Syllable-based Neural Machine Translation System for Myanmar-English Language Pair

Yi Mon Shwe Sin

University of Computer Studies, Yangon

October, 2019

Syllable-based Neural Machine Translation System for Myanmar-English Language Pair

Yi Mon Shwe Sin

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

October, 2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…..……………………………   ………….......…………………………

Date   Yi Mon Shwe Sin

ACKNOWLEDGEMENTS

First of all, I would like to thank the Union Minister, the Ministry of Education, for allowing me the advanced study and providing full facilities during the Ph. D Course at the University of Computer Studies, Yangon.

Secondly, I would like to express very special thanks to Dr. Mie Mie Thet Thwin, Rector of the University of Computer Studies, Yangon, for allowing me to develop this thesis and giving me general guidance during the period of my study.

I deeply and specially thank the external examiner, Dr. Tin Htar Nwe, Professor, University of Computer Studies, Magway, for her patience in critical reading, valuable suggestions and comments in the preparation of thesis.

I would also like to express my special appreciation and thanks to Dr. Khine Moe Nwe, Professor and Course-coordinator of the Ph.D 9th Batch, University of Computer Studies, Yangon, for the kindness, useful comments, advices and insight which are invaluable to me.

I would like to express my deepest gratitude to my supervisor, Dr. Khin Mar Soe, Professor, Natural Language Processing Lab., University of Computer Studies, Yangon, for her excellent guidance, caring, patience, and providing me with excellent ideas for doing research.

I would also like to extend my special appreciation and thanks to Dr. Win Pa Pa, Professor, Natural Language Processing Lab, the University of Computer Studies, Yangon, for giving enough consideration to my ideas and views.

I would like to express my respectful gratitude to Daw Aye Aye Khine, Associate Professor, Head of English Department for her valuable supports from the language point of view and pointed out the correct usage not only through the Ph.D. course work but also in my dissertation.

My sincere thanks also goes to all my respectful Professors for giving us valuable lectures and knowledge during the Ph.D. course work.

I also thank my friends from Ph.D. 9th Batch for providing support and friendship that I needed.

I am very much indebted to my family for always believing in me, for their endless love and support**.** They are always supporting and encouraging me during the years of my Ph.D. study. This accomplishment would not have been possible without them.

ABSTRACT

Without doubt, Internet has made for people to communicate easily with others because it is cheap and convenient. Besides, today's world has a growing demand for immediate and accurate information. However, language remains an important barrier that prevents all information from being spread across different cultures because of the high cost in terms of money and time that human translation implies. Therefore, the demands and challenges for translation from one language to another language is steadily growing. To overcome this demands and challenges, the researchers do research the machine translation to easily translate one language to several languages for the user. Therefore, machine translation is becoming a popular field in research trends and it make smaller the language barrier.

Machine Translation (MT), which is also known as Computer Aided Translation, is the task of specifically designing to translate both verbal and written texts between natural languages by a computer system. MT uses a machine translation engine to perform substitution of words or phrases or any other in one language for words or phrases or any other in another language. MT was one of the first conceive computer applications in the 1950's and there is still many challenges. MT is widely used in Natural Language Processing tasks such as online translation services applications in information extraction, document retrieval, intelligence analysis, electronic mail, and much more. Today, some machine translation applications are available in the market. The most widely used being applications are Statistical Machine Translation (SMT), Rule-Based Machine Translation (RBMT), Hybrid Systems, which combine RBMT and SMT and Neural Machine Translation (NMT).

Today, there have been very few studies on the machine translation from Myanmar language to another language. And Myanmar machine translation is still in its early stages and researchers are faced with many difficulties such as the lack of resources and there is only less amount of data corpus. Furthermore, the techniques for performing as pre-processing step for Myanmar language, such as segmentation are also currently in the process of being developed. Existing research on Myanmar translation has been either rule-based or more recently phrase-based techniques have been tried.

The research aims to develop Neural Machine Translation system for Myanmar to English language pair. In this work, there are two main parts. The first one is the building the Myanmar-English parallel corpus and Myanmar monolingual corpus. Although Myanmar language is one of the low resource languages, bilingual sentences are collected from the website and eBooks by crawling or copying. The second part introduces the attention-based neural machine translation system for Myanmar-English language pair. The experiments of the proposed model are done based on word-based neural machine translation model, Myanmar character-based neural machine translation model and Myanmar syllable-based neural machine translation model. Moreover, Myanmar monolingual corpus is also used to improve Myanmar syllable-based neural machine translation model. The experimental results show that Myanmar syllable-based neural machine translation model outweighs over other models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

Machine Translation (MT) is the automatic translation mechanism from one natural language into another language by means of a computerized system. In the late twentieth, the translation of natural languages by machine has been popular in the real world.

Generally, machine translation starts in the 1950s, although the researcher found the machine translation trend from earlier periods. During a long time from 1950 to 1980, machine translation was done using the linguistic information. Then, it generates the translations based on the dictionaries and grammars rules. This method is called rule-based machine translation (RBMT). Since initially documented, machine translation has revealed to be one of the most complex tasks to carry out in the field of natural language processing (NLP), considered one of the AI-hard problems. Therefore, machine translation is increasingly popular since then.

With the development of Statistics, Statistical Machine Translation (SMT) is becoming a popular research area in the late 1980s. SMT also translates based on the parallel corpora. This method achieved the better performance than other method such as RBMT. From the 1980s to until now, SMT controlled the machine translation field. Word-based, phrase-based, syntax-based and hierarchical phrase-based are the approaches based on SMT. Among them, phrase based statistical machine translation(PBSMT) also became a popular trend over the last few years[37].

Later in 2003, the researchers improved the language model by applying on neural networks models. These models addressed the data sparsity problem of previous SMT models. They decided to apply a foundation of neural networks for machine translation in future. Therefore, Neural Machine Translation(NMT) became a very popular research title now[1].

For Myanmar-English machine translation, the previous research mostly learnt by using rule-based as well as statistically based approach. However, the translation performance is not sufficient to be good accuracy due to the lack of a huge amount of corpus. Myanmar language is an under-resourced language (known as low-resource language) and there were not many parallel corpus or monolingual corpus. There are existing parallel corpora for Myanmar-English languages pair. They are Basic Travel

Expressions Corpus(BTEC) and Asian Language Treebank(ALT) Corpus. Both corpus are not more than 20,000 parallel sentences. Actually, the performance of the NLP tasks is depending on the size of the data.

In this research, firstly, large-scale Myanmar-English parallel corpus and Myanmar monolingual corpus are built. Myanmar-English parallel corpus, called UCSY-corpus, is a general domain which includes the parallel sentences of local news, travel domain, school text book and spoken text. There are over 200K parallel sentences in total. And Myanmar monolingual corpus is local news data and consists of 170K Myanmar sentences. These sentences are collected by crawling from the bilingual websites, by downloading from the websites and by copying from the eBooks.

Then, neural machine translation with attention models are built for Myanmar-English language pair in both directions. As a baseline systems, word to word NMT model are trained applying OpenNMT toolkits for Myanmar to English and English to Myanmar neural machine translation systems. In addition, Myanmar character-based NMT model and Myanmar syllable-based NMT model are also trained. For word to word neural machine translation model, UCSY_NLP lab segmenter is used to segment Myanmar sentences into word level. For Myanmar character-based neural machine translation model, python programming code is applied to segment Myanmar sentences into character level. For Myanmar syllable-based neural machine translation model, syllable segmenter is used to segment Myanmar sentences into syllable level. All models are trained on default setting of OpenNMT [80].

Moreover, Myanmar Monolingual Data are also used to improve the Myanmar syllable-based NMT model. When monolingual data is used, it is tested with two ways. They are copy monolingual data usage and other monolingual data usage for Myanmar-English neural machine translation models in both directions.

1.1 Problem Statement

Neural machine translation is an end-to-end architecture to translate by substitution of words or characters or any other in one language into another languages. Any machine translation systems, especially statistical machine translation system and neural machine translation system, requires a large amount of parallel corpus. Lack of large amount of parallel corpus for system development is the main issue in developing machine translation.

All other applications for Myanmar-English language pair are still very much at an early stage, and fall far short of yielding consistently accurate translation results. Although Myanmar NLP is struggling to be developed at the present time, large amount of resources available are very insufficient. Since the parallel corpus is the most important component of the system, obstacles were present right from the outset. This is due to the fact that there are very few available resources necessary for the construction of an effective system.

Myanmar language and English language are vastly different languages not only in terms of basic sentence structure but also in grammar, syntax, and morphology. This can cause great complications in any natural language processing task. Previous Myanmar-English machine translation system translate only the short sentences well because the existing corpus collected mostly short sentences.

NMT has seen very limited use with the Myanmar-English language pair, and there was no documented research work. Word level are considered as basic training unit in nearly all previous work in machine translation. To deal with unknown word problem, character level is used as units input and output. In Myanmar language, word level also used as basic training unit in a baseline system. However, Myanmar language is a morphologically rich language and it is necessary to deal with unknown words problems (UNK). Therefore, not only character level but also syllable level are used to address unknown words problems (UNK).

According to the mentioned problems, neural machine translation system is still necessary to develop in the area of Myanmar NLP research.

1.2 Motivation of the Research

One of the reasons why it is tried to research the machine translation is to generate the accurate translation results for Myanmar-English neural machine translation system in both direction. Furthermore, the lack of large scale parallel corpus is to be motivate to do the research. Although there are many data resources for other languages, there is no available parallel corpus for low resource languages, especially Myanmar language. For this reason, building a large scale Myanmar-English parallel corpus is prepared to construct.

Another point is that the previous studies on Myanmar-English machine translation system had mainly focused on rule-based and statistical approaches. In the current, there is no Myanmar-English machine translation system based on neural method. Besides, neural machine translation system has not been published for Myanmar-English language pair.

Therefore, this research is to build Myanmar-English parallel corpus, Myanmar monolingual corpus and to develop the Myanmar-English Machine Translation system.

1.3 The Objectives of the Research

The main objective of this research is to develop Myanmar-English neural machine translation system through the various experiments. Today, statistical approach or neural approach are usually used in machine translation system. However, these approaches need to train the many bilingual sentences. Myanmar-English machine translation is still at an early stage and requires to produce the improving translation results. This is due to the fact that does not have a few available parallel sentences for the Myanmar machine translation model. Therefore, a Myanmar-English parallel corpus have been concentrated to prepare in this system. By using these large-scale corpus, various experiments have been tested based on neural machine translation system for Myanmar-English language pair to get high result. The objectives of this research area are as follows:

(i) To build a Myanmar-English parallel corpus

(ii) To build a Myanmar monolingual corpus

(iii) To support the corpus into the other natural language processing(NLP) tasks

(iv) To train the corpus for Myanmar to English machine translation system

(v) To develop Myanmar-English neural machine translation model in both directions

(vi) To discover the best performance of neural machine translation model for Myanmar-English language pair

1.4 Contributions of the Research

Until recently, there were not many available Myanmar-English parallel corpora. There was an only several thousand sentences and this is too small to use in NLP tasks. Especially, Machine translation between Myanmar and English languages has so far been limited by a lack of parallel corpora. Actually, the parallel corpus is the most important component of the NLP tasks. Therefore, a number of large-scale parallel Myanmar-English sentences have been prepared. These data is from the local news, international news and Myanmar to English speaking sentences available from online. So, the first contribution is to build a large scale Myanmar to English parallel corpus.

Not only there is no sufficient resources but also no current machine translation system based on neural method for Myanmar-English language pair. These points are contributions of this systems. Therefore, the second contribution is to build the translation model based on NMT with attention for Myanmar-English language pair.

Neural machine translation reaches the great success of performance in some languages to translate by substitution of word or character or any other level. For Myanmar-English language pair, not only word level and Myanmar character level but also Myanmar syllable level has been investigated. Therefore, the third contribution is to introduce the Myanmar syllable-based neural machine translation model for both language pair.

1.5 Organization of the Research

This dissertation is organized with seven chapters, including introduction of machine translation, the existing Myanmar-English machine translation systems, about this research, objectives and contributions of the research work. The literature reviews on neural machine translation(NMT), and related words concerning existing Myanmar to English machine translation systems are described in Chapter 2. In Chapter 3, the nature of Myanmar language and the structure of Myanmar language is presented. Chapter 4 describes the building a parallel corpus called UCSY-corpus and Myanmar monolingual corpus for Myanmar-English language pair. In Chapter 5, encoder-decoder models, attention-based neural machine translation models are explained. The evaluation results are shown according to the experimental results in Chapter 6. Chapter 7 concludes with the limitations of the work and the future work.

CHAPTER 2
LITERATURE REVIEW

This chapter describes an introduction to natural language processing, machine translation systems, the literature review of existing Myanmar-English machine translation systems. In addition, the current different approaches on neural machine translation system are reviewed.

2.1 Natural Language Processing

Natural language processing (NLP) forms the backbone of every human language technology application. Natural language processing is concerned with natural or human languages which human beings use for day-to-day communications. Natural language processing is a subdivision of an artificial intelligence and computational linguistics. It learns the problem of natural human languages that is the automated machine generation and machine understanding. Natural language processing is concerned with the design and implementation of effective natural input and output components for computational systems. Therefore, the common critical problems is to work with the natural input and output.

Language is a system for communication. It embodies both verbal and written expression to help us communicate our thoughts and feelings. Language uses a wide range of sounds, sings, and symbols to create words, sentences, paragraphs and other media. Whether written or spoken, language is the medium we use for expression and organization what we know, think and feel.

In fact, most languages continue to evolve as society changes, as new technologies developments occur, and as changes in customs dictate. Formal languages are artificial or contrived languages deliberately developed for a special purpose. A good example is a computer language. If used properly, the language will permit a limited form of communication between people and computers.

The system of rules for putting words together to form complete sentence and thoughts is called grammar. Grammar is composed of two basic parts: syntax and semantics. Syntax arranges the words and phrases to form the sentences. It is a subfield of grammar that deals with the adding and sequencing of various words in the language such as nouns, verbs, adjectives and so on. Syntax is a method of putting words in a specific order or they will have the correct form according to the language.

Semantics refers to meaning in language. It is the study of relationships between words and the way they are assembled to present a particular through. Semantics provides us with ways of analyzing and interpreting what is being said. To keep in mind that meaning is also a function of syntax. The way words are used and ordered very much determined the meaning of the combination.

What we form sentences in a certain way and use specific words, we are referring to a certain model of the world that we have in mind. The mapping between a sentence and the conceptual mental model is the role of semantics. We usually speak or write in complete sentences. Each sentence expresses one complete thought. We then string the sentences together forming a paragraph to convey a particular idea. As a whole, the paragraph makes sense. But if we pull one sentence out of the paragraph and look at it in isolation, its meaning may not be fully understood. We say that the sentence is out of context.

Context refers to the complete idea or though surrounding any sentence in a paragraph. Together, all the sentences add up and make sense. Alone, each sentence contains only a piece of the whole and is often subject to interpretation is needed when the sentences are looked at together. Context clarifies meaning by explaining circumstances and relationship. Therefore, it is an important part of language and understanding.

Pragmatics refers to what people really mean by what say or write. Often, we say or write one thing but mean another. You may ask "what time it is", but you really mean "am I late for the meeting". Pragmatics tries to get at the true meaning or feeling. Both context and pragmatics play a major role meaning in understanding. It is one thing to communicate, but another to know the real meaning of the message. Context and pragmatics fill in the gaps often left by syntax and semantics.

The aim of natural language processing research is to create the computational models. The final goal is to be able to specify models that learn the human activities in the linguistics tasks of reading, listening and speaking.

In theory, natural language processing is a substitute method between human and computer. natural language processing do machines to understand human activities with intelligence. In order to do so, "Understanding" of the natural language sentences plays a vital role in the development of natural language processing program. Therefore, the primary goal of natural language processing was to understand how exactly the human beings understand, generate and learn languages [3].

2.1.1 Statistical Natural Language Processing

The statistical approaches to natural language processing have been remarkably successful over the past two decades. The availability of corpus has played a vital role in their success. And statistical methods rely on the amount of data. Statistical approaches work various mathematical techniques. It is often use corpora to generalize the models approximately. Natural language processing tasks requires to decide the annotations of the language normally. A statistical-based natural language processing approaches is good in predictions of the language annotations such as word sense, syntactic structure and etc. And, the statistical approach is able to learn the lexical and structural preferences from corpora [17]. Statistical models are robust, generalize well and behave gracefully in the presence of errors and new data [21]. Therefore, statistical models on natural language processing have led to provide the successful disambiguation in large scale systems with naturally occurring text. In addition, the parameters of statistical natural language processing models can often be estimated from text corpora. This reduces the human effort in producing natural language processing systems. And it raises interesting scientific issue regarding human language acquisition.

2.2 Machine Translation Systems

Machine translation is a computerized automated translation between the languages. Machine translation has a long history and researches sine 1950s various ups and downs. It has been steadily evolving with many new approaches and techniques until now. Over the years, researchers have applied different techniques to the face the challenge of machine translation. The most significant of these are outlined in the following subsections.

2.2.1 Rule-based Machine Translation System

In the field of machine translation, the rule-based approach is the first method that was developed. A rule-based machine translation system consists of collection of rules. These rules are called grammar rules and also called a bilingual or multilingual lexicon rules. Another called software programs to process the rules. Nevertheless, building rule-based machine translation system requires a human assistant to generate all of the linguistics resources such as part-of-speech taggers, syntactic parsers, bilingual dictionaries and source to target transliteration. Therefore, rule-based machine translation system always is extensible and maintainable. Building rules play a vital role in syntactic processing, semantic interpretation, contextual processing and etc., which are various stages of translation. Generally, ruled are generated with the linguistic information. Therefore, this approach is based on dictionary entries, word-by-word translation. The meanings of these words are not always interchangeable. Rule-based machine translation system is categorized into three different approaches. They are transfer-based machine translation, dictionary-based machine translation and interlingua machine translation.

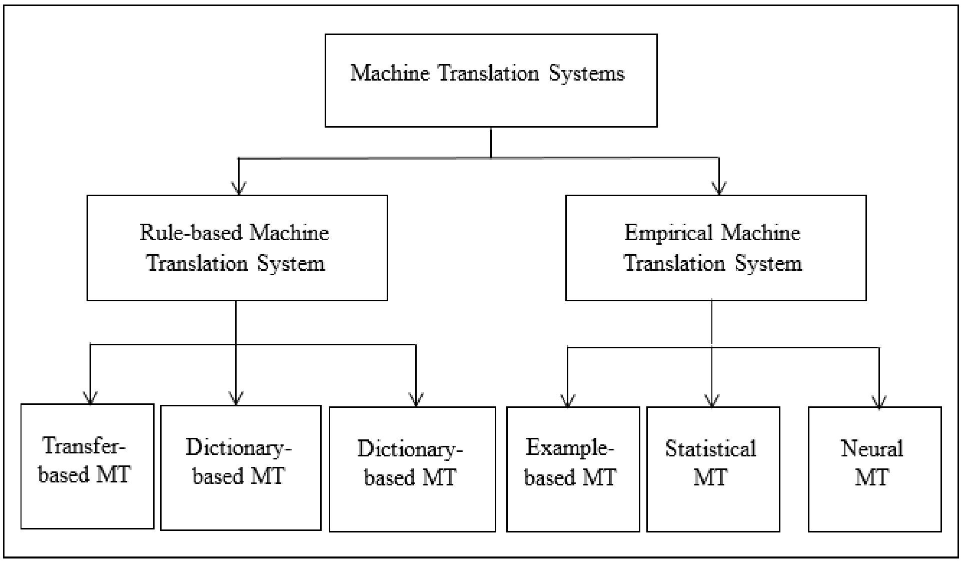

Machine Translation Systems
Rule-based Machine Translation System
Empirical Machine Translation System
Transfer-based MT
Dictionary-based MT
Dictionary-based MT
Example-based MT
Statistical MT
Neural MT

Figure 2.1 Types of Machine Translation

2.2.2 Transfer-based Machine Translation System

Transfer-based machine translation based on the idea of interlingua and an intermediate representation. It collects the meaning of source sentence and generates the correct translation. Transfer-based approaches locate between interlingua and direct machine translation. Syntactic transfer-based approach is closer to the direct approaches and analysis on the source sentences to generate the syntactic representations and target sentences output. Semantic transfer-based approach is closer to the interlingua approaches and copes the ambiguities after syntactic semantic analysis.

2.2.3 Direct Machine Translation System

Another type of machine translation is direction machine translation. Direct machine translation approaches directly translate from source language into target language by word-to-word. The advantage of this approach is that they do not need sophisticated syntactic and semantic analysis generally. However, it often ignores the meaningful linguistic information. This approach requires a large amount of bilingual sentences. To train translation models, these bilingual sentences are used. Example-based machine translation and statistical machine translation are typical approaches in this category.

2.2.4 Interlingua Machine Translation System

Interlingua-based approaches consists three steps. The first step is to analyze the source sentences. In the next step. it generated Interlingua which means a language independent semantic representation. The last step is to produce the target language translation which is based on the semantic representation. This approach is appropriate for multilingual translation language pairs where Interlingua analysis and language generation are developed for each language only once. However, the disadvantages of interlingua machine translation system requires human effort when the domain is getting larger and broader. Therefore, interlingua machine translation system is only suitable for specific domains. Typical Interlingua-based systems include [23, 66, 97, 100].

2.2.5 Example-based Machine Translation System

In 1994, Nagao Makoto firstly proposed the example-based machine translation system. Exampled-based approach is often use the bilingual corpus at run time as its main knowledge base. It is essentially a translation by analogy. And this approach can be viewed as an implementation of the case-based reasoning approach. Firstly, people translation decomposes a sentence into certain phrases. And then it translates these phrases and finally composes the properly translated phrases into one long sentence.

