focuses on Speech Translation in Natural Languages. She is strongly interested in the areas of Natural Language Processing (NLP) (NLP) such as machine translation, Speech Processing, Image Processing, Machine Learning, and Deep Learning.

Thida San is a Ph.D candidate at University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin and a faculty at Myanmar Institute of Information Technology (MIIT) Myanmar. Her current doctoral thesis research focuses on Text to Speech between Myanmar Braille and Myanmar written text. She is interested in the research area of natural language processing (NLP), speech processing, big data analysis, and deep learning.

Zun Hlaing Moe is a faculty at Myanmar Institute of Information Technology (MIIT) and also a Ph.D candidate at University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar. Her current doctoral thesis research focuses on machine translation between Myanmar Braille and Myanmar written text and vice versa. She is interested in the area of natural language processing (NLP) such as machine translation, big data analysis and deep learning.

Hnin Aye Thant is currently working as a Professor and Head of Department of Information Science at the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin Township, Mandalay Division, Myanmar. She got Ph.D (IT) Degree from University of Computer Studies, Yangon, Myanmar in 2005. The current responsibilities are managing professional teachers, doing instructional designer of e-learning content development and teaching. She has 14 years teaching experiences in Information Technology specialized in Programming Languages (C,C++, Java and Assembly), Data Structure, Design and Analysis of Algorithms/Parallel Algorithms, Database Management System, Web Application Development, Operating System, Data Mining and Natural Language Processing. She is a member of research group in "Neural Network Machine Translation between Myanmar Sign Language to Myanmar Written Text" and Myanmar NLP Lab in UTYCC. She is also a Master Instructor and Coaching Expert of USAID COMET Mekong Learning Center. So, she has trained 190 Instructors from ten Technological Universities, twelve Computer Universities and UTYCC for Professional Development course to transform teacher-centered approach to learner-centered approach. This model is to reduce the skills gap between Universities and Industries and to fulfill the students' work-readiness skills.