2.2.6 Statistical Machine Translation System

The statistical approach derives from the empirical machine translation (EMT) systems. These systems rely in the large amount of parallel aligned corpora. Statistical machine translation system is a framework for translating text from one language to another. These are based on the knowledge and statistical models extracted from parallel corpora. In statistical machine translation system. bilingual or multilingual sentences of the source and target language or languages are entailed. A statistical machine learning algorithm is used to build the statistical tables. This process is called the training and the statistical tables consists of statistical information. In the decoding step, these statistical information is used to find the best result. There are three different statistical approaches in machine translation. They are word-based translation, phrase-based translation, and hierarchical phrase based model.

2.3 Existing Myanmar-English Machine Translation Systems

The research of automatic translation on Myanmar-English language pairs begins 2010. However, the study of automatic translation of Myanmar to English is quite few. Most of all previous research is applied to the rule-based and the statistical based approaches.

Thet Thet Zin et al. [120,121] presented improving Myanmar to English translation model. Baseline system used N-gram that are based target and source language model. To reformulate the translation probability of phrases, Bayes rule is also used. This system use the small amount of training data, so morphological analysis is also applied to solve the Myanmar morphological problems. Their results presents that half of the tested dataset are the Out-of-Vocabulary (OOV) words in 2000 training sentences. Therefore, translation model used syntactic structure to reduce the number of OOV words in translation. By adding these features. it had shown that the system can achieve a better result. The system showed that the proposed method improved over the baseline system.

Soe and Thida [89] presented the Myanmar verb phrase identification and Myanmar to English translation. For Myanmar verb phrase identification, the system uses rule-based maximum matching approach. For Myanmar to English verb translation, Markov model is used to reformulate the translation probability. And the system is based on syntactic structure and morphology of Myanmar language by using Myanmar-English bilingual lexicon. The results showed that the proposed system improves the translation quality by applying morphological analysis on Myanmar Language.

Ye Kyaw Thu et al. [93] used the statistical machine translation approach to contribute the first large scale evaluation of translation between Myanmar and twenty other languages, in both directions. The twenty languages are Arabic, Chinese, English, German, Hindi, Indonesian, Italian, Japanese, Korean, Malaysian, Mongolian, Nepali, Portuguese, Russian, Sinhala, Spanish, Tagalog, Thai, Turkish and Vietnamese. The system experimented three different statistical machine translation approaches. They are phrase-based, hierarchical phrase-based, and the operation sequence model (OSM) between Myanmar and twenty languages from the multilingual Basic Travel Expressions Corpus (BTEC). For Myanmar sentence segmentation, three different segmentation schemes are used. They are syllable segmentation, maximum matching word segmentation with dictionary, and CRF word segmentation. The hierarchical phrase-based statistical machine translation system (HPBSMT) showed the highest translation quality in almost all language pairs.

Win Pa Pa et al. [74] studied the five-state-of-the-art statistical machine translation methods for under resourced languages. The five-state-of-the-art methods are phrase-based, hierarchical phrase-based, the operational sequence model, string-to-tree, tree-to-string statistical machine translation methods. And the under resourced languages are Lao, Myanmar and Thai. The system translated between English and these under resourced languages in both directions. The system trained these statistical machine translation systems using the ASEAN-MT parallel corpus for each language pair. The main motivation is to investigate machine translation performance with the dominant statistical machine translation(SMT) approaches on under resourced languages. In their experiments, the phrase-based statistical machine translation (PBSMT) method generally gave the highest BLEU scores.

Hammam Riza et al. [81] introduced the ALT project, Asian Language Treebank. ALT project includes six institutes such as, Indonesian, Japanese, Khmer, Malay, Myanmar and Vietnamese. The project was to make a parallel treebank for these languages and English. The texts were about 20,000 sentences sampled from the

English Wikinews. These sentences were translated into the six languages. ALT includes word segmentation, POS tagging, syntax analysis annotations, together with word alignment links among these languages. A survey of Asian natural language processing resources was also presented to highlight the contributions of ALT to the community.

Thandar Nwet and Khin Mar Soe [71] performed the Myanmar-English translation model. And the system performed the comparative study using Asian Language Tree-bank (ALT) corpus. The main motivation is to investigate the statistical based- Myanmar-English machine translation system. In Myanmar language, words are not delimited by spaces. Therefore, in the preprocessing step, the system used Myanmar Language Segmenter which was implemented by UCSY NLP Lab [73].

2.4 Creation of Corpus

The quality of neural machine translation systems is strongly depended on the size of parallel corpus. Although there is a lot of parallel sentences in rich resource languages, poor resource languages have little or no readily available parallel sentences. As a result, most machine translation researchers focus on the collecting of large amounts of parallel sentences.

Liang Tian et al. [96] constructed a large scale and high quality parallel corpus, UM-Corpus, for English and Chinese language pairs. The constructed corpus is designed to eight different kinds of domains and contain about 15 million parallel sentences. Eight different kinds of domains are News, Spoken, Laws, Thesis, Educational Materials, Science, Speech/Subtitles, and Microblog. Therefore, UM-Corpus is a multi-domain and balanced parallel corpus. The system aimed to create a large-multi-domain for statistical machine translation models. And this corpus is to serve as an important resource of the study of statistical machine translation domain adaptation. To construct UM-Corpus, the first step is to identify the appropriate bilingual websites and crawl. The second step is to extract the content and the next step is to aligned the document and the sentences. Finally, the collected data is cleaned by the removing of noisy text and duplication sentences with human experts. In total, there are 15,792,666 sentences.

Matt Post et al. [79] constructed a collection of parallel corpora between English and six low-resource and under-studied languages. These languages are Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. For each parallel corpus, the system first manually assigned each of the Wikipedia documents for each language with nine categories. They are events, language and culture, people, places, religion, sex, technology, things or misc. The system collected the parallel corpus with three steps. The first step was to build a bilingual dictionary. And these dictionaries were used to bootstrap the experimental controls in the collection of four translations of each source sentence. Finally, as a measure of data quality, the system independently collect votes on the which of the four redundant translations is the best.

Huidan LIU et al. [54] built a large scale text corpus for Tibetan language. The system found the Tibetan web pages with a crawler. And the system select the three biggest sites and topic pages with a rule based method by checking the url. The three biggest sites were China Tibet New, China Tibet Online and Tibetan's web of China. And the different domains are art, culture, economy, education, finance & economy, history & geometry, music, news, photo, picture, policy, politics & law, rural life, social life, special issues, technology & education, Tibetan Buddhism, Tibetan food, Tibetan medicine, tour, video and other. And then the selected pages are analyzed and extracted related topic information. Implementing this method, the system collected the parallel sentences starting on 12nd January in 2011. Consequently, the system collected more than 65 thousands documents. Nearly 1.59 million sentences and 35 million syllables existed in this corpus.

Aditya Kumar Pathak et al. [77] created automatic parallel corpus for Hindi-English language pairs. The parallel corpus was created from a comparable corpus by crawling from the web news translation tasks with the fuzzy string matching algorithm. Firstly, the system extracted Hindi news according to the range of month and year. It took as an input and these data are cleaned. And then the crawled news were translated English language using Google translate. After that, this translated English news was used as a baseline system. In the next step, English news contents were extracted. Ant then, the respective news documents are aligned manually on Hindi news documents, translated English news and English news documents. The token length, abbreviations and numbers synthesize using a rule based system by the system. Parallel corpora with respect to different threshold values based on the fuzzy string matching algorithm.

Eray Yıldız et al. [110] studied the effect of English-to-Turkish parallel corpus quality and size in the statistical machine translation model. English-Turkish parallel corpus consists of one million sentences. These sentences were from the various sources such as news text, literature text, subtitles text and web crawled text by varying quality levels. The system applied a machine learning based classifier to a parallel corpus to classify as high or poor quality.

2.5 Neural Machine Translation Systems

Machine translation is regarded as one of the most challenged and difficult tasks in natural language processing literature. In statistical machine translation settings, the problem is treated as calculating the probabilities of the target language model and the translation model of the language pair separately. Furthermore, the translation model has to deal with alignment problems and so on. However, in neural machine translation (NMT) , it is made possible to use a general, end-to-end model to solve this task. Now, it will review on some research neural machine translation methods; especially for attention-based neural machine translation.

Dzmitry Bahdanau et al. [4] first introduced the Attention mechanism. This approach learns a alignment model between the source and target characters or words. Neural machine translation has been developed in recent years and shows sufficiently results in full resources languages. Many neural machine translation models based on the encoder and decoder framework. This framework consists of one encoder and one decode recurrent neural networks (RNNs). The attention-based neural machine translation models proposed recently by a family of encoder and decoders framework. The encoder do to encode a source sentence into a fixed-length vector and the decoder generates a translation from it. Attention mechanism aims at building a single neural network. This can be jointly tuned to maximize the translation performance. The proposed model searches the most appropriate information in a source sentence for each time and predicts a target word according to these source information and all the previous generated target words. The system evaluated on the tasks of English-to-French translation with the parallel corpora, contained 850M words in total, provided by ACL WMT'14. The system train two types of models: RNN Encoder-Decoder(RNNencdec) and the proposed model(RNNsearch) with the same settings. Each model twice trained : first with the sentences of length up to 30 words and 50 words. The encoder and decoder of both model have 1000 hidden units each, and minibatch of 80 sentences. It takes approximately for 5 days for each model. The proposed models yield the good result on longer sentences. Unlike with the traditional machine translation systems, all of the pieces of the translation system, including the alignment mechanism, are jointly trained towards a better log-probability of producing correct translations. And The proposed RNNsearch outperforms the conventional encoder–decoder model (RNNencdec) significantly.

Jing Wu et al. [105] applied three methods of subword training, monolingual data and a neural machine translation correction model. The system aims to boost the low-resource Mongolian-Chinese language pairs bases on the attention-based neural machine translation models. A public Mongolian-Chinese parallel corpora CWMT'2009 is used which is the only public Mongolian-Chinese. It consists of 65K parallel sentences and the length is 50. The system also built phrase-based statistical machine translation model using Moses toolkits. The alignment is performed by GIZA++ toolkit and the decoder is performed by MERT. For attention-based neural machine translation, open-source GroundHog is used as Baseline 2. The system trained the attention-based neural machine translation with 1024 dimensional for word embedding, 1024 hidden units per layer for both encoder and decoder, the source vocabulary to 50K, the target vocabulary to 10K and beam search is 10 on GPU of NVIDIA Tesla K80. The proposed methods were effective to improve the Mongolian-Chinese neural machine translation model.

Yonatan Belinkov and James Glass [7] compared the phrase-based statistical machine translation system and neural system. The systems experimented the external segmentation tools and subword modeling by character-level neural models on Arabic-Hebrew translation tasks. All models were trained with a number of large-scale parallel Arabic-Hebrew corpora, OpenSubtitles which is mostly from spoken language. For phrase-based statistical machine translation, the system use Moses toolkit and grow-diag-final for word alignment, lexical reordering follows msd-bidirectional-fe and sentences are longer than 80 words. The system trained a 5-gram language model using KenLM and tune with MERT to optimize BLEU. For neural machine translation, the system use a Torch OpenNMT with the default settings. Although phrase-based statistical machine translation system and neural machine translation system reached the comparable performance, the neural machine translation has a small advantages in this system.

2.5.1 Word-based Neural Machine Translation System

The word-based neural machine translation model is the very first attempt in the neural methods. The disadvantage of this system may reduce the performance of the translation system because it translates the word by word translation of sentences. Luong and Manning [55] presented hybrid word-character neural machine translation system. This system mostly translates the word level and the character components for rare words to achieving open vocabulary neural machine translation system. The system trained three types with the WMT'15 English to Czech translation task: purely word-based, purely character-based, and hybrid by following the global attention mechanism together with similar hyper parameters. It is shown in Table 2.1.

Table 2.1 Hyper-parameters of Hybrid Word-Character NMT System

Hyper-parameter	**NMT models**
Number of hidden units	1024 cells
embeddings	1024-dimensional
epoch	6
learning rate	1.0
mini-batches	128
dropout	0.2

The gradient is rescaled whenever its norm exceeds 5. The hybrid approach offers an addition boost of BLEU points over models that already handle unknown words. The best system achieves 20.7 BLEU score.

2.5.2 Character-based Neural Machine Translation System

Nowadays, the popular machine translation systems, phrase-based or neural-based, have made experiments on word-level model basically. In neural machine translation, the result has reached to the great success. Despite their remarkable success, word-level neural machine translation models suffer from several weaknesses. The researchers addresses for this weaknesses with rare word model, out-of-vocabulary words model in rich morphology languages such as Czech, Finnish and

Turkish. In addition, they have tried to solve this issues bases on the character level neural machine translation systems.

Wang Ling et al. [52] introduced character-based on both side of source and target sides neural translation system using attention model. The system is capable of interpreting and generating unseen word forms and removes the challenges of tokenization of preprocessing step. The system implements in two datasets. In the English-Portuguese language pair, there are 600K sentence pairs for training from Europarl, 500 sentence pairs for development and 500 sentence pairs for testing. In the English-French language pair, there are 20K sentence pairs for training from BTEC, development sets is from CSTAR03 and test sets is from IWSLT04 ,respectively. The system showed that its methods can improve over equivalent word-based neural translation models.

Junyoung Chung et al. [13] experimented a subword-level encoder and a character-level decoder with an attention-based mechanism. The system trained with four language pairs. using the parallel corpora from WMT'15. In this corpus, they contained 12.1M, 4.5M, 2.3M and 2M sentence pairs, respectively. And the system tested three models settings: BPE!BPE, BPE!Char (base) and, BPE!Char. For forward direction and reverse direction, the encoder has 512 hidden units. For forward direction and reverse direction, the decoder has 1024 hidden units per layer. Each model is computed using a minibatch of 128 sentence pairs and trained using stochastic gradient descent with Adam. The system showed that the models with a character-level decoder outperform on a subword-level decoder.

Marta R. Costa-juss`a and Jos´e A. R. Fonollosa [15] proposed the standard encoder and decoder neural machine translation system with attention, and introduce the character-based source word embedding to use unlimited-vocabulary. The system used the German-English WMT data. For preprocessing step. the system used tokenizing, truecasing, normalizing punctuation and filtering sentences. For phrase-based system, it is built using Moses with grow-final-diag, Good-Turing smoothing of the relative frequencies, 5-gram language modeling using Kneser-Ney discounting, and lexicalized reordering, among others. The neural-based system was built using DL4MT2 tool. The system used an embedding of 620, a dimension of 1024, a batch size of 32, and no dropout, vocabulary size of 90 thousand words in German-English. The system showed that the proposed method is improved.

Shenjian Zhao and Zhihua Zhang [116] presented an efficient architecture to train a character-level neural machine translation by introducing a decimator and an interpolator. Before encoding, the decimator is used for the source sequences and the interpolator is used to resample after decoding. The system implemented the model using Theano tool and Blocks tools. The system train on the task of English-to-French translation, which is the bilingual, parallel corpora provided by ACL WMT'14. contains totaling 850M words. The system used newstest2013 as the development set and evaluate the models on the newstest2014 which consists of 3003 sentences not present in the training data. Preprocessing consists only the usual tokenization. The detailed trainings are shown in Table 2.2. As a result, the character-level modeling had solved the out-of-vocabulary (OOV) issue that word-level models suffer from.

Table 2.2 Hyper-parameters of Character-level NMT System

Hyper-parameter	**NMT models**
Number of hidden units	1024 cells
word embedding	620
frequent words	30,000
character embedding	64
frequent characters	120
mini-batches	80

Zhen Yang et al. [109] proposed a novel character-aware neural machine translation mode that learn to encode at the character level. The proposed model is also applied to the language without explicit word segmentation, because this model did not rely on the boundaries. The model was trained on Chinese-English translation tasks which consists of 2.3M parallel sentences and use the open-source neural machine translation system, GroundHog, as the baseline system. The system tested the four model, namely, RNNsearch-Word, RNNsearch-Char, character-aware-forward and character-aware-recurrent. The system used the BLEU metric to evaluate the translation quality and test the translation. Experimental results on Chinese-English translation tasks show that the proposed character-aware neural machine translation model can achieve comparable translation performance with the traditional word based neural machine translation models.

Table 2.3 Hyper-parameters of Character-to-Character NMT System

Hyper-parameter	**bpe2char**	**char2char**
Vocab size	54,544	400
Source embedding	512	128
Target embedding.	512	512
Pool stride	-	5
Highway	-	4 layers
Encoder	1-layer 512 GRUs	1-layer 512 GRUs
Decoder	2-layer 1024 GRUs	2-layer 1024 GRUs

Jason Lee et al. [48] introduced a character-to-character neural machine translation model without any segmentation. The character-to-character model outperforms a recently proposed baseline with a subword-level encoder on WMT'15. And it gives comparable performance on FI-EN and RU-EN. The system experiment two different scenarios: a bilingual setting and a multilingual setting. The system trained on a single language pair for bilingual setting and a single model on data from all four language pairs for multilingual setting. A single language pairs are German-English, Czech-English, Finnish-English and Russian- English. A single model to translate four languages pairs are German, Czech, Finnish and Russian to English, called multilingual setting. The system used all available parallel data on the four language pairs from WMT'15 and tested (a) bilingual bpe2bpe, (b) bilingual bpe2char, (c) bilingual char2char, (d) multilingual bpe2char and (e) multilingual char2char. Table 2.3 show the detail training settings. In this multilingual setting, the character-level encoder significantly outperforms the subword-level encoder on all the language pairs.

2.5.3 Subword Level Neural Machine Translation System

One of the main problems of machine translation systems is the out of vocabulary(OOV) problem. Most languages are restricted to the fixed size vocabulary. One of the common solutions is to use back-off methods and subword unit model. The same problem also appears in neural machine translation.

As a consequence, Ilya Sutskever et al. [58] proposed and implemented an effective techniques to address the rare word problem in neural machine translation system. The system presents a simple alignment-based technique. The system train an neural machine translation system on data that is augmented by the output of a word alignment algorithm, allowing the neural machine translation system to emit, for each OOV word in the target sentence, the position of its corresponding word in the source sentence. They had demonstrated empirically that on the WMT'14 English-French translation task. And the system showed that their technique outputs a consistent and substantial improvement of up over various neural machine translation systems. The hyper parameter that is used in the training multi-layer deep long short-term memory(LSTMs). Each of LSTM has 1000 cells with 1000 dimensional embeddings. And the parameters are initialized uniformly in [-0.08, 0.08] for 4-layer models and [-0.06, 0.06] for 6-layer models. SGD has a fixed learning rate of 0.7, 8 epochs and 128 batch size. Training takes about 10-14 days on an 8-GPU machine.

In recent year, Rico Sennrich et al. [85] first described a byte pair encoding approach. This approach is to aggregate frequent adjacent characters into n-grams subwords. This is based on the various word classes that are translatable via smaller units than words. The paper contributed that neural machine translation systems can translate rare and unseen words as a sequence of subword units. The system showed that subword models improve over a back-off dictionary baseline. And the system also showed performance gains over the baseline with both BPE segmentation, and a simple character bigram segmentation.

2.5.4 Low Resource Neural Machine Translation System

The quality of machine translation models relies exclusively on available parallel sentences. When large quantities of parallel sentences is not available, machine translation quality tends to be poor. In recent years, neural machine translation model has achieved a great success in rich resources languages, which has a great deal of parallel corpora. However, it does not work well for under resource languages. In order to solve this problem, the researchers had tried with the various ways such as using the monolingual corpus, training a transfer learning method, etc.

As mentioned in section 2.4, construction of parallel corpus is expensive and expertise. And there are often nonexistent for low-resource languages. Conversely, monolingual corpus is much easier to collect and the significant amounts of monolingual data are in many languages. Caglar Gulcehre et al. [26] used monolingual corpus in neural machine translation system. The system proposed two methods, shallow fusion and deep fusion, for adding monolingual data into an existing neural machine translation system. The system applied the proposed approaches on four language pairs: Czech to English, Turkish to English, German to English and Chinese to English. Chinese to English language pairs from OpenNMT'15 are sentence-aligned pairs from three domains such as SMS/CHAT, conversional telephone speech and newsgroups/weblogs. The training set consists of 430K sentence pairs in total. Turkish to English languages pairs use WIT parallel corpus and SETimes parallel corpus. These corpus are as a part of IWSLT’14 and consists of the sentence-aligned subtitles of TED and TEDx talks. The training data for German to English and Czech to English languages pairs is from WMT'15, and these languages pairs are high-resource languages. The monolingual corpus is the English Gigaword corpus by the Linguistic Data Consortium (LDC), which mainly consists of newswire documents. The system shows that incorporating monolingual corpora can improve a translation system on a low-resource language pair (Turkish-English) and a domain restricted translation problem (Chinese-English SMS chat). In addition, they showed that these methods improve the performance on the relatively high-resource German-English and Czech-English translation tasks.

Rico Sennrich et al. [84] investigated the use of target side monolingual data for neural machine translation system. The system explored strategies to train with monolingual data without changing the neural network architecture of the previous work. By pairing monolingual training data with an automatic back translation, the system can serves it as additional parallel training data, and they obtain substantial improvements on the WMT 15 task English-German, and for the low-resourced IWSLT 14 task Turkish-English, obtaining new state-of-the-art results. The system also shows that small amounts of in domain monolingual data, back-translated into the source language, can be effectively used for domain adaptation.

Previous researchers have proven that the target-side monolingual data can greatly improvement for neural machine translation. Jiajun Zhang† and Chengqing Zong [115] proposed two approaches: a self-learning algorithm and a new multi-task learning framework, to use the source side monolingual data in neural machine translation systems. The self-learning algorithm generates the synthetic parallel corpus and enlarge the bilingual training data to enhance the encoder model of neural machine translation. The multi-task learning framework performs machine translation on bilingual data and sentence reordering on source-side monolingual data by sharing the same encoder network. The dataset is performed two tasks on Chinese-to-English translation: one for small data set (0.63M sentence pairs) and the other for large-scale data set (2.1M sentence pairs including the small training data). For the source-side monolingual data, there are about 20M Chinese sentences. Each neural machine translation model is trained on GPU K40. The system used mini batch size of 32, the word embedding dimension of source and target language is 500 and the size of hidden layer is set to 1024. The training time for each model ranges from 5 days to 10 days for small training data set and ranges from 8 days to 15 days for large training data set8. We use case-insensitive 4-gram BLEU score as the evaluation metric. The extensive experiments demonstrate that the proposed methods obtain significant improvements over the strong attention-based neural machine translation. And the proposed multi-task learning framework performs better than the self-learning algorithm at the expense of low efficiency. Furthermore, the experiments also demonstrated that neural machine translation is more effective for incorporating the source-side monolingual data than conventional statistical machine translation. It is also observed that more monolingual data does not always improve the translation quality and only relevant data does help.

Guillaume Lample et al. [47] presented a new approach where translation model is learned using monolingual corpus only, without any aligned sentences or documents for both source side and target side monolingual corpus. The principle of the approach is based on word-by-word translation model. The model is trained on two widely used datasets and two language pairs. And this system had shown that the proposed model get BLEU scores up to 32.8.

Another way to resolve the problem of lack of the resources in low resource languages is to use the transfer learning method. Generally, transfer learning method is to first train a high-resource language pair, called the parent model. And then these model parameters are transferred to the low-resource language pair, called the child model to initialize and constrain training. Therefore, Barret Zoph et al. [119] firstly described a transfer learning method for improving low resource neural machine translation systems. In this paper, the system used French to English language pairs for parent model and Hausa to English, Turkish to English, Uzbek to English, and Urdu to English languages pairs for child model. For parent model training, it is trained with a dropout probability of 0.2. The learning rate is 0.5, decay rate is 0.9, minibatch size is 128, hidden size is 1000, a target vocabulary size is 15K and a source vocabulary size is 30K. For child model training, it is trained with a dropout is 0.5, the learning rate is 0.5, decay rate is 0.9, minibatch size is 128, hidden size is 1000, a target vocabulary size is 15K, a source vocabulary size is 30K and 100 epochs. Overall, the transfer learning method improves neural machine translation scores on low-resource languages. Our experiments suggest that there is still room for improvement in selecting parent languages that are more similar to child languages, provided data for such parents can be found.

Tao Feng et al. [24] exploited encoder-decoder framework with attention mechanism for parent model and takes some parameters of the parent model to initialize the child model. There are two language pairs on transfer learning method. The first one is that French-English neural machine translation model is the parent model and Vietnamese-English model is child model. In the other one, English-Chinese model is the parent model and Mongolian-Chinese model is the child model. French-English corpus contains 2 million sentence pairs, 50 million English words and 52 million French words, that is from the WMT14 workshop. English-Chinese corpora contains 2 million sentence pairs, 22 million English words and 24 million Chinese words from the WMT2017. Vietnamese-English corpora contains 133K sentence pairs, 2.7 million English words and 3.3 million Vietnamese words, is provided by IWSLT2015. Mongolian-Chinese corpus contains 67K sentence pairs, 848K Chinese words and 822K Mongolian words, is provided by CWMT 2009. The system used a two-layers bi-directional RNN in the encoder, another two-layers uni-directional RNN in the decoder, LSTM cells with 1024 units, the dimensionality of word embedding is set to 1024, minibatch size is 128, sentence length of 50, dropout for parent model is a probability of 0.2 and 0.5 for child model. The experiments demonstrated that the proposed method can achieve the excellent performance on low-resource machine translation by weight adjustment and retraining.

Toan Q. Nguyen and David Chiang [70] presented a simple method to improve neural translation of Turkish-English and Uyghur-English languages pairs with the help of Uzbek-English neural translation model. The system did the experiments the models based on word-level translation, but not always significantly. In advance, the system firstly split words into subwords using Byte Pair Encoding to increase vocabulary overlap in not only source languages but also target languages. And then it trained a parent model and transfer its parameter to a child model. When used of a much stronger BPE baseline, it yields larger and statistically significant improvements. The system trained the models with a minibatch size of 32 and dropout rate is 0.2. For the Uzbek-English to Turkish-English experiment, the parent and child models were trained for 100 and 50 epochs, respectively. For the Uzbek-English to Uyghur-English experiment, the parent and child models were trained for 50 and 200, respectively. As mentioned above, the BPE models were trained for half as many epochs because their data is duplicated.

2.5.5 Evaluation Metrics for Machine Translation Systems

Human evaluations of machine translation are extensive but expensive [118]. There are many evaluation metrics for machine translation: BLEU, EBLEU, NIST, METEOR, METEOR-PL, TER and RIBES. Previously, BLEU (Bilingual Evaluation Understudy) was one of the popular evaluation metrics. Alternatively, the RIBES(Rank-based Intuitive Bilingual Evaluation Score) evaluation metrics is more sensitive for reordering and has shown to have better correlations with human judgements. BLEU scores can be computed either at a document level or at a sentence level. They range between 0 (or 0% – lowest quality = completely irrelevant to the reference) and 1 (or 100% – highest quality = same as the reference). RIBES score was introduced by adding a rank correlation coefficient prior to unigram matches. The focus of the RIBES metric is word order. It uses rank correlation coefficients based on word order to compare statistical machine translation and reference translations. In this research, BLEU and RIBES are used to evaluate the quality of neural machine translation systems.

CHAPTER 3

Myanmar Language

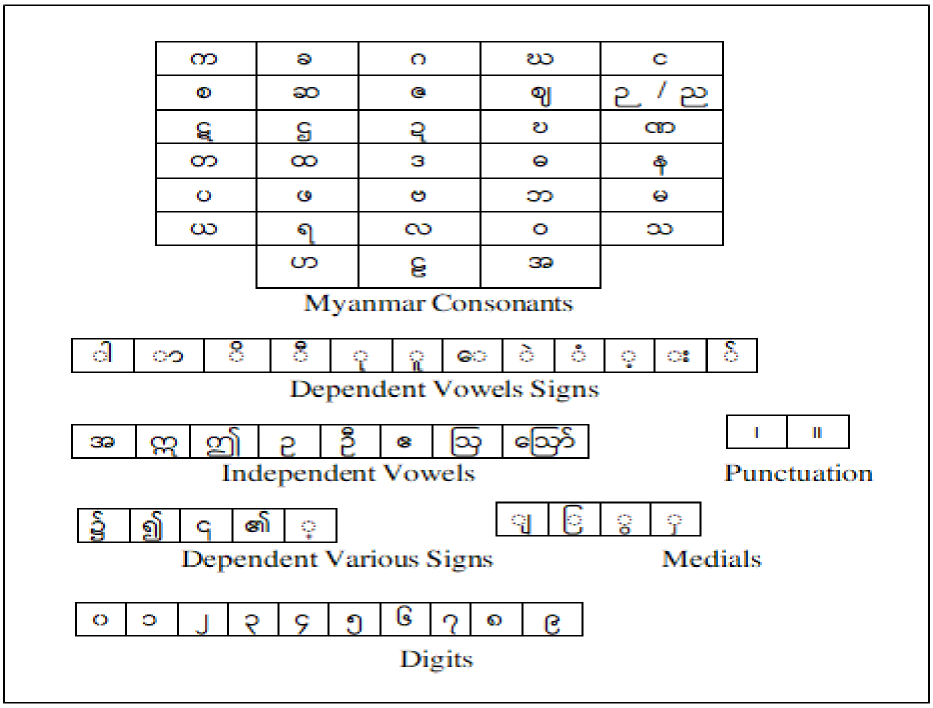
This chapter presents the introduction to Myanmar language. In addition,, nature of Myanmar language and Myanmar grammar. Introduction of Unicode, Importance of Unicode and Myanmar Unicode are also discussed.

3.1 Introduction to Myanmar Language

Myanmar language is the official language of Republic of the Union of Myanmar. And it is also the native language of Myanmar. Myanmar is a group of the Tibeto-Burman group. Myanmar language is spoken about 34.5 million people as their first language. Although ethnic groups have their own mother tongue, they speak the Myanmar language as the second language. Myanmar Language has its own script and it is syllable based.

3.2 Nature of Myanmar Language

According to the history, the Myanmar script was originally adopted from the Mon script. This Mon script was derived from Pali that is the ancient Indian language of the text of Theravada Buddhism.



က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ဉ / ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
	ဟ	ဠ	အ	

Myanmar Consonants

◌ါ	◌ာ	◌ိ	◌ီ	◌ု	◌ူ	◌ေ	◌ဲ	◌ံ	◌့	◌း	◌်

Dependent Vowels Signs

အ	ဣ	ဤ	ဥ	ဦ	ဧ	ဩ	ဪ

Independent Vowels

၊	။

Punctuation

၌	၍	၎	၏	◌့

Dependent Various Signs

◌ျ	◌ြ	◌ွ	◌ှ

Medials

၀	၁	၂	၃	၄	၅	၆	၇	၈	၉

Digits

Figure 3.1 Myanmar Character Patterns

Myanmar script consists of (33) consonants, Independent vowels, Dependent consonant signs (also known as Medials), Dependent vowels signs, Dependent various signs (also known as Parli Word), punctuation and digits. They can be seen in Figure 3.1. Myanmar language is written from left to right. There are two language styles: Spoken style and Written style. And the structure of the sentences is subject-object-verb (SOV). Like other Southeast Asia languages, Myanmar language does not define how to place the spaces between words. It is usually written continuously without using space. Sometimes, it is written in spaces between phrases. Sentences can be easily determined with sentence boundary maker "။" which is called ပုဒ်မ and pronounced as "Pou ma ". However, there is no rule how to write definitely in Myanmar language[67].

Table 3.1 Formation of Myanmar Sentence

English Sentence	The doctor gave me this prescription.												
Myanmar Sentence	ဒီဆေးညွှန်းကဆရာဝန်ငါ့ကိုပေးလိုက်တာ။												
Myanmar Phrases or clauses	Noun Phrase				Noun Phrase			Noun Phrase		Verb Phrase			Punctuation
	ဒီဆေးညွှန်းက				ဆရာဝန်			ငါ့ကို		ပေးလိုက်တာ			။
Myanmar Word	ဒီ	ဆေးညွှန်းက			ဆရာဝန်			ငါ့ကို		ပေးလိုက်တာ			။
Myanmar Syllables	ဒီ	ဆေး	ညွှန်း	က	ဆ	ရာ	ဝန်	ငါ့	ကို	ပေး	လိုက်	တာ	။
Myanmar Characters	ဒ ဆ ည န က ဆ ရ ဝ င ပ လ တ ◌ီ ◌ေ ◌း ◌ွ ◌် ◌ာ ◌ါ ◌့ ◌ိ ◌ု ။												

A word consists of one or more syllables. And a syllables is composed of one or more character. But sometimes, a word may be only consonant without any vowel. The formation of Myanmar sentence is shown in Table 3.1 and the positioning of characters in a Myanmar syllable is shown in Figure 3.2.

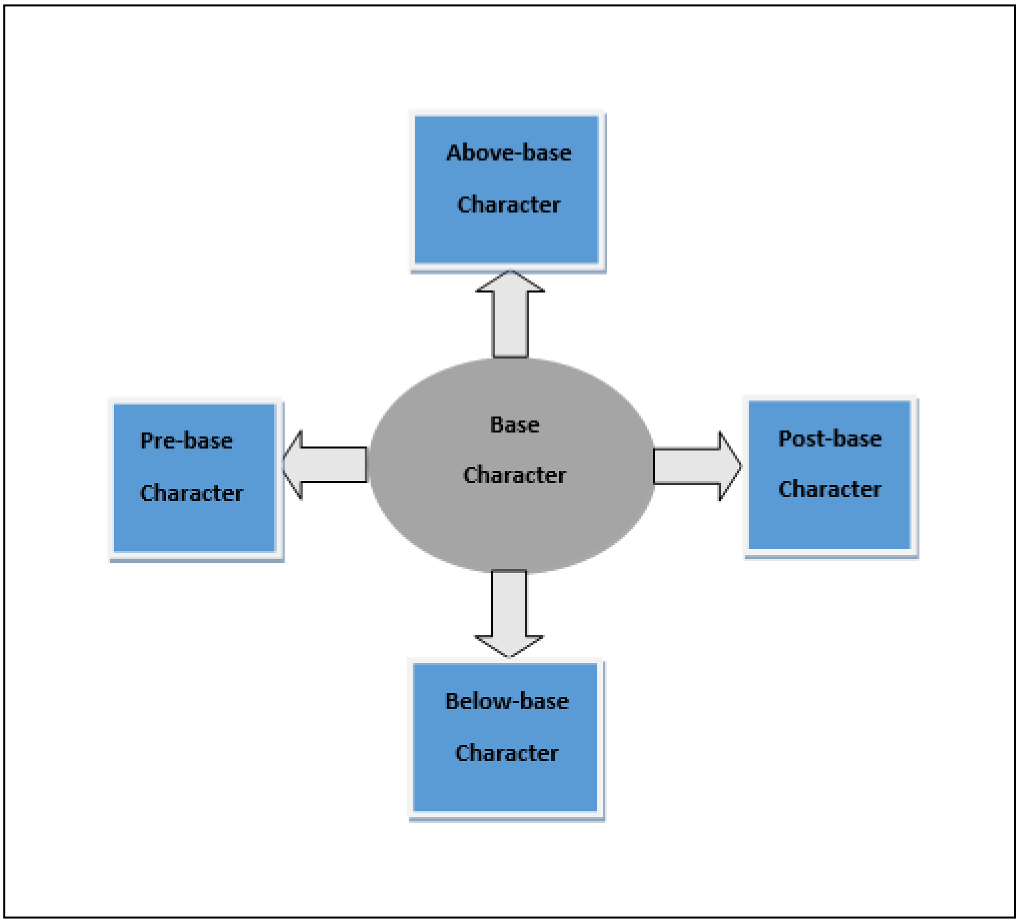


Figure 3.2 Positioning of Characters in a Myanmar Syllable

3.3 Myanmar Grammar

Morphologically, Myanmar words can be combined to make new word. Therefore, Myanmar language may be agglutinative and inflective language. Like Chinese, morphemes can be combined freely with no changes. However, Myanmar language is syntactically quite similar to Japanese and Korean languages because Myanmar language is typically head-final language.

Grammatically, there are nine main part-of-speeches in Myanmar language, which are shown in Figure 3.3. They are Noun, Pronoun, Adjective, Verb, Adverb, Particle, Post-positional, Conjunction, and Interjection.

3.3.1 Noun

In general, nouns refer to a person, an object, or abstract ideas. Nouns can be a word or compound word. Nouns in Myanmar grammar are classified into two types. They are the four types of meaning or representation and the four types of constructions. The four types of meaning or representation are proper noun, abstract noun, common noun and collective noun. The four types of construction are indivisible noun, compound noun, verb modification noun and qualitative noun.

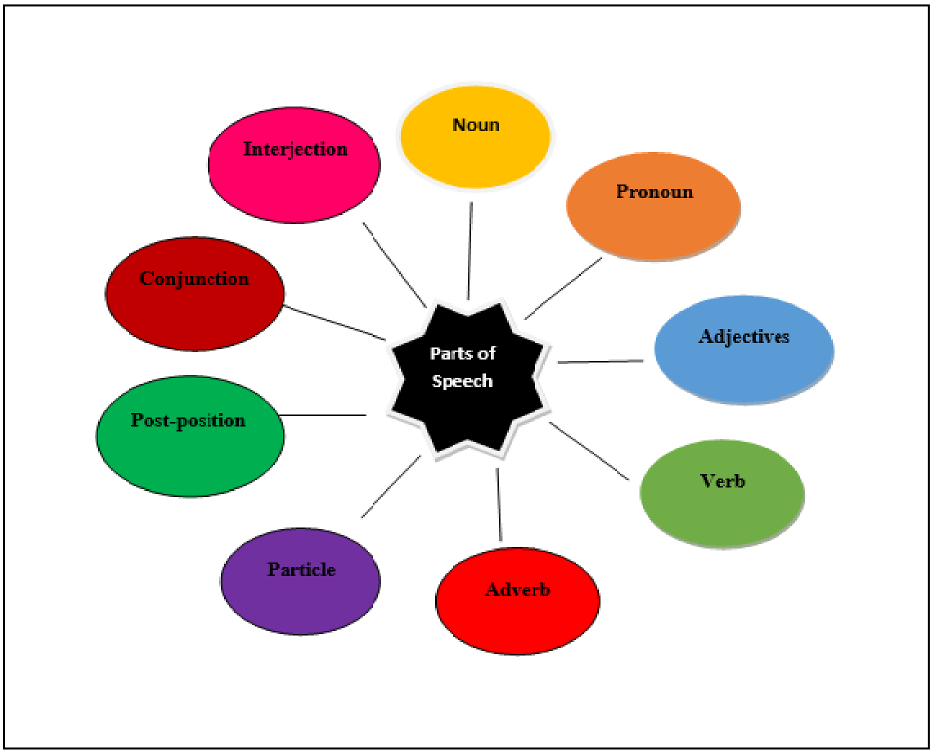


Figure 3.3 Nine main parts of speech in Myanmar Grammar

In Myanmar language, some nouns are usually started with အ(အပြုံး= smile) and ended with မှု(လှပမှု=beauty), ခြင်း(ကျန်းမာခြင်း= health), ရေး(ကျန်းမာရေး=health), ချက်(အားနည်းချက်= weakness). In Myanmar language, plural nouns are formed by suffixing the particle "များ၊တို့၊တွေ" into singular noun. "များ" is used in written style. "တို့၊တွေ" is used in spoken style. The Myanmar language recognizes no grammatical or artificial gender, but that only which consists in the distinction of the sexes, viz, the masculine and the feminine. Particles as gender indicators are prefixed before noun. "ဦး၊ကို၊မောင်၊ထီး" is used for male gender and "ဒေါ်၊မ" is used for female. Particles are suffixes to the noun to describe the type of noun, known as measure words. For example, "ယောက်၊ဦး" is used for people. "ကောင်" is used for animals. "ခု" is used for the general classifiers. "လုံး" is used for round objects. "ပြား" is used for flat objects and "စု" is used for group objects. And, some nouns are identified with the combination of particle to the verb or the adjective. Likewise the particles, post-positional makers (PPM) are also suffixed to the end of the noun. They are explained in the postposition.

3.3.2 Pronoun

Pronoun is used to address or refer to people or things. In Myanmar language, there are four classifications of pronouns: personal pronoun, referential pronoun, question pronoun and mathematical pronoun.

Personal pronoun is used in place of a person. "I= ကျွန်တော်", "You= သင်", "He= သူ", "She= သူမ", "It= ၎င်း" is the Personal Pronoun in Myanmar language. Referential pronoun means the pointing to something or someone such as, "this = ဤ ", "that= ထို ". Question pronouns refers to question words that are equivalent to English words such as "what", "who", "where". For example, in the Myanmar question: "What does she like?", "what= ဘာ" refers to "the thing that she likes" (noun), and it is considered a question pronoun. By the same logic, "where" in the question: "Where did he go?", "where= ဘယ်" refers to "the place that he went" (noun), and it is a question pronoun. Mathematical pronoun equivalent to "one person", "three cups", "four items", "some", "few", "all", "half", etc.

3.3.3 Adjective

Adjective is a word used to modify the noun. Myanmar adjectives are usually ended with the particles "သော". Myanmar adjectives are classified into 3 levels. In normal adjectives is ended with "သော". In comparative adjectives, the particles "ပို၍၊ သာ၍" is prefixed to the adjectives. In superlatives adjectives, the particles "ဆုံး" is suffixed to the adjectives.

Adjectives in Myanmar grammar are classified into two types. They are the four types of meaning or representation and the two types of constructions. The four types of meaning or representation are qualitative adjectives, referential adjectives, mathematical adjectives and questionnaires of adjectives. The two types of construction are indivisible adjectives (fast= မြန်သော) and compound adjectives (clear= သန့်ရှင်းသော).

Qualitative adjectives are words used to modify the quality of the noun as how to something or someone is. For example (“rich= ချမ်းသာသော"). Referential adjectives are words that make reference to something or someone. ဤ/this, သည်/this, ထို/that, အခြား/other, etc., are referential adjectives. Mathematical adjectives are words used to describe "how many" of something or someone, "what position" in an ordered list of the something, and unspecified numbers are in this category. Furthermore, it is classified into three types. They are quantitative, ordinal numbers and unspecified numbers. Quantitative adjectives are the words that described numbers followed by measure words. For example: (two cats= ကြောင်နှစ်ကောင်). Ordinal numbers are the words that show position in the ordered list of numbers such as "first, second and third." For example: (21st birthday= ၂၁ကြိမ်မြောက်မွေးနေ့), there, မြောက် is an ordinal number of adjective. Unspecified number of adjectives is the words that are used as quantifiers without numbers. အားလုံး/all, အချို့/some, etc are unspecified number of adjectives. Questionnaires of adjectives are equivalent of English words. မည်မျှ/how many, မည်သို့သော/which, etc., are questionnaires of adjectives.

3.3.4 Verb

Verb is a word that shows the action, event, and condition. Myanmar verb can usually be identified by the combination of root word, prefix and suffix. In Myanmar language, the roots of verbs are always followed or suffixed with one or more particle. These particle conveys tense information, intention information, politeness information, mood information, etc. The suffix "သည်၊၏၊ပြီ" can be used as a marker making the present tense statement and also as a verb postpositional marker. The suffix "ခဲ့သည်၊ခဲ့၏၊ခဲ့ပြီ" can be used as a marker making the past tense statement and as a verb postpositional marker. The suffix "နေသည်" can be used to describe an action in progression of happening and equivalent to the English ‘-ing’.

The suffix "မည်၊လိမ့်မည်၊လတ္တံ့၊အံ့" can be used as a marker making the future tense statement and as a verb postpositional marker. Myanmar verbs are negated by the particle "မ", and "မ" is prefixed or infixed to the verb. Usually the marker "ပါ၊ဘူး" are used with negative verb. For example; (သွား =go, မသွား = does not go),(စာရေး = write, စာမရေး =does not write).

3.3.5 Adverb

Adverb is a word used to modify the verb. Myanmar adverbs are usually ended with the particles "စွာ". There are five types of Myanmar adverbs.

စောစော/early, ကြာမြင့်စွာ/so long time, မကြာခဏ/sometimes, ချက်ချင်း/at once, ယခင်/previously are identified as time indicator of adverb. ရိုသေစွာ/respectfully, လျင်မြန်စွာ/quickly, ခပ်ပြုံးပြုံး/smiley are Manner indicator of adverb. ဧကန်/surely, စင်စစ်/surely, မုချ/surely, သည်းထန်စွာ/heavily are situational indicator of adverb. အလွန်/very, နည်းနည်း/a little, လုံးဝ/all are quantity indicator of adverb. And မည်မျှ/how many, မည်သို့/which, ဘယ်လို/how, ဘယ်လို/how are questionnaires indicator of adverb. Additionally, some adverbs are formed by combining opposite words: positive and negative. For example, ကောင်းမကောင်း/good or bad and လုပ်သင့်မလုပ်သင့်/should do or not.

3.3.6 Particle

Particle is a word that are serving to qualify on a noun, pronoun, adjective, verb, and adverb. Particles are suffixes that after noun, pronouns, verb, adjective and adverb. And particles are untranslatable words. Some Particles are များ, တို့, သော, သည့်, မည့်, သာ, သင့်, ပင်, ဝံ့, ရက်, ရှာ, တော့, နှင့်, ချည်း, ပါ, နှယ် and etc.

3.3.7 Post-positional(PPM)

Post-positional are a word that are followed or suffixed to a noun, pronoun and verb. Noun with PPM and pronoun with PPM designate as the subject and object.

Verb with PPM indicate time and mood. Although the particles are untranslatable words, some post-positional makers are translated words except subject maker and object maker. The following Table 3.2(a) is post-positional makers of the noun.

Table 3.2(a) Noun PPM

No	Makers	PPM	Examples
1.	Subject Makers	သည်၊က၊မှာ	ကျွန်တော်သည်(I)
2.	Object Makers	ကို	နှင်းဆီပန်းကို(rose)
3.	Receiver Makers	အား	လူနာအား(patient)
4.	Place Makers(Location)	၌ ၊ မှာ၊ တွင်၊ ဝယ်၊ က	at, on, in
5.	Place Makers(Departure)	မှ၊ က	from
6.	Place Makers(Destination)	သို့၊ ကို	to
7.	Place Makers(Direction)	သို့၊ ကို ၊ ဆီသို့	to, towards
8.	Place Makers (Continuation of place)	တိုင်တိုင် ၊ အထိ	until, till
9.	Time Makers	မှာ၊တွင်	at, on, in
10.	Continuous of Time	တိုင်တိုင်၊အထိ	up to, till
11.	Instrumentality Makers	ဖြင့်၊နှင့်အတူ	by, with
12.	Cause Makers	ကြောင့်၊သဖြင့်	because, because of
13.	Possessive Makers	၏၊ ရဲ့	's
14.	Accordance Makers	အလိုက်၊ အရ	as, according to
15.	Accompaniment Makers	နှင့်၊နှင့်အတူ၊ နှင့်အညီ	and, with
16.	Choice Makers	တွင်၊အနက်၊မှ၊ထဲမှ	between, among
17.	Purpose Makers	ဖို့၊အတွက်၊ရန်	to, for

There are two kinds of verb PPM. They are three types of tense and four kinds of types. Verb PPM is showed in Table 3.2(b).

3.3.8 Conjunction

Conjunction is a word to connect between words, phrases, and sentences. To join similar items, conjunctions are used. They are also used to show continuity things in English. For example "and= နှင့်", "with= နှင့်အတူ", "moreover= ထို့အပြင်", etc. To describe contrasts, conjunctions are also used such as "however= သို့ရာတွင်", "but= သို့ပေမယ့်" and "nevertheless= မည်သို့ပင်ဆိုစေကာမူ". Besides, conjunctions are also used to express constraints and challenges like the English words. For example "despite= သို့ဖြစ်စေကာမူ" and "in spite of= သော်လည်း". Conjunctions are also used to show "or else" choice, logical consequence such as "therefore= ထို့ကြောင့်", and "as a result= ရလဒ်အနေဖြင့်", or to describe the desired end result such as "in order to= ဖြစ်စေရန် " and "so that= စေရန်".

Table 3.2(b) Verb PPM

No	Makers	PPM
	Three types of tense	
1	Present Tense	သည်၊၏ ၊ပြီ
2	Past Tense	ခဲ့သည်၊ခဲ့၏ ၊ခဲ့ပြီ
3	Future Tense	မည်၊လိမ့်မည်၊လတ္တံ့
	Four types of tense	
1	command [literary] Maker	လော့
2	Consensus Maker	စို့၊ရအောင်
3	sympathy and mercy Maker	ပါ၊ရစေ
4	Judgment Maker	စေ

3.3.9 Interjection

An interjection expresses some emotional word. An interjection can be used as filler words. Interjections do not have rule related to a grammatical function of the sentence. Moreover, these words are not related to the other parts of the sentence. If an interjection word is omitted in the sentence, this sentence still makes meaningful. It is able to stand alone. For example: (အမလေး=Alas!, မီး....မီး, လိုက်ဟ.....လိုက်ဟ..... , အို=Oh!, ဟာ, ဟင်).

3.4 Introduction of Unicode

Unicode is a standard character set encoding. It is developed and implemented by the Unicode consortium. It is computing for industry deployment. The Unicode character set is the capacity to support over one million characters. And the Unicode character set aims to support from all scripts and many symbols as a single character set. The Unicode is used around the world today and in the past. And it is well on the way to become a dominant and ubiquitous standard. Currently, the standard character set encoding supports over 96,000 characters that are from a large number of scripts. Computers represent a character that are stored as numbers. And these characters are provided by Unicode as a unique. It is platform, and language independent.

Unicode achieved success at unifying character sets. It is widespread and predominant used around the world in computer field. The Unicode standard has been implemented in many technologies. The Unicode standard has been developed in modern operating systems, XML, programming languages, and the .NET Framework. The Unicode standard can be defined by several character encodings. The common usages are UTF-8, UTF-16, and UTF-32. The most commonly used encodings are UTF-8. UTF-8 is the dominant character encoding on the World Wide Web.

To support Myanmar language, Unicode fonts required Zawgyi-One and other pseudo-Unicode fonts twice as many characters sets. Unicode font is unlike Zawgyi-One and other pseudo-Unicode fonts. It uses intelligent rendering to stack consonants and combine diacritics[98].

3.5 Importance of Unicode

Unicode is an critical step towards standardization from a translation point of view. Unicode is a single software product. This is designed for multiple platforms, languages and countries. These can lead to a signification reduction in cost over the use of legacy character sets. And Unicode can be used in many systems that do not corrupt data. Therefore, Unicode represents a single encoding scheme for all languages and characters. It is the preferred encoding scheme used by XML-based tools and applications. Therefore, Unicode is an important for translation applications[98].

3.6 Myanmar Unicode

The Myanmar Unicode Code defined in Unicode 5.1. And the Myanmar Unicode Code points in the code space from U+1000 to U+109F. Unicode always uses the same code point for the same character or semi-vowel, even if it changes shape depending on the context. Myanmar font scripts which follow Unicode rules are Myanmar3, Parabaik, Padauk, Thanlwin, WinuniInnwa, Win Myanmar, MyMyanmar, Yunghkio, Panglong, and Tharlon. It is the international accepted standard by the World Wide Web Consortium, the main international standards organization for the World Wide Web. The fonts necessary to view and edit are freely available. Search is seamless with Unicode. Unicode makes it extremely easy to translate the Wikipedia's interface. Unicode fonts support 11 languages that use the Myanmar script: Burmese, 2 liturgical languages: Pali and Sanskrit, 8 minority languages: Mon, Shan, Kayah, four Karen languages and Rumai Palaung.

CHAPTER 4
Building Myanmar-English Parallel Corpus and Myanmar Monolingual Corpus

A corpus is a very large collection of text or speech produced by real users of the natural language and may contain texts or speech in a single language, called monolingual corpus or in two languages, called parallel corpus or in many languages, called multilingual corpus. The scope of the corpus is endless in computational linguistics and natural language processing[32].

Monolingual corpus is the most frequent type of corpus and contains texts in one language only. A parallel corpus is a collection of text or speech in one language and their translation languages. Parallel corpora can be bilingual corpus or multilingual corpus. In most cases, parallel corpora contain data from only two languages, the texts of one corpus are the translation of another corpus. The order of the translation may be sentence by sentence, phrase by phrase, and word by word and the sentences, phrases, and words are needed to be aligned and matched. A parallel corpus is very useful for language learning process, cross-language information retrieval, and electronic dictionary; especially for machine translation system.

Creation of parallel corpus is an essential step for machine translation systems[109]. It is also the first step in building a translation model for low resource languages where there is no pre-created parallel corpus. Although there are many parallel corpus available for high resource languages, creation of large one manually is hard and time consuming. And creation of small parallel corpora is manually simpler and easier, but it is a very tedious, time consuming and expensive task for large one manually[77]. As a consequence, the researchers use comparable corpora as a resource which affects the performance and accuracy of the system.

Nowadays, the machine translation is a very challenging research task in NLP and the demand of it is growing in the world. Lots of machine translation systems have been developed all over the world using several pairs of major natural languages, such as English to (Arabic, Bengali, Chinese, French, Hindi, Japanese, Spanish, and Urdu)[32]. Myanmar language is one of the low resource languages and there is no proper data for Myanmar-English machine translation system.

4.1 Existing Myanmar-English Parallel Corpus

Today, there are already many parallel corpora in many languages, especially for rich resource languages. However, only a few parallel corpora for Myanmar-English language pairs are publicly available because Myanmar language is one of the low resource languages. And there were not many available parallel corpora between Myanmar and English languages.

There are existing parallel corpora for Myanmar-English languages pair. They are Basic Travel Expressions Corpus(BTEC) and Asian Language Treebank(ALT) corpus. BTEC corpus is collected about travelling data and it consists of nearly about 20,000 sentences. The ALT corpus is one part from the Asian Language Treebank project[81]. The ALT corpus began about 20,000 sentences from English Wikinews, and then these sentences were translated into other nine languages such as Filipino, Indonesian, Japanese, Khmer, Laotian, Malay, Myanmar, Thai, and Vietnamese. Nevertheless, the size of these corpus are not more than 20,000 parallel sentences. Actually, the performance of the NLP tasks is depending on the size of the data. Therefore, the performance of natural language processing tasks is still low.

4.2 UCSY-corpus: Myanmar-English Parallel Corpus

The UCSY-corpus is built aiming to promote machine translation research on Myanmar language. The UCSY-corpus is a general domain. The corpus consists of 200K Myanmar-English parallel sentences collected from different domains.

4.2.1 Data Collection

Myanmar language is regarded as a low resource language, so there are some difficulties to build a parallel corpus. Creating a Myanmar-English parallel corpus for Myanmar language is conditioned by various factors like the availability of the texts in that language.

Some parallel sentences are crawled electronically. If blogs and websites are available in both language (Myanmar and English), then crawl the data. And manual sanitization shall be needed to weed out insignificant data. Table 4.1 shows the names of the websites that are data collected.

Table 4.1 The Names of the Collected Websites

No	Website Name
1.	www.president-office.gov.mm
2.	www.pyidaungsu.hluttaw.mmthe
3.	www.pyithuluttaw.gov.mm
4.	www.amyotha.gov.mm
5.	www.mofa.gov.mm
6.	www.myanmarmoha.org
7.	www.ucsb.gov.mm
8.	www.oag.gov.mm
9.	www.unionsupremecourt.gov.mm
10.	www.moe.gov.mm
11.	www.statecounselloe.gov.mm
12.	www.moi.gov.mm
13.	www.myanmar-now.org/mm
14.	www.moi.gov.mm/npe/mal
15.	www.moi.gov.mm/npe/nlm
16.	www.ibiblio.org/obl/show.php
17.	www.english4mm.org
18.	www.bumarlibrary.org

Because there are not so many Myanmar-English bilingual websites and bilingual web pages are distributed, it is practical to use manually one site by site and one book by one book. Some parallel sentences are collected by downloading from the websites and copying form the eBooks. Some parallel sentences are collected from newspapers that write in the language in its script. This is a little hard because it might need to type the sentences in hand.

All the sources of the collected websites that contain Myanmar and English sentences are carefully selected and manually verified. The parallel sentences obtained from local news, Wikipedia news, travel domain, school text books and spoken text. Local news is from the government official websites such as Myanma Alin, The Global New Light of Myanmar. It starts to collect from January 2016 to

December 2017. Travel domain contains about 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). School text books and spoken text mainly contain the widely used spoken English. We get more than 200K sentences. Table 4.2 shows the statistics of UCSY parallel corpus. The corpus may be used to make many kinds of experiments such as word-based, Myanmar character-based and Myanmar syllable-based for translation models. Moreover, these corpus can also be used for other language processing tasks. Figure 4.1 shows the proportion of different domains of UCSY parallel corpus.

Table 4.2 Statistics of UCSY Parallel Corpus

Domain	**Number of Sentence pair**	**Number of Word**		**Number of Myanmar Syllable tokens**
		Myanmar	**English**	
Local News	72,720	1,902,346	1,500,305	3,675,211
ALT domain	20,123	698,347	436,923	1,138,297
Travel domain	94,414	830,925	567,151	1,263,330
School Text Book	19,433	173,676	122,842	273,962
Spoken Text	17,971	135,188	107,979	218,617
Total	224,661	3,740,482	2,735,200	6,569,417

4.2.2 Creation Process of UCSY-corpus

The creation process of UCSY corpus follows as illustrated in Figure 4.2. The creation process of UCSY-corpus has five major steps generally.

The **initial** step is to select the appropriate resources. These resources contain the Myanmar and English parallel sentences.

In the **second** step, these data are downloaded and copied or crawl the documents. These documents are ready for bilingual sentences.

And then, the type of documents is categorized into different domains types.

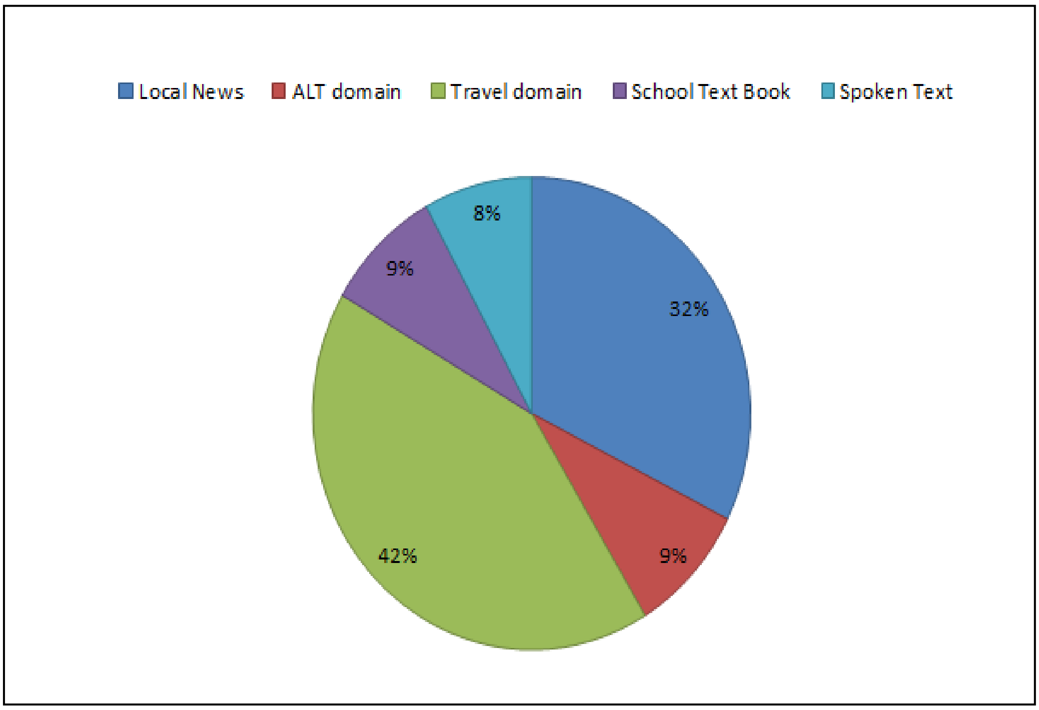


Figure 4.1 Proportion of Different Domains of UCSY Parallel Corpus

In order to be useful, a parallel corpus must be aligned at a certain level such as document level, paragraph level, sentence level or word level. A parallel corpus should be aligned at least at the sentence level. Sentences in a raw corpus are sometimes misaligned sentences[106]. As a result, the translated sentences do not corresponded with these sentences. Collected sentences are separated into sentences one per line. In this work, sentences are aligned manually. Therefore, sentences alignment stage is needed in the creation of parallel corpus.

The collected sentences include sometimes noise data. So, collected data is needed to be clean.

“Noise” not only can refer to the extraneous information in the plain texts, but also can indicate the illegal data[96]. Web pages often contain HTML tags, the color, font, style attribute of each piece of text and etc., This kind of noises can be removed. The next task, before sentence segmentation, is to remove the extra spaces in the collected sentences.

Although English or other similar language could only have a single space, Myanmar language does not define how to write these spaces. Therefore, the removal of extra whitespaces is important in Myanmar language. By removing such extra whitespaces, it not only can transform the various text formats into a unified one, but also can improve the accuracy of natural language processing tasks that use this corpus. And the duplicate sentences are also removed.

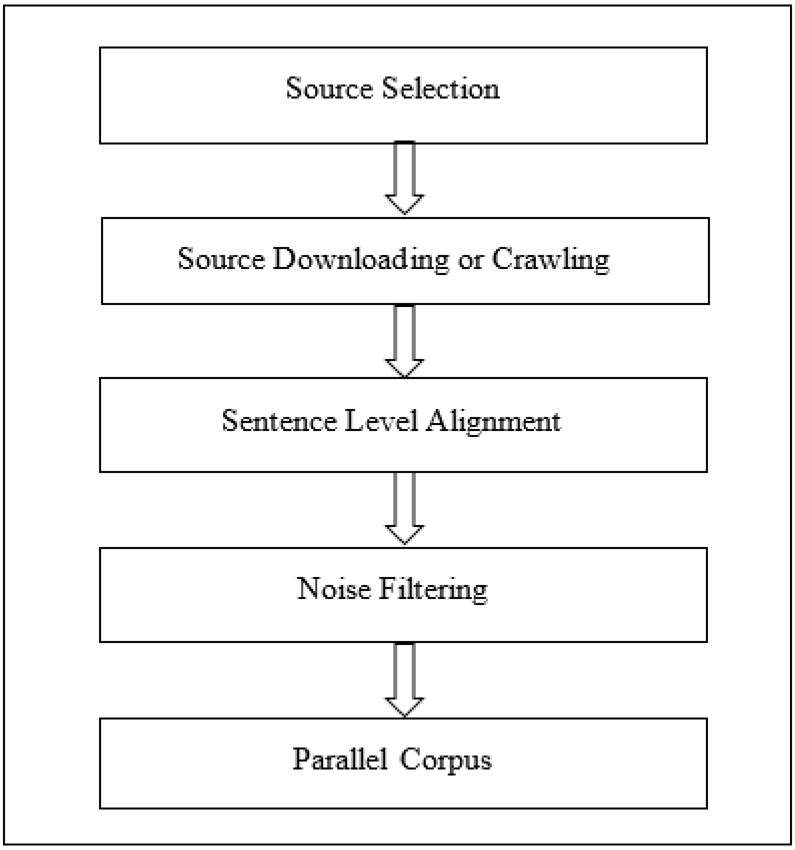


Figure 4.2 Creation of UCSY-corpus

Another problem is that this dataset contains spelling errors. Therefore, the spelling errors are modified. The common spelling errors are shown in Table 4.3 and modified according to [122].

Table 4.3 Common Spelling Error

No.	Common Spelling Error	Common Spelling Word
1.	ပဲ နှင့် ဘဲ	ကန့်သတ်ခြင်းပြ ==> ပဲ ပဲ ==> မင်းပဲ၊ ငါပဲ၊ ယခုပဲ၊ ဒါပဲ။
		ငြင်းဆိုခြင်းပြ ==> ဘဲ (မ+ကြိယာ+ဘဲ) ဘဲ ==> မလုပ်ဘဲ၊ မသိဘဲ၊ မရှိဘဲ။

2.	ပစ် နှင့် ပြစ်	ပစ် ==> စွန့်ပစ်၊ လွှင့်ပစ်၊ သေနတ်ပစ်။
		ပြစ် ==> ပြေပြစ်၊ ပြစ်မှား၊ အပြစ်ဒဏ်။
3.	ပ နှင့် ပု	ပ ==> ပခုံး၊ ပလွေ၊ ပလိပ်ရောဂါ၊ ပဝါ၊ သံပရာ။
		ပု ==> ပုလင်း၊ ပုရပိုက်၊ ပုထိုး၊ ပုဆိုး၊ ပုခတ်၊ ပုဆိန်၊ ပုတီး။
4.	ငုတ် နှင့် ငုပ်	ငုတ်(အပေါ်သို့ထွက်) ==> လက်ငုတ်လက်ရင်း၊ ငုတ်တုတ်ထိုင်၊ သစ်ငုတ်၊ ငုတ်ရိုက်။
		ငုပ်(အထဲသို့ဝင်) ==> ရေငုပ်၊ ချွေးငုပ်၊ ငုပ်လျှိုး
5.	ပျံ့ နှင့် ပြန့်	ပျံ့ ==> ပျံ့နှံ့၊ သင်းပျံ့၊ ပျံ့လွင့်၊ ပျံ့ပျူး၊ ပျံ့နှံ့ခြင်း။
		ပြန့်==> ပြန့်ကျဲ၊ ပြန့်ပွား၊ သတင်းပြန့်၊ ပြန့်ပြူး
6.	ယာ နှင့် ရာ	ယာ ==> ကွမ်းယာ၊ လယ်ယာ ၊ ကားဘီးတာယာ။
		ရာ ==> အိမ်ရာ၊ အမှုအရာ၊ ကြယ်တာရာ၊ အိပ်ရာ။
7.	မဲ့ နှင့် မယ့်	မဲ့ ==> ကင်းမဲ့၊ ဒါပေမဲ့၊ လက်နက်မဲ့၊ အပြစ်မဲ့၊ ပြောချင်ပေမဲ့။
		မယ့်(အနာဂတ်) ==> တောက်မယ့်မီးခဲတရဲရဲ ၊ လာမယ့်နှစ်၊ မင်းပြန်လာမယ့်နေ့။
8.	ဌ နှင့် ဋ္ဌ	ဌ ==> ဌာန၊ ဌာန္တရ၊ ဌာပနာ၊ ဌာန်၊ သူဌေး။
		ဋ္ဌ ==> ဥက္ကဋ္ဌ၊ ဆဋ္ဌမ၊ ပြဋ္ဌာန်း၊ ကမ္မဋ္ဌာန်း၊ အဋ္ဌကထာကျမ်း၊ အနိဋ္ဌာရုံ၊ အဓိဋ္ဌာန်၊ သန္နိဋ္ဌာန်။
9.	ခတ် နှင့် ခပ်	ခတ် ==> ခြင်းလုံးခတ်၊ ပစ်ခတ်၊ ပစ်စလက်ခတ်၊ ဘောင်ဘင်ခတ်၊ ယောက်ယက်ခတ်၊ လျှော်ခတ်၊ ပျာယာခတ်၊ လက်ထိပ်ခတ်၊ တိုက်ခတ်၊

		ကြက်ခြေခတ်၊ လှုပ်ခတ်၊ ယပ်ခတ်၊ ဒဏ်ခတ်၊ လက်ခမောင်းခတ်၊ မျက်စဉ်းခတ်၊ အကဲခတ်၊ ဆားခတ်၊ ကျွဲခတ်၊
		ခပ် ==> ကူးခပ်၊ တစ်ဖက်ကမ်းခပ်၊ အရာခပ်သိမ်း၊ ဇွန်းဖြင့်ခပ်၊ ရေခပ်။
10.	ငြာ ၊ ညာ ၊ ငြား ၊ ညား	ငြာ ==> ကြော်ငြာ၊ ငြာသံပေး။
		ညာ ==> ကြေညာ၊ အကြေအညာ၊ ညာလက်ရုံး၊ ညာသန်၊ လိမ်ညာ၊ ညီညာ။
		ငြား ==> သော်ငြားလည်း။
		ညား ==> ခံ့ညား၊ ခန့်ညား၊ လင်မယားညား။
11.	ချင်း နှင့် ခြင်း	ချင်း ==> ကိုယ်ချင်းစာ၊ ချက်ချင်း၊ နှစ်ချင်းပေါက်၊ ကိုယ့်မင်းကိုယ်ချင်း၊ လူသားချင်းစာနာမှု၊ ဆွေမျိုးသားချင်း၊ တစ်မုဟုတ်ချင်း၊ မကုန်မချင်း၊ အရည်အချင်း၊ မျက်နှာချင်းဆိုင်၊ တခဏချင်း၊ နေ့ချင်းညချင်း၊ တစ်ယောက်ချင်း၊ ပွဲချင်းပြီး၊ ဘေးချင်းတိုက်၊ ခြေချင်းဝတ်၊ အချင်းချင်း၊ မသေမချင်း၊ သွေးချင်း၊ မွေးချင်း၊ မြင်မြင်ချင်း၊ အိမ်နီးချင်း။
		ခြင်း ==> သေခြင်းဆိုး၊ အခြင်းအရာ၊ ခြင်းခြင်းနီ၊ အကြောင်းချင်းရာ၊ စည်းလုံးခြင်း၊ ခြင်းတောင်း၊ နှစ်ခြင်းသာသနာ၊ ပျောက်ခြင်းမလှပျောက်။
12.	ကျ နှင့် ကြ	ကျ ==> ကုန်ဈေးနှုန်းကျ၊ အမြင့်မှကျ၊ ကျခံ၊ ကျဆုံး၊ စာမေးပွဲကျ၊ သင်္ကြန်ကျ၊ စီးပွားရေးကျ၊

		ကျသင့်ငွေ။
		ကြ ==> သွားကြ၊ လာကြ။
13.	စီး၊ စည်း ၊ ဆီး ၊ ဆည်း	စီး ==> ခေါင်းစီး၊ စီးချင်းထိုး ၊ နိုင့်ထက်စီးနင်း၊ စီးဖြန်း၊ ပျက်စီး၊ ရေစီး။
		စည်း ==> ကလာပ်စည်း၊ ကြက်တောင်စည်း၊ ခေါင်းစည်းကြိုး၊ ထင်းစည်း၊ စည်းကြပ်ဒိုင်၊ သွေးစည်းညီညွတ်၊ ဆုံစည်း ။
		ဆီး ==> တားဆီး၊ ဆီးကြို၊ ဖမ်းဆီး၊ ဖျက်ဆီး၊ အတားအဆီး၊ ထုပ်ဆီးတိုး။
		ဆည်း ==> ဆည်းကပ်၊ ဆည်းပူး၊ ဖြည့်ဆည်း၊ သိမ်းဆည်း၊ ဆည်းလည်း။
14.	ယက် နှင့် ရက်	ယက် ==> ဂယက်ရိုက်၊ ဝဲဂယက်၊ လုယက်၊ လက်ယက်တွင်း။
		ရက် ==> ကွန်ရက်၊ ရက်ကန်းရက်၊ ထရံရက်၊ လက်ရက်ထည်။
15.	လှန်း နှင့် လှမ်း	လှန်း ==> နေလှန်း၊ အဝတ်လှန်း။
		လှမ်း ==> ခြေလှမ်း၊ လျှောက်လှမ်း။
16.	အ နှင့် အာ	အ ==> အရုဏ်ဦး၊ အရုဏ်တတ်၊ အရုဏ်ဆွမ်း။
		အာ ==> အာရုံ၊ အာရုံပြု၊ အာရုံကြော။
17.	ထုတ် နှင့် ထုပ်	ထုတ် ==> ထုတ်နုတ်၊ ထုတ်ပြန်၊ ထုတ်ပယ်၊ ထုတ်ဖော်၊ ထုတ်လုပ်၊ ခွဲထုတ်၊ ပွဲထုတ်၊ ထုတ်ဝေ၊ အားထုတ်၊ ထုတ်ချင်းပေါက်။
		ထုပ် ==> ဂေါ်ဖီထုပ်၊ ထုပ်ဆီးတိုး၊ ဓာတ်ထုပ်၊ ကောက်ညှင်းထုပ်၊ ထုပ်ပိုး၊ ဦးထုပ်။

18.	လည် နှင့် လယ်	လည် ==> နှစ်ပတ်လည်၊ လိပ်ပတ်လည်၊ ပတ်ပတ်လည်၊ မျက်စိလည်၊ တစ်ပတ်လည်၊ ကျင်လည်။
		လယ် ==> နေ့လယ်၊ အလယ်ခေါင်၊ နေ့လယ်စာ။
19.	မျှား ၊ မြား ၊ မြှား	မျှား ==> ငါးမျှား၊ မျှားခေါ်သည်။
		မြား ==> လေးမြား၊ မြားမြောင်၊ မြားတံ၊ အမြောက်အမြား။
		မြှား ==> ဒိုင်းမြှား။
20.	ကျေ နှင့် ကြေ	ကျေ ==> ကျေနပ်၊ ကမ္ဘာမကျေ၊ အောက်သက်ကျေ၊ ဝတ်ကျေတမ်းကျေ၊ လောကွတ်ကျေ၊ အခဲမကျေ၊ ဆယ်လီကျေ၊ ကျေအေး။
		ကြေ ==> ကြေညာ၊ စိစိညက်ညက်ကြေ၊ မွမွကြေ၊ ကြေညက်၊ ကြေမွ၊ ကြေကွဲ၊ တွန့်ကြေ၊ အစာမကြေ။
21.	နုတ် နှင့် နှုတ်	နုတ် ==> ဝမ်းနုတ်၊ အမွေးနုတ်၊ ကောက်နုတ်၊ ထုတ်နုတ်၊ အပေါင်းအနုတ်၊ ရာထူးမှနုတ်ထွက်၊ ပျိုးနုတ်။
		နှုတ် ==> နှုတ်ခမ်းနီ၊ နှုတ်ကျိုး၊ နှုတ်တိုက်ရွတ်၊ နှုတ်ဆက်၊ နှုတ်လုံ။
22.	လဲ နှင့် လည်း	လဲ(မေးခွန်း) ==> ဘာစာအုပ်ဖတ်နေလဲ။ ဘယ်တော့လဲဆိုတာ။ လည်း(မေးခွန်းမဟုတ်လျှင်)==> ဒါပေမဲ့လည်း၊ ကျွန်တော်လည်း၊ သို့သော်လည်း၊ မပြောလည်း။

23.	ဉနှစ်လုံးဆင့် အသံထွက်ပုံများ ပါဠိစကားလုံးများတွင် ညသည်ဉနှစ်လုံး ထပ်ထားသောဗျည်းတွဲ ဖြစ်ပြီးဉတစ်လုံးသည်ရှေ့က ဗျည်း၏အသတ်ဖြစ်သည်။ ရှေ့ဗျည်း၏ စာလုံးပေါင်း ကို မူတည်၍ ဉသတ်သံ ထွက်ရသည်။	ပညာ = ပင်ညာ(ပါဠိသံ) ၊ ပျင်ညာ(မြန်မာသံ) သာမည = သာမင်ည ပုည= ပုန်ည ပဋိညာဉ် = ပဒိန်ညင် ဝိညာဉ် = ဝိန်ညင် သညာ = သင်ညာ အညမည = အင်ညမင်ည အညတြ = အင်ညတရ အဘိညာဉ် = အဘိန်ညင်

Finally, more than 200K sentences (204,539) are collected for the UCSY-corpus after filtering, and shown in Table 4.4. The corpus preparation is a very important task to train the NLP tasks. All Myanmar-English parallel sentences have been converted into UTF-8 text file format and Myanmar3 font. After further cleaning the UCSY-corpus, it is ready to use in the research of the Natural Language Processing tasks.

4.3 Building Myanmar Monolingual Corpus

Myanmar monolingual corpus is constructed to improve the performance of machine translation systems. The collected sentences of Myanmar monolingual corpus is local news data and 170K Myanmar sentences. These sentences are collected from the government official websites. This corpus is also removed the noise and extra spaces. Spelling errors in this corpus are also modified. All Myanmar sentences have been converted into UTF-8 text file format and Myanmar3 font. This corpus is also ready to use in the research of the Natural Language Processing tasks. Table 4.5 shows the monolingual data about the corpus.

Table 4.4 Myanmar-English Parallel Corpus

Myanmar Sentences	English Sentences
မြန်မာကင်းထောက်ရာပြည့်အခမ်းအနားကိုအလင်္ကာကျော်စွာဟင်္သာတမြင့်ငွေဂုဏ်ပြုစပ်ဆိုသည့်ကင်းထောက်ဂုဏ်ရည်သီချင်းဖြင့်ဖွင့်လှစ်သည်။	The ceremony was opened by singing the song titled, virtue of scout — Kin Htauk Gonyi composed by Hinthada Myint Ngwe.
ကင်းထောက်လူငယ်များ၏ လှုပ်ရှားမှုဆိုသည်မှာလူငယ်များစိတ်ဓါတ်ဖွံ့ဖြိုးရေးအတွက်ပြုစုပျိုးထောင်ရေးနည်းတစ်မျိုးပင်ဖြစ်ပါသည်။	Activity of scout youths is a kind of nurturing moral development among youths.
ကောင်းသည့်ကိစ္စများကိုအားလုံးအတူတူဝိုင်းဝန်းဆောင်ရွက်ခြင်းသည်တစ်ဘဝအတွက်ခိုင်မြဲသည့်မိတ်ဆွေများရဲဘော်ရဲဘက်များရှာဖွေခြင်းပင်ဖြစ်ပါသည်။	Collective contribution in a good cause is a kind of finding life-long comrades or alter egos.
ကောင်းသည့်ကိစ္စများကိုအတူတူလုပ်ဆောင်ခြင်းထက်ပို၍ခိုင်မြဲနိုင်သည့်အဆက်အသွယ်မရှိနိုင်ပါ။	Nothing can supersede collective contribution in a good cause.
ထို့ကြောင့်ကင်းထောက်လူငယ်များယခုကဲ့သို့စုစည်းနိုင်သည့်အခြေအနေပြန်လည်ရောက်ရှိလာသည်ကိုဝမ်းမြောက်မိပါသည်။	Therefore, I am very glad to see our scout youths assemble like this.
ကင်းထောက်တစ်ယောက်အနေဖြင့်ငယ်ငယ်ကတည်းကသင်ပေးလိုက်သည့်စံနှုန်းစံထားများသည်တစ်ဘဝစွဲလန်းသွားနိုင်ပါသည်။	Standardized attitudes and mindsets indoctrinated since young age can impress anyone's mind for ever.
၎င်းတို့သည်မိမိကိုယ်တိုင်အတွက်သာမကမိမိတို့အသိုင်းအဝန်းအတွက်နိုင်ငံအတွက်ကမ္ဘာအတွက်များစွာအကျိုးပြုနိုင်မည်ဖြစ်ပါသည်။	These will surely benefit not only oneself but also one's society, one's country and the whole world.
မြန်မာနိုင်ငံကင်းထောက်အဖွဲ့ကို၁၉၁၆ခုနှစ်တွင်စတင်ဖွဲ့စည်းခဲ့သည်။	The Myanmar scouts movement was formed in 1916.
၂၀၁၂ခုနှစ်တွင်မြန်မာနိုင်ငံကင်းထောက်အဖွဲ့ကိုပြန်လည်ဖွဲ့စည်းခဲ့သည်။	The Myanmar scouts was reformed in 2012.

Table 4.5 Statistics of Monolingual Corpus

Domain	**Number of Sentence**	**Number of Myanmar word tokens**	**Number of Myanmar Syllable tokens**
Local News	179,443	4,378,522	7,924,296

CHAPTER 5

Attention-based Neural Machine Translation System

Neural Machine Translation (NMT) is a new technology to remarkably improve the machine translation. Previously, sequence-to-sequence models are developed and nowadays improved upon into attention-based models. Therefore, neural machine translation has now become a popular method to success for machine translation, as well as serves for other related NLP tasks.

Neural machine translation is designed as the entire machine translation process and it requires a parallel corpus like statistical machine translation (SMT). Moreover, there is a few preprocessing steps before a translation model can be built. And the neural machine translation is end-to-end architecture so it can remove the problem of sub components in SMT systems. This chapter presents the encoder and decoder approach. "An alignment model" is added to it, that is, the advanced methods of the encoder-decoder approaches.

5.1 Encoder-Decoder Approach

Encoder-decoder models consist of two parts: an encoder and a decoder. The encoder reads the whole input sequence. And then it encodes into the representation of the input sequence. The decoder uses the encoded representation as the decoder inputs to produce the output sequence.

The encoder-decoder models based on recurrent neural networks (RNNs) are sequence to sequence models. An RNN encoder-decoder consists of two parts. An RNN encoder takes input to a sequence. An RNN decoder outputs to another sequence. For example, a sentence in Myanmar can be considered as a sequence which can be input, and English translation of the sentence is generated as an output which again is a sequence of words.

From the perspective view of machine translation, the task of translation learns the conditional distribution p(f | e) of a target sentence f given a source sentence e. Generating all previous words, a model predicts the next word. When reaching the end of the sentence, the translation of the sentence is proceeded to predict, one word at a time. Figure 5.1 shows Myanmar to English sequence-to-sequence encoder-decoder model. Once processing having predicted the end of the input sentence marker </s>, the hidden state encodes its meaning. In other words, the vector holding the values of the nodes of this final hidden layer is the input sentence embedding. This is the encoder phase of the model. Then this hidden state is used to produce the translation in the decoder phase. During the decoder phase, not only does it need to have enough information to predict each next word, there also needs to be some accounting for what part of the input sentence has been already translated, and what still needs to be covered [43].

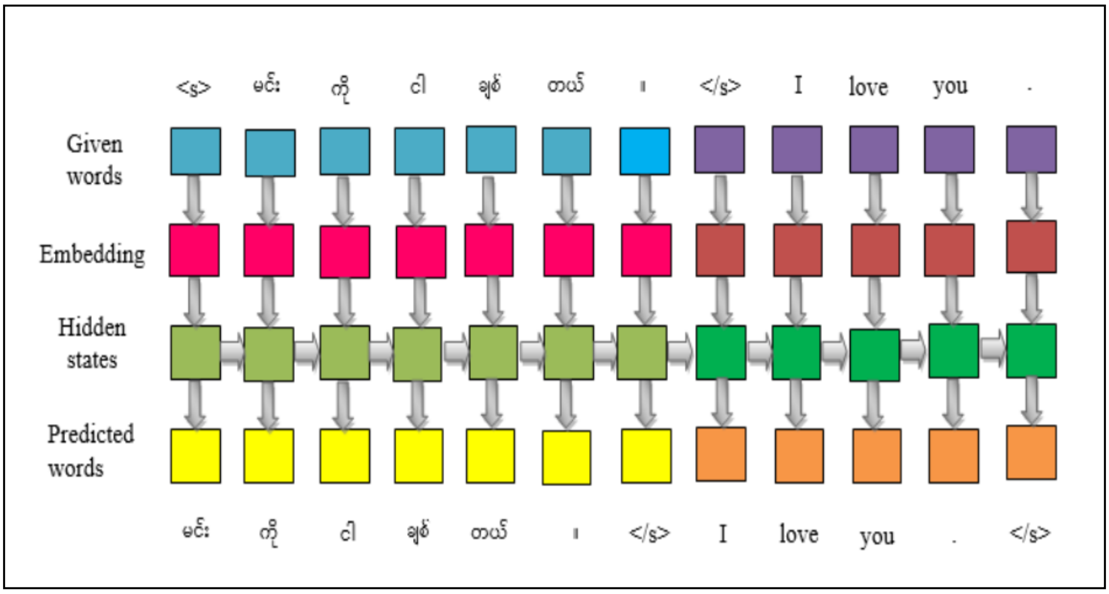


Figure 5.1 Myanmar to English Sequence-to-sequence Encoder-decoder Model

From the mathematical view of machine translation, an encoder reads the input sentence, a sequence of vectors $x = \left(x_1, \dots \dots, x_{T_x}\right)$, into a vector c:

$$h_t = f(x_t, h_{t-1}) \qquad \text{Equation (5.1)}$$

and

$$c = q\left(\{h_1, \dots, h_{T_x}\}\right), \qquad \text{Equation (5.2)}$$

where $h_t \epsilon\ R^n$ is a hidden state at time t, and c is a vector generated from the sequence of the hidden states. f and q are some nonlinear functions. The decoder is often trained to predict the next word y_t, given the context vector c and all the previously predicted words $\{y_1, \dots \dots, y_{t'-1}\}$. In other words, the decoder defines a probability over the translation y by decomposing the joint probability into the ordered conditionals:

$$p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, \dots \dots, y_{t-1}\}, c), \qquad \text{Equation (5.3)}$$

where $y = \left(y_1, \dots \dots \dots, y_{T_y}\right)$. With an RNN, each conditional probability is modeled as

$$P(y_t | \{y_1, \dots \dots \dots, y_{t-1}\}, c) \ = g(y_{t-1}, s_t, c), \qquad \text{Equation (5.4)}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_t, and s_t is the hidden state of the RNN.

In practice, the encoder-decoder models well translate only for short sentences up to, say, 10–15 words and fails for long sentences. Although there is a different way to enable the translation the long sentences, another idea is to reverse the order of the output sentences. In this way, the last words of the input sentences are close to the last words of the output sentence.

5.2 Adding an Alignment Model

Alignment is the one problem in machine translation problems. Alignment is identifying the relationships information between the input words and output words whereas translation is the process to use these information to select the appropriate output. In the neural machine translation field, this alignment is called "attention" and encoder-decoder model with attention is widely used in these days. Previous approach in section 5.1 contains encoder and decoder. Now these versions contain not only encoder and decoder but also attention mechanism[43].

5.2.1 Encoder

In this model, the encoder takes an input sequence x like $[x_1, \dots\dots, x_{T_x}]$ and summarize both of the preceding words and the following words. In this model, bidirectional recurrent neural network (BiRNN) is also used especially. A BiRNN consists of forward recurrent neural network and backward recurrent neural network. The forward recurrent neural network $\overrightarrow{f}$ takes the input sequence of words as in direction from left to right. And then it calculates a sequence as forward hidden states $(\overrightarrow{h_1}, \dots \dots, \overrightarrow{h_{T_x}})$. The backward recurrent neural network $\overleftarrow{f}$ also takes the sequence in the reverse order that is direction from right to left. More precisely, it means from the end of the sentence to the beginning. And it results in a sequence of backward hidden states $(\overleftarrow{h_1}, \dots \dots, \overleftarrow{h_{T_x}})$. An annotation for each word x_j is obtained by joining the forward hidden state and the backward hidden state, i.e., $h_j = \left[\overrightarrow{h_j^T}; \overleftarrow{h_j^T}\right]^{\mathrm{T}}$.

In this way, the annotation h_j contains the summaries of both the preceding words and the following words. Due to the tendency of RNNs to better represent recent inputs, the annotation h_j will focus on the words around x_j . This sequence of annotations is used by the decoder and the alignment model to compute the context vector. In the equation mentioned above, a generic function f for a cell is used in the recurrent neural network. This function may be a typical feed-forward neural network layer — such as f(x) = tanh(Ax + b)—or the more complex gated recurrent units (GRUs) or long short term memory cells (LSTMs). The original paper proposing this approach used GRUs, but lately LSTMs have become more popular[43].

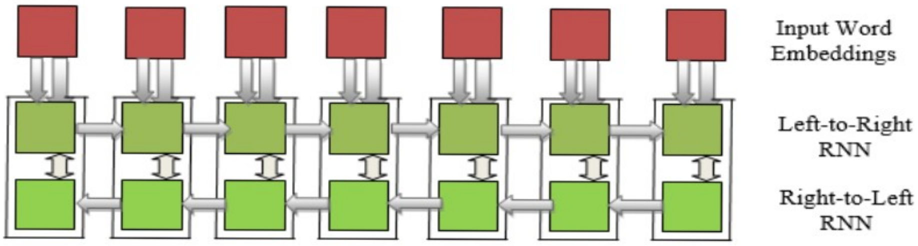


Figure 5.2 Bidirectional Recurrent Neural Network(BiRNN) Encoder and Two Hidden States

5.2.2 Decoder

The decoder is also a recurrent neural network. It takes some representation of the input context, the previous hidden state and output word prediction as inputs. And then it generates a new hidden decoder state and a new output word prediction.

In a new model architecture, each conditional probability are defined as:

$$p\ (y_i \mid y_1, \dots \dots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i), \qquad \text{Equation (5.5)}$$

Mathematically, recurrent neural network maintains a sequence of hidden states s_i which are computed from the previous hidden state s_{i-1}, the embedding of the previous output word y_{i-1}, and the input context c_i

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \qquad \text{Equation(5.6)}$$

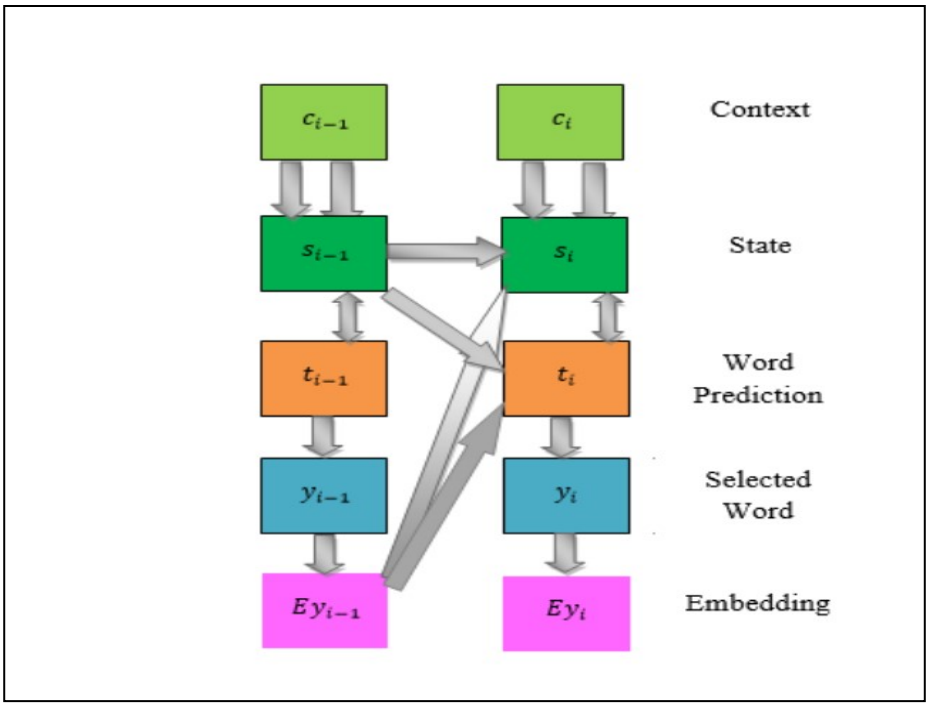


Figure 5.3 Output Decoder

Again, there are several choices for the function f that combines these inputs to generate the next hidden state: linear transforms with activation function, GRUs, LSTMs, etc. Typically, the choice here matches the encoder. So, if LSTMs are used for the encoder, then LSTMs for the decoder are also used. From the hidden state, the output word is now predicted. This prediction takes the form of a probability distribution over the entire output vocabulary. The prediction vector t_i is conditioned on the decoder hidden state s_{i-1} and, again, the embedding of the previous output word y_{i-1} and the input context c_i.

$$t_i = softmax\big(W(s_{i-1} + y_{i-1} + c_i)\big) \qquad \text{Equation (5.7)}$$

Note that we repeat the conditioning on y_{i-1} since we use the hidden state s_{i-1} and not s_1. This separates the encoder state progression from s_{i-1} to s_i from the prediction of the output word t_i.

The softmax is used to convert the raw vector into a probability distribution, where the sum of all values is 1. Typically, the highest value in the vector indicates the output word token y_i. Its word embedding y_{i-1} informs the next time step of the recurrent neural network. During training, the correct output word y_i is known, so training proceeds with that word. The training objective is to give as much probability mass as possible to the correct output word. The cost function that drives training is hence the negative log of the probability given to the correct word translation.

$$cost = -\log t_i \,|y_i| \qquad \text{Equation (5.8)}$$

Ideally, we want to give the correct word the probability 1, which would mean a negative log probability of 0, but typically it is a lower probability, hence a higher cost. Note that the cost function is tied to individual words, the overall sentence cost is the sum of all word costs.

During inference on a new test sentence, we typically chose the word y_i with the highest value in t_i use its embedding y_i for the next steps. But we will also explore beam search strategies where the next likely words are selected as y_i, creating a different conditioning context for the next words.

5.2.3 Attention Mechanism

Attention model was first presented by Dzmitry Bahdanau, et al., which proposed the alignment model as a natural extension of their previous work on the Encoder-Decoder model. Attention model addresses as a solution to the limitation of the Encoder-Decoder model encoding the input sequence to one fixed length vector from which to decode each output time step. This issue is believed to be more of a problem when decoding long sequences. Therefore, attention model is proposed as a method to both align and translate[43].

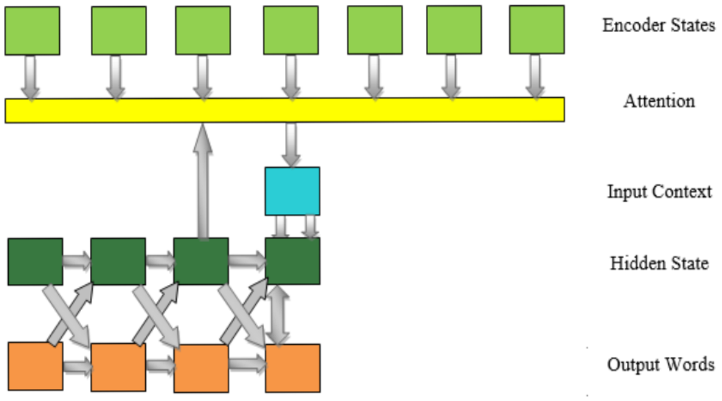


Figure 5.4 Attention Model

The decoder gave the output as a sequence of word representations $h_j = (\overleftarrow{h_j}, \overrightarrow{h_j})$ and the decoder expects a context c_i at each step i. Figure 5.4 gives at least an idea what the input and output relations are. The attention mechanism is informed by all input word representations $(\overleftarrow{h_j}, \overrightarrow{h_j})$ and the previous hidden state of the decoder s_{i-1} and it produces a context state c_i.

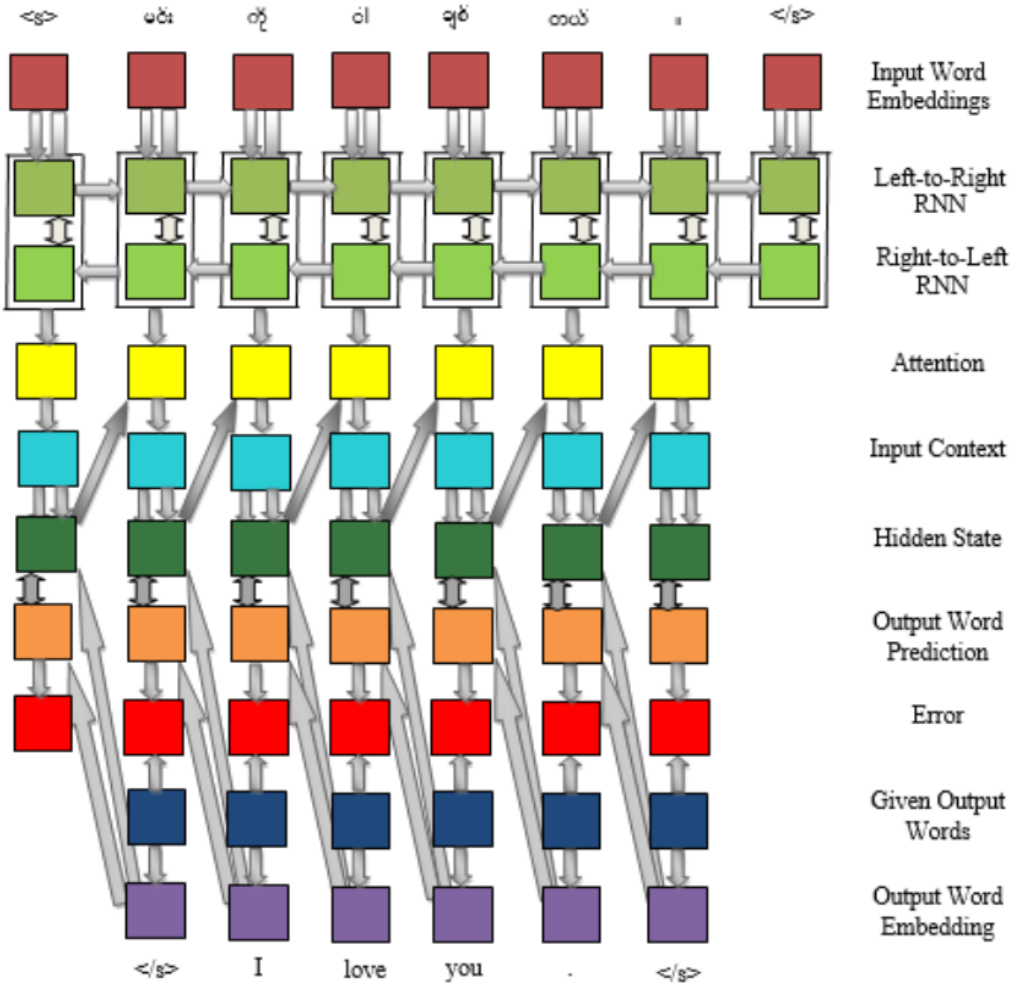


Figure 5.5 Fully Computation Graph for Training Example with 7 Input Tokens and 5 Output Tokens.

Mathematically, this association with a feedforward layer is computed firstly

$$a(s_{i-1}, h_j) = \omega^{aT} s_{i-1} + \mu^{aT} h_j + b^a \qquad \text{Equation (5.9)}$$

The output of this computation is a scalar value, indicating how important input word j is to produce output word i. And then, this attention value is normalized, so that the attention values across all input words j add up to one, using the softmax.

$$\alpha_{ij} = \frac{\exp\left(a(s_{i-1}, h_j)\right)}{\sum_k exp(a\ (s_{i-1}, h_k))} \qquad \text{Equation (5.10)}$$

Now the normalized attention are value use to weigh the contribution of the input word representation h_j to the context vector ci and are done.

$$c_i = \sum_j \alpha_{ij} h_j \qquad \text{Equation(5.11)}$$

Simply adding up word representation vectors may at first seem an odd and simplistic thing to do. But it is a very common practice in deep learning for natural language processing. Researchers have no qualms about using sentence embeddings that are simply the sum of word embeddings and other such schemes. The fully computation graph for a short sentence pair is shown in Figure 5.5. Practical training of neural machine translation models requires Graphics User Interfaces (GPUs) which are well suited to the high degree of parallelism inherent in these deep learning models.

CHAPTER 6
EXPERIMENTAL RESULTS

In this chapter, the dataset, preprocessing tools and Myanmar-English neural machine translation with attention models are described in both directions. In both direction of Myanmar-English neural machine translation models, word-based model are described as a baseline system. And Myanmar character-based model and Myanmar syllable-based model are presented. Moreover, the results and evaluation details of the systems are described. In order to improve the performance of Myanmar syllable-based models, the experiments are done by adding the monolingual data into existing training data.

6.1 Experimental Setting

The attention-based neural machine translation approach has been evaluated on the task of Myanmar-English language pair in both directions. This section describes the dataset, preprocessing tools and the neural machine translation models.

6.1.1 Dataset

Myanmar language is regarded as a low resource Asian language, so there are not many Myanmar-English parallel sentences. In the experiments, two parallel corpora are used: the ALT corpus and the UCSY corpus. ALT corpus consists of twenty thousand Myanmar-English parallel sentences from Wiki news articles. The UCSY corpus consists of 200K Myanmar-English parallel sentences collected from different domains, including Local news, Travel Domain, school text books and spoken text. There are 220K parallel sentences in total. The corpus is randomly divided into training data, development data and test data. Therefore, the data for Myanmar-English and English-Myanmar translation tasks is a mix domain data collected from different sources. Besides, this corpus is general corpus covering difference domains. Table 6.1 shows data statistics used for the experiments.

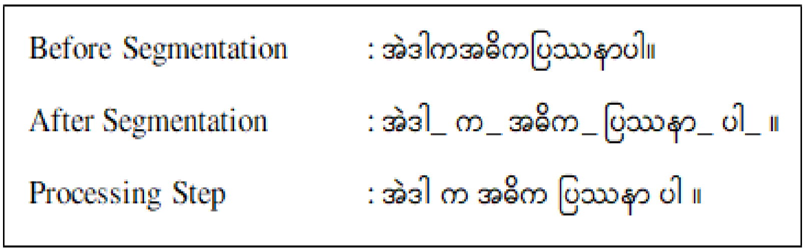
6.1.2 Preprocessing Tools

Myanmar Language is an unsegmented language and there is no clear definition of word boundaries. As it does not contain white space to delimit the words like English, the proper text segmentation is essential for Myanmar Language. It is also essential for every language as it is the fundamental step for linguistic processing. It may also be necessary to allow multiple correct segmentations of the same text, depending on the requirements of further natural language processing steps, such as Machine translation from Myanmar to other languages. In the experiments, three neural machine translation models are tested with attention: Word-based model, Myanmar character-based model, and Myanmar syllable-based model.

Table 6.1 Data Statistics of the Corpus

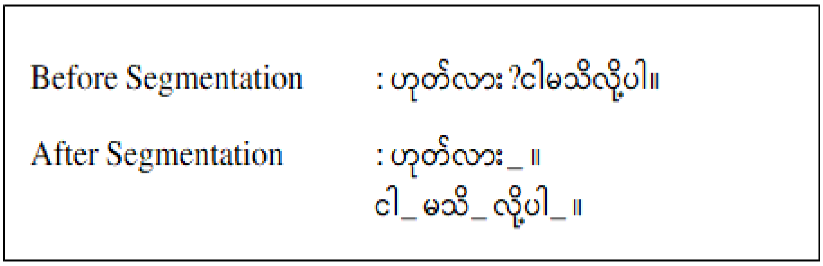
Domain	**No. of Training Sentences**	**No. of Development Sentences**	**No. of Test Sentences**
Local News	72,190	300	230
ALT domain	19,683	300	140
Travel domain	93,314	800	300
School Text Book	18,733	400	300
Spoken Text	17,271	400	300
Total	221,191	2200	1270

Word Segmentation is not a trivial task for Myanmar text, same as other Asian languages. It is necessary for high level language analysis including name entity recognition and syntactic parsing that are used in many natural language processing (NLP) applications such as machine translation system. For word-based neural machine translation model, UCSY_NLP lab segmenter [73] is used to segment the Myanmar sentence into word level. UCSY_NLP lab segmenter is implemented a combined model, bigram and word juncture. This segmenter works by longest matching and bigram method with a pre-segmented corpus of 50,000 words collected manually from Myanmar Text Books, Newspapers, and Journals. The corpus is in Unicode encoding. After segementing the Myanmar sentence by UCSY_NLP lab segmenter, the "_ " from the result is removed and replaced with space. Figure 6.1 shows the process of UCSY_NLP lab segmenter. It is not able to segment when "?" and "%" contains in Myanmar sentences. Examples are shown in Figure 6.2 and Figure 6.3. These sentences are segmented manually.



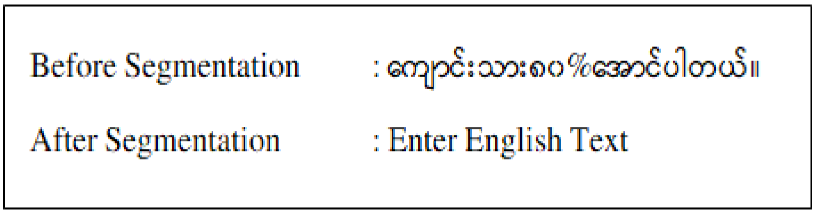
Before Segmentation	: အဲဒါကအဓိကပြဿနာပါ။
After Segmentation	: အဲဒါ_ က_ အဓိက_ ပြဿနာ_ ပါ_ ။
Processing Step	: အဲဒါ က အဓိက ပြဿနာ ပါ ။

Figure 6.1 The Process of Word Level Segmentation



Before Segmentation	: ဟုတ်လား ?ငါမသိလို့ပါ။
After Segmentation	: ဟုတ်လား_ ။ ငါ_ မသိ_ လို့ပါ_ ။

Figure 6.2 Sentence that are Manually Segmented

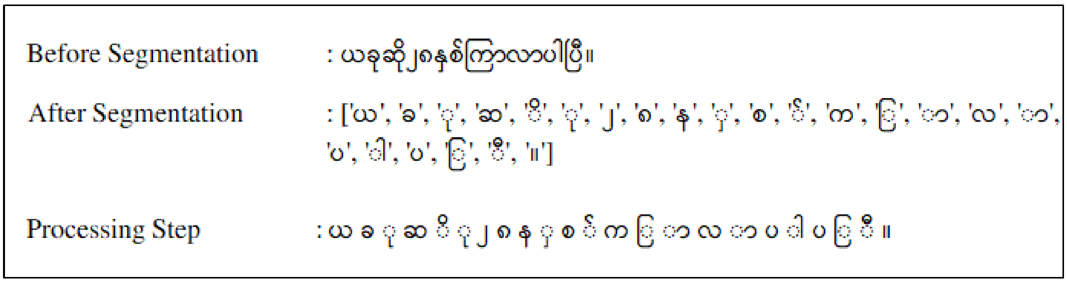


Before Segmentation	: ကျောင်းသား၈၀%အောင်ပါတယ်။
After Segmentation	: Enter English Text

Figure 6.3 Sentence that are Manually Segmented

For Myanmar character-based neural machine translation model, python programming code is used. After segementing the Myanmar sentence into character segmentation, the ', ' , [' and '] from the result is removed and replaced with space. Figure 6.4 shows the process of character segmentation for Myanmar character-based NMT model.

For Myanmar syllable-based neural machine translation model, "sylbreak" [92] is used to segment the Myanmar sentence into syllable level. Syllable segmentation is an important preprocess for many natural language processing (NLP) such as romanization, transliteration and grapheme-to-phoneme (g2p) conversion. "sylbreak" is a syllable segmentation tool for Myanmar language (Burmese) text encoded with Unicode (e.g. Myanmar3, Padauk). After segmenting the Myanmar sentence into syllable segmentation, the "|" from the result is removed and replaced with space and leading the trim process. Figure 6.5 shows the process of syllable segmentation for Myanmar syllable-based NMT model.

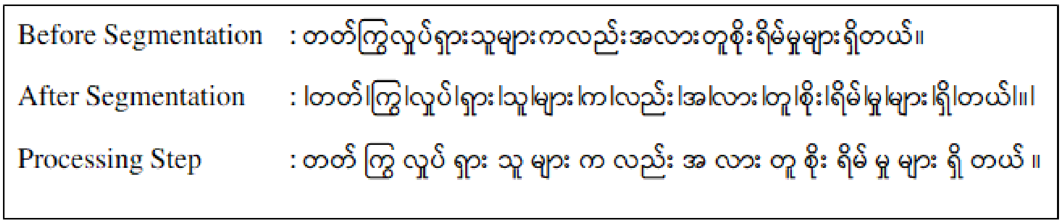


Before Segmentation : ယခုဆို၂၈နှစ်ကြာလာပါပြီ။

After Segmentation : ['ယ', 'ခ', 'ု', 'ဆ', 'ိ', 'ု', '၂', '၈', 'န', 'ှ', 'စ', '်', 'က', 'ြ', 'ာ', 'လ', 'ာ', 'ပ', 'ါ', 'ပ', 'ြ', 'ီ', '။']

Processing Step : ယ ခ ု ဆ ိ ု ၂ ၈ န ှ စ ် က ြ ာ လ ာ ပ ါ ပ ြ ီ ။

Figure 6.4 The Process of Character Level Segmentation



Before Segmentation : တတ်ကြွလှုပ်ရှားသူများကလည်းအလားတူစိုးရိမ်မှုများရှိတယ်။

After Segmentation : |တတ်|ကြွ|လှုပ်|ရှား|သူ|များ|က|လည်း|အ|လား|တူ|စိုး|ရိမ်|မှု|များ|ရှိ|တယ်|။|

Processing Step : တတ် ကြွ လှုပ် ရှား သူ များ က လည်း အ လား တူ စိုး ရိမ် မှု များ ရှိ တယ် ။

Figure 6.5 The Process of Syllable Level Segmentation

6.1.3 Models

Myanmar-English neural machine translation models are built with attention in both directions. A year later, in 2016, a neural machine translation system won in almost all language pairs. In 2017, almost all submissions were neural machine translation systems. At the time of writing, neural machine translation research is progressing at rapid pace. There are many directions that are and will be explored in the coming years, ranging from core machine learning improvements such as deeper models to more linguistically informed models. More insight into the strength and weaknesses of neural machine translation is being gathered and will inform future work.

There is an extensive proliferation of toolkits available for research, development, and deployment of neural machine translation systems. At the time of writing, the number of toolkits is multiplying, rather than consolidating. So, it is quite hard and premature to make specific recommendations. Nevertheless, some of the promising toolkits are: Theano, Torch/pyTorch, Sockeye and Tensorflow. Many other systems such as GroundHog, Blocks, tensorflowseq2seq, lamtram, and seq2seq-attn, exist mostly as research code. These libraries provide important functionality but minimal support to production users. NMT toolkits can provide a foundation to build upon. A toolkit should aim to provide a shared framework for developing and comparing open source systems, while at the same time being efficient and accurate enough to be used in production contexts. Currently there are several existing NMT implementations.

With these goals in mind, Pytorch OpenNMT [80] available in GitHub is applied, to build the Myanmar-English neural machine translation models. Three models with attention, namely, word-based NMT model(baseline system), Myanmar character-based NMT model and Myanmar syllable-based NMT model are trained.

6.1.3.1 Word-based Neural Machine Translation System(Baseline System)

As a baseline system, word-level neural machine translation model is built in both directions. For the translation system, the sequence-to-sequence model are trained with attention and a 2-layer long short-term memory with 500 hidden units each on both the encoder/decode are used. Drop-out was set to 0.3. Each direction is computed using 1.0 learning rate, 64 batches size. A vocabulary size of 26,133 words and 50,004 words for Myanmar to English NMT model respectively and 50,002 words and 26,135 words for English to Myanmar NMT model are used. This model is trained for at least one day with Tesla K80 GPU.

6.1.3.2 Myanmar Character-based Neural Machine Translation System

Myanmar Character-based Neural Machine Translation Systems is a re-implementation of Myanmar character and English word level with attention model. This model is trained with Pytorch openNMT [80] toolkit in both directions. All weights are initialized from a uniform distribution. Each model is trained on Tesla K80 GPU. The following settings are generally used: 2-layer long short-term memory, 500 hidden units, 1.0 learning rate, 64 batches size and 0.3 dropout. And 198 characters and 31,976 words for Myanmar to English NMT model are also used. For English to Myanmar NMT model, 31,974 words and 200 characters are used, respectively.

6.1.3.3 Myanmar Syllable-based Neural Machine Translation System

The further investigation is the usage of Myanmar syllable and English word level NMT model with attention in both directions. The vocabulary size 5,072 syllables and 50,004 words for Myanmar to English NMT model have also been used respectively and 50,002 words and 5,074 syllables are used for English to Myanmar NMT model. For the experiment, the following settings are also used: 2-layer long short-term memory, 500 hidden units, 1.0 learning rate, 64 batches size and 0.3 dropout.

6.2 Myanmar to English Neural Machine Translation System Results and Details

Two automatic criteria for the evaluation of the translation output are used. One is the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) and the other is the Rank-based Intuitive Bilingual Measure (RIBES). In Myanmar to English neural machine translation, it is clear that the evaluation results of Myanmar Syllable-based neural machine translation model are much better than those of the others. Table 6.2 shows the evaluation result of Myanmar to English neural machine translation models. It is observed that Myanmar syllable-based translation is surprisingly effective when translating into English, obtaining a BLEU score of 21.88 and RIBES score of 0.640361. Generally, it improves the performance of machine translation up to 5 BLEU over baseline systems.

Table 6.2 Evaluation Results of Myanmar to English Neural Machine Translation Models

Model	BLEU	RIBES
word-based NMT	16.94	0.581001
Myanmar Character-based NMT	15.18	0.543013
Myanmar Syllable-based NMT	**21.45**	**0.638706**

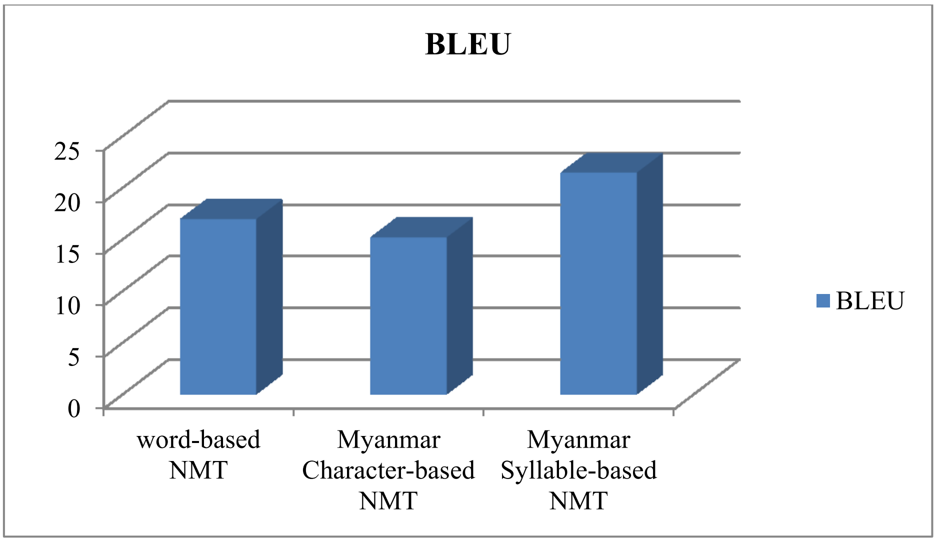


Figure 6.6 The Comparison of Myanmar to English Translation in BLEU

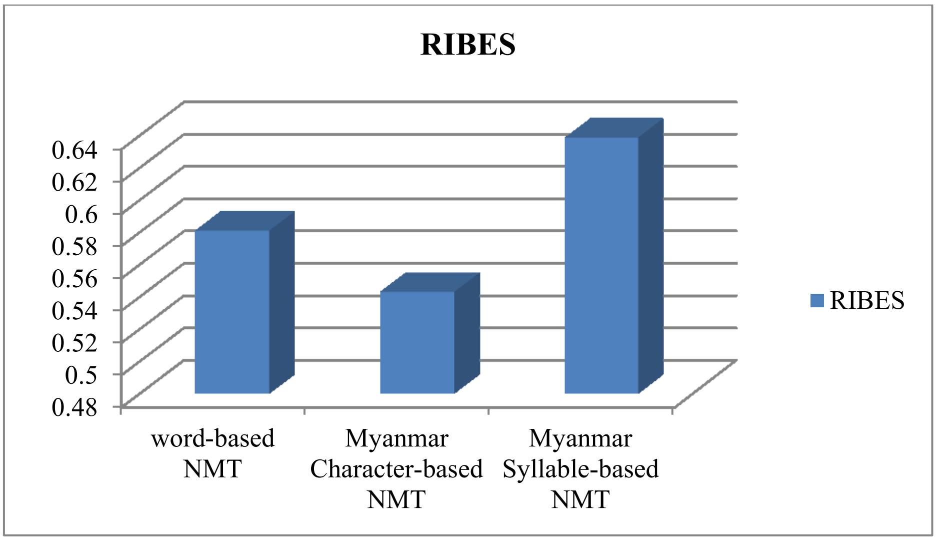


Figure 6.7 The Comparison of Myanmar to English Translation in RIBES

Compared with word-based NMT, Myanmar character-based NMT and Myanmar Syllable-based NMT, Syllable-based NMT generates a better translation. Though word-based NMT keeps most meaning of the source sentence, it fails to identify the subject of the whole sentence. Although it is expected Myanmar character-based NMT model to handle novel and unseen morphological inflections well, the translation output is not sufficient to address this problem. On the other hand, a character-level model has a much better chance recovering the original word or sentence. Indeed, the proposed model is robust against a few spelling mistakes and inaccurate translation. The Myanmar syllable-based model performs better on these long, concatenative words with ambiguous segmentation. Myanmar syllable-based model is capable of achieving compared to the other systems. Figure 6.6 and 6.7 show the comparisons of models in BLEU and RIBES in Myanmar to English translation, respectively.

As a human translator, Myanmar syllable-based model has an impact in learning a better translation which includes better alignment, reordering, morphological generation and disambiguation. Mainly, Myanmar syllable-based neural machine translation system reduces the source unknown words and handles the morphological variations. Examples of translation on Myanmar to English NMT models are shown in Table 6.3.

Table 6.3 Examples of Translations on Myanmar to English Neural Machine Translation Models

1	Source	နော်ဝေနိုင်ငံ၏ တယ်လီနောကုမ္ပဏီကလူ့အခွင့်အရေးကို လေးစားမည်ဖြစ်ကြောင်းကတိပြုထားသည်ဟုဆိုသည်။
	Word-based NMT	It is a violation of human rights.
	Myanmar Character-based NMT	The issue of the country is found to respect the human rights.
	Myanmar Syllable-based NMT	It is said that the Telenor of Norway will respect human rights.
	Reference	The Telenor of Norway says its commitment to respecting human rights.
2	Source	နိုဝင်ဘာ၈ရက်နေ့တွင်မြန်မာနိုင်ငံ၏ အထွေထွေရွေးကောက်ပွဲကိုကျင်းပမည်။
	Word-based NMT	It is to amend the 2008 constitution.
	Myanmar Character-based NMT	It is the issue of Myanmar infrastructure.
	Myanmar Syllable-based NMT	The Myanmar’s general elections will be held on November 8.

	Reference	Myanmar's general election on November 8 will be held.
3	Source	အမှန်တော့ကျွန်တော်တို့စိုးရိမ်ပါတယ်။
	Word-based NMT	In fact, we are very afraid of this.
	Myanmar Character-based NMT	But we are very worried.
	Myanmar Syllable-based NMT	Of course, we are worried.
	Reference	Of course, we're worried,".
4	Source	အစိုးရဟာKIOကိုတရားမဝင်အဖွဲ့အစည်းအဖြစ်သတ်မှတ်ထားပါတယ်။
	Word-based NMT	But the government has failed to support the Kachin independence organization (KIO).
	Myanmar Character-based NMT	The government is constituted as the state.
	Myanmar Syllable-based NMT	The government considers the KIO an unlawful association.
	Reference	The government considers the KIO an unlawful association.
5	Source	မြို့ပြစစ်ပွဲတွေစတင်ခဲ့ပြီးနောက်ကလေးတွေရဲ့ပညာရေးကအလွန်ခက်ခဲမှုတွေဖြစ်လာခဲ့တယ်လို့သူမကပြောခဲ့သည်။
	Word-based NMT	She said the children had become very difficult and <unk>
	Myanmar Character-based NMT	It started the picturesque of the urban war.
	Myanmar Syllable-based NMT	After the civil war began, children's education became very difficult, she said.
	Reference	After the civil war began, children's education became very difficult, she said.
6	Source	စာတိုက်ရုံးကိုဘယ်မှာတွေ့ရမလဲပြောပြပါလား။

	Word-based NMT	Would you tell me where to find the post office?
	Myanmar Character-based NMT	Could you tell me where to get to the post office?
	Myanmar Syllable-based NMT	Could you tell me where to find the post office?
	Reference	Could you tell me how do I get to the post office?

In the examples 3,4,5,6, the translation results of Myanmar Syllable-based NMT model is correct exactly. Although the translation results of example 1 and 2 are not same to the references exactly, their results are meaningful.

6.3 English to Myanmar Neural Machine Translation System Results and Details

In English to Myanmar neural machine translation, the evaluation results of Myanmar Syllable-based neural machine translation models are the best. Table 6.4 shows the evaluation result of English to Myanmar neural machine translation models. In this experiment, Myanmar syllable-based NMT model is surprisingly improved and obtained a BLEU score of 28.70 and RIBES score of 0.793571. Myanmar syllable-based NMT improves the performance up to 15 BLEU over baseline system.

Table 6.4 Evaluation Results of English to Myanmar Neural Machine Translation Models

Model	BLEU	RIBES
word-based NMT	13.38	0.692246
Myanmar Character-based NMT	26.18	0.783013
Myanmar Syllable-based NMT	**28.70**	**0.793571**

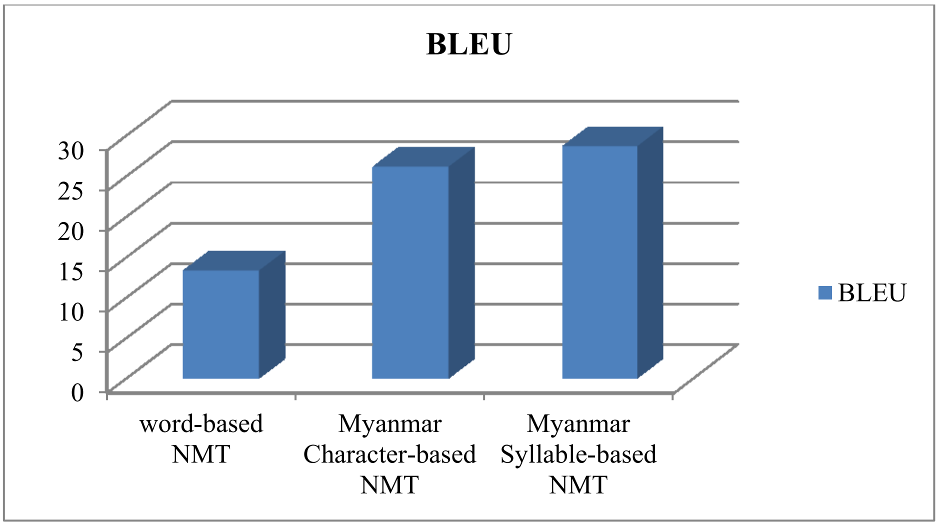


Figure 6.8 The Comparison of English to Myanmar Translation in BLEU

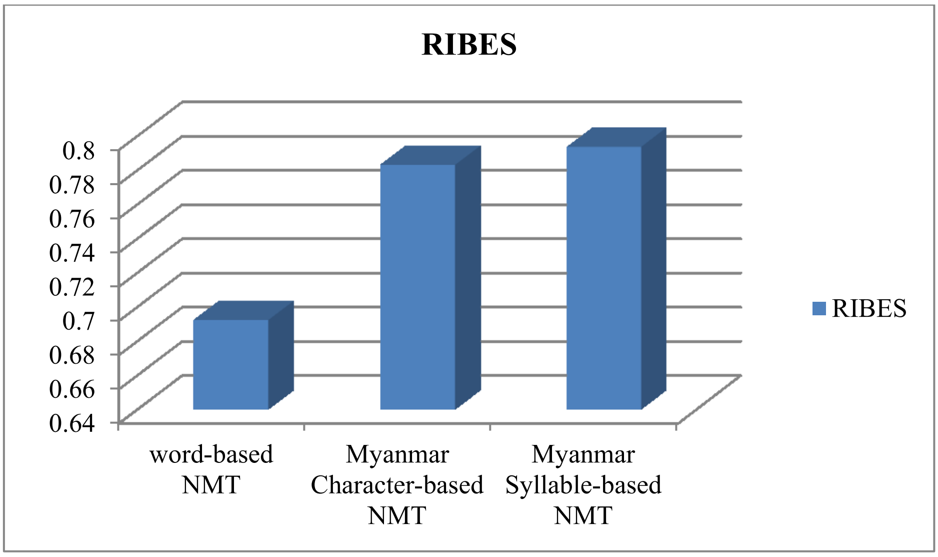


Figure 6.9 The Comparison of English to Myanmar Translation in RIBES

Table 6.5 summarizes the translation results of English to Myanmar NMT model for all the three models. Myanmar Syllable-based NMT model is capable of achieving compared to baseline systems. Though the work-based NMT model and Myanmar character-based NMT model translate the same part of the sentence twice, Myanmar syllable-based NMT model translates generally. Figure 6.8 and 6.9 show the comparisons of three models in BLEU and RIBES in English to Myanmar translation respectively.

Table 6.5 Examples of Translation on English to Myanmar Neural Machine Translation Models

1	Source	What are the women in important decision-making roles?
	Word-based NMT	၎င်း သည် မြန်မာနိုင်ငံ တွင် အမျိုးသမီးများ ၏ ပါဝင် လုပ်ဆောင် မှု ကို အဟန့်အတား ဖြစ်စေ ပါသည် ။
	Myanmar Character-based NMT	အ ရ ေ း အ က ြ ီ း ဆ ု ံ း မ ိ န ် း မ တ ွ ေ ဖ ြ စ ် ခ ဲ ့ တ ဲ ့ အ မ ျ ိ ု း သ မ ီ း တ ွ ေ ပ ါ ။
	Myanmar Syllable-based NMT	အ ရေး ကြီး ဆုံး ဆုံး ဖြတ် ချက် ချ ရာ မှာ အ မျိုး သ မီး တွေ ကဏ္ဍ က ဘာ တွေ လဲ ။
	Reference	အရေးကြီးတဲ့ဆုံးဖြတ်ချက်တွေချရမယ့်နေရာတွေမှာလည်း အမျိုးသမီးတွေရဲ့ကဏ္ဍကဘာလဲ။
2	Source	But I have lost this chance due to the situation.
	Word-based NMT	ဒါပေမဲ့ ဒီ အခွင့်အရေး ကို လည်း မ ချိုးဖောက် ဘူး ။
	Myanmar Character-based NMT	အ ခ ု အ ခ ြ ေ အ န ေ က တ ေ ာ ့ ဘ ယ ် လ ိ ု မ ှ မ ပ ျ ေ ာ က ် ဆ ု ံ း ဖ ြ စ ် ဘ ူ း ။
	Myanmar Syllable-based NMT	ဒါ ပေ မဲ့ အ ခု အ ခြေ အ နေ ကြောင့် ကျွန် တော် ဒီ အ ခွင့် အ ရေး ကို ဆုံး ရှုံး သွား တယ် ။
	Reference	ဒါပေမဲ့အခုအခြေအနေကြောင့်ဒီအခွင့်အရေးကိုကျွန်တော်ဆုံး ရှုံးရပြီ။
3	Source	Sayargyi U Hla comes back from school.
	Word-based NMT	ဦးလှ သည် ကျောင်း မှ ပြန် လာသည် ။
	Myanmar Character-based	ဦ း လ ှ ၊ က ျ ေ ာ င ် း သ ည ် ပ ြ န ် လ ာ

	NMT	သ ည ် ။
	Myanmar Syllable-based NMT	ဆ ရာ ကြီး ဦး လှ သည့် ကျောင်း မှ ပြန် လာ သည်
	Reference	ဆရာကြီးဦးလှသည့်ကျောင်းမှပြန်လာပါသည်။
4	Source	I wanted to attend Myitkyina university.
	Word-based NMT	တက်ရောက် လာ တဲ့ KIO ကိုယ်စားလှယ် တွေ က ငြိမ်းချမ်းရေး ဆွေးနွေးပွဲ တွေ ကို တက်ရောက် တယ် ။
	Myanmar Character-based NMT	က ျ ွ န ် တ ေ ာ ် တ က ္ က သ ိ ု လ ် က ိ ု မ က ြ ာ ခ င ် တ က ် ခ ျ င ် လ ိ ု ့ ။
	Myanmar Syllable-based NMT	ကျွန် တော် မြစ် ကြီး နား တက္က သိုလ် ကို တက် ချင် တယ် ။
	Reference	ကျွန်တော်မြစ်ကြီးနားကတက္ကသိုလ်ကိုတတ်ရောက်ရန်ဆန္ဒရှိခဲ့တယ်။
5	Source	He lost both his legs.
	Word-based NMT	သူ့ ခြေထောက် နှစ် ဦးလုံး ပျောက် သွား ပါတယ် ။
	Myanmar Character-based NMT	သ ူ သ ည ် သ ူ ့ ခ ြ ေ န ှ စ ် ခ ျ ေ ာ င ် း န ှ စ ် ယ ေ ာ က ် ရ ှ ိ ခ ဲ ့ တ ယ ် ။
	Myanmar Syllable-based NMT	သူ သည် သူ ၏ ခြေ နှစ် ဖက် ဆုံး ရှုံး ခဲ့ ရ သည် ။
	Reference	သူသည်သူ့ခြေနှစ်ဖက်ဆုံးရှုံးခဲ့ရသည်။

6.4 Using Monolingual Data for Improvements of Myanmar Syllable-based NMT model

In this section, Myanmar syllable-based NMT model is described using monolingual data. Using monolingual data aims to improve the performance of Myanmar-English neural machine translation systems in both directions. Myanmar language is one of the low resource languages and there is no corpus that contains millions sentences. Using monolingual data helps to improve the performance of machine translation systems in low resource languages.

In this research, two ways are used in Myanmar syllable-based NMT model using monolingual data. The first way is the copying monolingual data from the existing training data. Firstly, Myanmar syllable-based NMT model are trained. Secondly, Myanmar sentences of the trained data copy and translate these copy sentences into English language using Myanmar syllable-based NMT model that are previous step. In addition, these translate sentences added to the existing training data. Finally, the Myanmar syllable-based NMT model using monolingual data is trained again. The other way is using other monolingual data. In this way, Myanmar syllable-based NMT model is trained. Then, other monolingual sentences are translated into English sentences using Myanmar syllable-based NMT model that are previous step. In addition, these translated sentences are added to the existing training data. Finally, the Myanmar syllable-based NMT model using monolingual data is trained again. Both ways do not changed the configuration setting.

Table 6.6 Evaluation Results of Myanmar to English Neural Machine Translation Models using Monolingual Data

Model	**BLEU**	**RIBES**
Myanmar Syllable-based NMT	**21.45**	**0.638706**
Myanmar syllable-based NMT model + copy monolingual	21.15	**0.652204**
Myanmar syllable-based NMT model + monolingual	**21.88**	**0.642924**

Table 6.6 and 6.7 are the Evaluation results of Myanmar-English NMT models in both directions using Monolingual data. In Myanmar to English NMT model using Monolingual data, Myanmar syllable-based NMT model with other monolingual data is a little improvement in both of BLEU score and RIBES score. In Myanmar syllable-based NMT model with copy monolingual data, BLEU score does not improve but RIBES score does a little improve. Figure 6.10 and 6.11 show the

comparisons of three models in BLEU and RIBES in Myanmar to English translation, respectively.

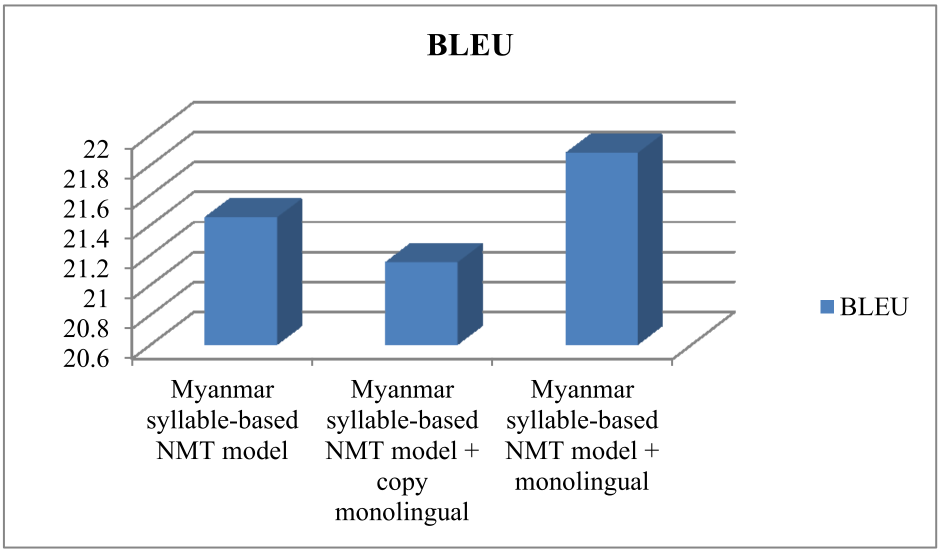


Figure 6.10 The Comparison of Myanmar to English Translation Using Monolingual Data in BLEU

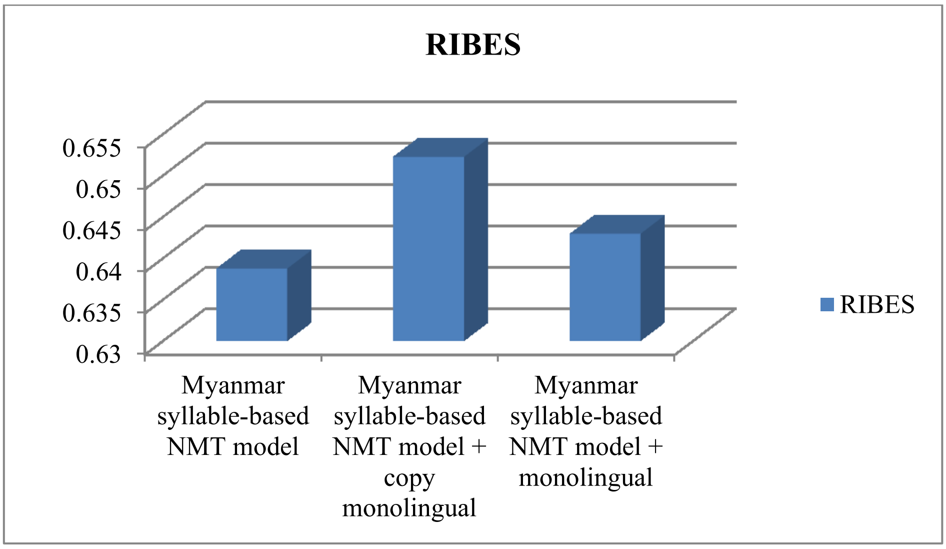


Figure 6.11 The Comparison of Myanmar to English Translation Using Monolingual Data in RIBES

In English to Myanmar NMT model with monolingual, Myanmar syllable-based NMT model with other monolingual data are significantly improved up to 5 BLEU score over the baseline system. In Myanmar syllable-based NMT model with copy monolingual data improves in both of BLEU and RIBES scores. Figure

6.12 and 6.13 show the comparisons of three models in BLEU and RIBES in English to Myanmar translation, respectively.

Table 6.7 Evaluation Results of English to Myanmar Neural Machine Translation Models using Monolingual Data

Model	**BLEU**	**RIBES**
Myanmar Syllable-based NMT	**28.70**	**0.793571**
Myanmar syllable-based NMT model + copy monolingual	29.39	**0.800856**
Myanmar syllable-based NMT model + monolingual	**33.17**	**0.758152**

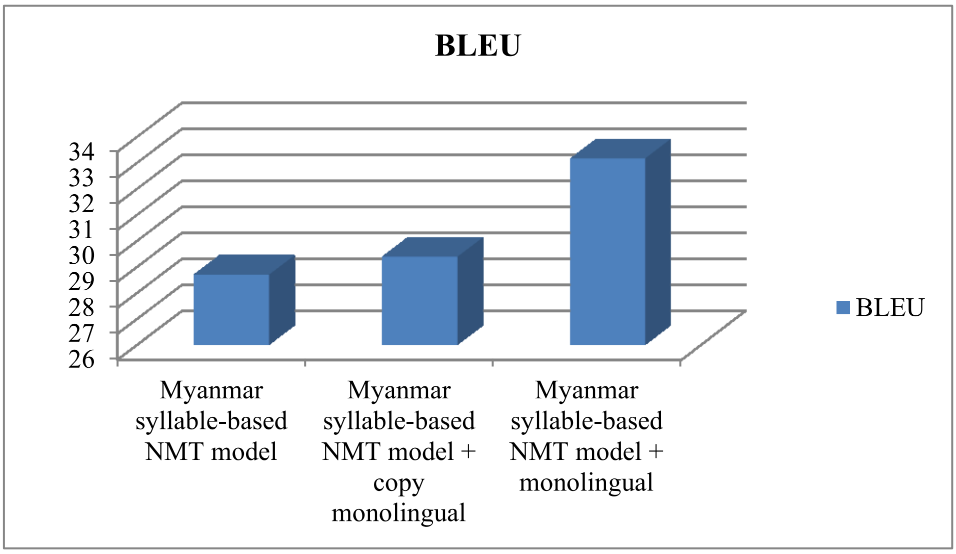


Figure 6.12 The Comparison of English to Myanmar Translation Using Monolingual Data in BLEU

6.5 Error Analysis

The most common problems of NMT are that it translates the duplicate words and unknown words. Previous approaches to unknown word problems are roughly categorized into three types: character-based, subword-based, and copy-based approaches. Although the unknown problem can be dealt with by using character-based and subword-based, there still remains the unknown words. When 1270 translated sentences of the test data have been manually checked, it is found that there are 250 unknown words of 132 sentences.

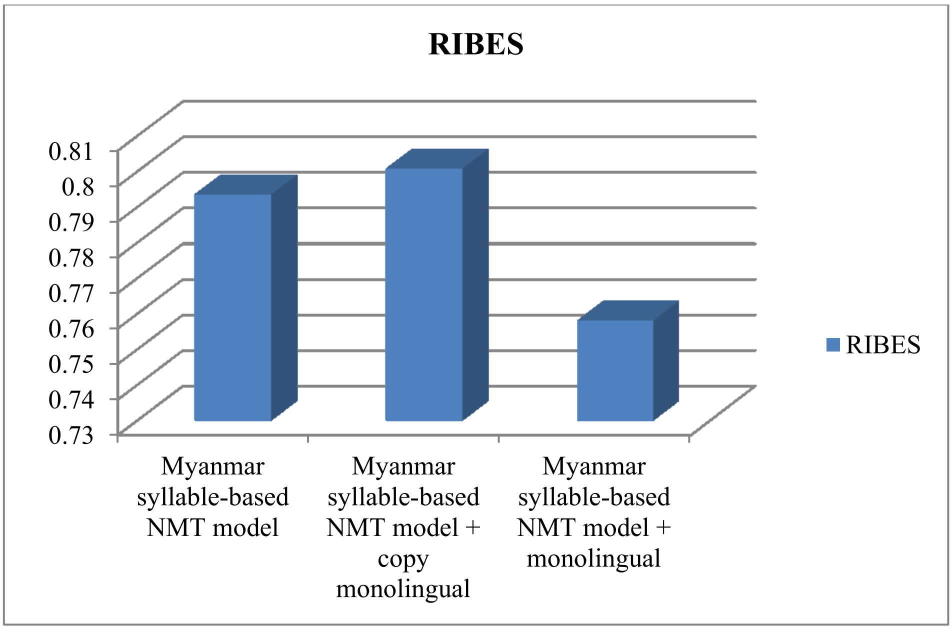


Figure 6.13 The Comparison of English to Myanmar Translation Using Monolingual Data in RIBES

CHAPTER 7
CONCLUSION AND FUTURE WORK

This chapter describes summarization of the research work and its limitations. It also indicates promising avenues for future research on Myanmar-English neural machine translation systems.

In this research, large scale Myanmar-English parallel corpus is built and proposed the neural machine translation system with attention model for Myanmar-English language pairs.

Myanmar is being regarded as a low-resourced language[93]. And there is no freely and commercially available Myanmar-English parallel corpus. Myanmar-English machine translation is still at an early stage and require to produce the improving translation results. This is due to the fact that do not have a few available parallel sentences for the Myanmar machine translation model. Therefore, the work on the thesis started with the collection of parallel sentences. In this work, large-scale Myanmar-English parallel corpus and Myanmar monolingual corpus that serves for the natural language processing tasks are built. The parallel corpus, called UCSY-corpus, is a general domain which includes the parallel sentences of local news, travel domain, school text book and spoken text. These parallel sentences are collected from the government official websites, online speaking websites and eBooks by crawling, downloading and copying manually. Therefore, the 200K parallel sentences are collected. And Myanmar monolingual corpus is local news data and consists of 170K Myanmar sentences.

As a next work, three NMT model with attention, namely, word-based NMT model(baseline system), Myanmar character-based NMT model and Myanmar syllable-based NMT model are trained. And the experimental results shows that the BLEU and RIBES of Myanmar syllable-based NMT models are the best performance in both directions.

To improve the baseline system, Myanmar Monolingual corpus is also used on Myanmar syllable-base NMT model, which is the best performance of Myanmar-English language pair. This setup contains two tasks for Myanmar-English Neural Machine Translation models in both directions. One is copy monolingual data usage and other is outside monolingual data. Monolingual data is trained without changing the existing neural network architecture. In Myanmar to English neural machine translation model, not only the BLEU score of Myanmar syllable-based NMT model with outside monolingual data is rarely higher but also the RIBES score of Myanmar syllable-based NMT model with copy monolingual data is rare higher. Moreover, both the BLEU score of Myanmar syllable-based NMT model with copy monolingual data and outside monolingual data are higher then the baseline system. The BLEU score of Myanmar syllable-based NMT model with outside monolingual data improve up to 5 BLEU over the baseline system. In the RIBES scores, the Myanmar syllable-based NMT model with copy monolingual data is the highest, while Myanmar syllable-based NMT model with outside monolingual data does not have much improvement than the baseline system.

7.1 Limitations of the System

The most common problems of NMT are that it translates the duplicate words and unknown words. To address the unknown word problems, the researchers solve with the previous approaches. These approaches are roughly categorized into three types: character-based, subword-based, and copy-based approaches. Although the unknown problem are addressed by using character-based NMT model, there still remains the unknown words. When 1270 translated sentences of the test data have been manually checked, it is found that there are 250 unknown words of 132 sentences.

It is also found that the third one of the training data is long sentences and phrases, and the other is only long sentences. In the translation results, the short sentences can translate well. Although the longer sentences cannot translate well exactly, their translation meaning understand as a human translator.

When Myanmar-English machine translation system outputs are carefully studied, Myanmar language is difficult from basic level. The difficulties are starting from its writing system up to linguistic structures including morphology information and the syntactic structures. In addition, all ways are complex to discourse the language structures. If there is the integration the source-side sentences information such as named entities and morphological analyzer, the performance of Myanmar-English machine translation system will further improve.

7.2 Future Work

To improve the quality of machine translation, large amount of training data are needed. By collecting the new bilingual training data, the translation results would significantly increase for training NMT as well as SMT systems. However, although the sheer amount of data, the quality of the data is important to be the best. To reach this purpose, the good quality and cleaned data should be collected to improve translation quality. In the future, existing Myanmar-English parallel corpus and monolingual corpus are not sufficient to train the translation model. Therefore, the collection of more data and more work of other neural models are necessary to increase the translation performance.

And UCSY-corpus is general domain and the data is mix in this corpus. Therefore, the bilingual sentences will be collected as the dominant corpus in the future.

In the area of the Myanmar-English NMT model, it is necessary to increase the improvement and to decrease the loss of accuracy. Thus, researchers have been trying to use the enriching bilingual dictionary in machine translation systems such as OOV remover and morphological analyzers as well as in the training data for Transliterator.

AUTHOR'S PUBLICATIONS

[p1] Yi Mon Shwe Sin, Yuzana, Khin Mar Soe, “Identifying Myanmar Phrases Using Conditional Random Field”, 15th International Conference on Computer Applications (ICCA), Yangon, Myanmar, 16th-17th February, 2017. Pg 91-95.

[p2] Yi Mon Shwe Sin, Khin Mar Soe, “Syllable-based Myanmar-English Neural Machine Translation”, 16th International Conference on Computer Applications (ICCA), Yangon, Myanmar, 22nd-23th February, 2018. Pg 228-233.

[p3] Yi Mon Shwe Sin, Khin Mar Soe, Khin Yadanar Htwe, “Large Scale Myanmar to English Neural Machine Translation System”, Proceeding of the IEEE 7th Global Conference on Consumer Electronic(GCCE2018), Nara, Japan, 9th-12th October, 2018.

[p4] Yi Mon Shwe Sin, Thazin Myint Oo, Hsu Myat Mo, Win Pa Pa, Khin Mar Soe, Ye Kyaw Thu, “UCSYNLP-Lab Machine Translation Systems for WAT2018”, 32nd Pacific Asia Conference on Language, Information and Computation (The 5th Workshop on Asian Translation PACLIC 32-WAT 2018), Hong Kong, 1st–3rd December 2018, pp 1127-1132.

[p5] Yi Mon Shwe Sin, Khin Mar Soe, "Attention-based Syllable level Neural Machine Translation System for Myanmar to English Language Pair" , International Journal on Natural Language Computing (IJNLC), Vol 8, No 2, April, 2019.

[p6] Yi Mon Shwe Sin, Khin Mar Soe, Dim Lam Cing, "Creation of Myanmar-English parallel corpus for Myanmar Natural Language Processing tasks", Myanmar Universities' Research Conference (MURC2019), 24th -25th May, 2019.

BIBLOGRAPHY

[1] N. T. Alsohybe, N. A. Dahan, F. M. Ba-Alwi, "Machine-Translation History and Evolution: Survey for Arabic-English Translations", Current Journal of Applied Science and Technology, 23(4): 1-19, 2017; Article no. CJAST.36124, Previously known as British Journal of Applied Science & Techonology. ISSN: 2231-0843, NLM ID:101664541.

[2] M. P. Aung, A. L. Moe, "New Phrase Chunking Algorithm for Myanmar Natural Language Processing."

[3] B. Babych, A. Hartley, "Extending the BLEU MT Evaluation Method with Frequency Weightings."

[4] D. Bahdanau, K. Cho, Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", Published as a conference paper at ICLR 2015.

[5] R. Bawden, R. Sennrich, A. Birch, B. Haddow, "Evaluating Discourse Phenomena in Neural Machine Translation", Proceedings of NAACL-HLT 2018, pages 1304–1313, New Orleans, Louisiana, June 1 - 6, 2018.

[6] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, J. Glass, "What Do Neural Machine Translation Models Learn about Morphology?"

[7] Y. Belinkov, J. Glass, "Large-Scale Machine Translation between Arabic and Hebrew: Available Corpora and Initial Results", Proceedings of the Workshop on Semitic Machine Translation, pages 7–12, Austin, Texas, USA, November 1, 2016.

[8] S. N. Ben, F. J. Och, G. Leusch, H. Ney, “An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research”, in Proceedings of the 2nd Language Resources and Evaluation Conference (LREC), 2000.

[9] F. Bond, S. Wang, E. H. Gao, H. S. Mok, J. Y. Tan, "Developing Parallel Sense-tagged Corpora with Wordnets", Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 149–158,

Sofia, Bulgaria, August 8-9, 2013.

[10] M. Chinea-Rios, ´A. Peris, F. Casacuberta, "Adapting Neural Machine Translation with Parallel Synthetic Data", Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 138–147 Copenhagen, Denmark, September 711, 2017. ©2017 Association for Computational Linguistics.

[11] J. Cho, H. Garcia-Molina, "Parallel Crawlers", WWW2002, May 7–11, 2002, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005.

[12] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Shwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation."

[13] J. Chung, K. Cho, Y. Bengio, "A Character-level Decoder Without Explicit Segmentation for Neural Machine Translation", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1693-1703, Berlin, Germany, August 7-12, 2016.

[14] M. Collins, "Phrase-based Translation Models."

[15] M. R. Costa-jussa, J. A. R. Fonollosa, "Character-based Neural Machine Translation", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 357–361, Berlin, Germany, August 7-12, 2016.

[16] F. Cromieres, C. Chu, T. Nakazawa, "Kyoto University Participation to WAT 2016", Proceedings of the 3rd Workshop on Asian Translation, pages 166–174, Osaka, Japan, December 11-17 2016.

[17] Y. David, "Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", Proceedings of the 14th conference on Computational linguistics- Nantes, France, 1992- Vol.2.pp.454.

[18] Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar. (2005)

[19] B. Dorr, P. W. Jordan and J. W. Benoit, "A Survey of Current Paradigms in Machine Translation" .Tech. Rep. CS-TR-3961,1998.

[20] K. Duh, G. Neubig, K. Sudoh, H. Tsukada, "Adaptation Data Selection Using Neural Language Models: Experiments in Machine Translation."

[21] B. Eric, "Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", Computational Linguistic. -1995.-Vol.21-pp. 543-565.

[22] M. Fadarr, A. Bisazza, H. Monz, "Data Augmentation for Low-Resource Neural Machine Translation."

[23] D. Farewell and Y. Wilks, "Ultra: A Multilingual Machine Translation", Technical Report MCCS-90-202, Computing Research Laboratory, New Mexico State University, 1990.

[24] T. Feng, M. Li, L. Chen, "Low-resource Neural Machine Translation with Transfer Learning."

[25] J. Gu, G. Neubig, K. Cho, V. O. K. Li, "Learning to Translate in Real-time with Neural Machine Translation", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1053–1062, Valencia, Spain, April 3-7, 2017.

[26] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Baeeault, H. Lin, F. Bougares, H. Schwenk, Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation."

[27] K. Hashimoto, A. Eriguchi, Y. Tsuruoka, "Domain Adaptation and Attention-Based Unknown Word Replacement in Chinese-to-Japanese Neural Machine Translation", Proceedings of the 3rd Workshop on Asian Translation, pages 75–83, Osaka, Japan, December 11-17 2016.

[28] T. H. Hlaing, Y. MIKAMI, "Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer."

[29] H. H. Htay, K. N. Murthy, “ Myanmar Word Segmentation using Syllable

level Longest Matching” , The 6th Workshop on Asian Language Resources, pp 41-48, 2008.

[30] P. Huang, F. Liu, S. Shiang, J. Oh, C. Dyer, "Attention-based Multimodal Neural Machine Translation", Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 639–645, Berlin, Germany, August 11-12, 2016.

[31] A. Imankulova, T. Sato, M. Komachi, "Improving Low-resource Neural Machine Translation with Filtered Pseudo-parallel Corpus", Proceedings of the 4th Workshop on Asian Translation, pages 70–78, Taipei, Taiwan, November 27, 2017.

[32] S. Islam, A. Paul, B. S. Purkayastha. I. Hussain, "Construction of English-Bodo Parallel Text Corpus for Statistical Machine Translation", International Journal on Natural Language Computing (IJNLC) Vol.7, No.5, October 2018.

[33] P. Jaimai, O. Chimeddorj, "Corpus Building for Mongolian Language", The 6th Workshop on Asian Language Resources, 2008.

[34] S. Jean, K. Cho, Y. Bengio, "On Using Very Large Target Vocabulary for Neural Machine Translation", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1–10, Beijing, China, July 26-31, 2015.

[35] A. R. Johansen, J. M. Hansen, E. K. Obeid, C. K. Sponderby, O. Winther, "Neural Machine Translation with Characters and Hierarchical Encoding."

[36] M. Johnson, M. Schuster, Q. V. Le, M. Krikum, Y. Wu, Z. Chen, N. Thorat, "Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation."

[37] D. Jones, A. Eisele, "Phrase-based Statistical Machine Translation between English and Welsh",

[38] T. Kajiwara, M. Komachi, "Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word

Embeddings", Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1147–1158, Osaka, Japan, December 11-17 2016.

[39] D. Kanojia, M. Shrivastava, "PaCMan : Parallel Corpus Management Workbench", S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 162–166, Goa, India. December 2014.

[40] A. Karakanta, J. Dehdari, J. V. Genabith, "Neural Machine Translation for Low-resource Languages without Parallel Corpora", DOI 10.1007/s10590-017-9203-5.

[41] G. Klein, Y. Kim, Y. Deng, J. Senellaart, A. M. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation."

[42] T. Kocmu, D. Varis, O. Bojar, "CUNI NMT System for WAT 2017 Translation Tasks", Proceedings of the 4th Workshop on Asian Translation, pages 154–159, Taipei, Taiwan, November 27, 2017.

[43] P. Koehn, "Neural Machine Translation."

[44] P. Koehn, R. Knowles, "Six Challenges for Neural Machine Translation", Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, Canada, August 4, 2017.

[45] P. Koehn, F. J. Och, D. Marcu, "Statistical Phrase-Based Translation."

[46] S. M. Lakew, M. Cettolo, M. Federico, "A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation."

[47] G. Lample, L. Denoyerm, M. A. Ranzato, "Unsupervised Machine Translation Using Monolingual Corpora Only", Under review as a conference paper at ICLR 2018.

[48] J. Lee, K. Cho, T. Hofmann, "Fully Character-Level Neural Machine Translation without Explicit Segmentation", Transactions of the Association for Computational Linguistics, vol. 5, pp. 365–378, 2017. Action Editor:

Adam Lopez. Submission batch: 11/2016; Revision batch: 2/2017; Published 10/2017.

[49] P. Levin, N. Dhanuka, T. Khalil, F. Kovalev, M. Khalilov, "Toward A Full-scale Neural Machine Translation in Production: the Booking.com use case."

[50] H. Li, "Adequacy-Fluency Metrics (AM-FM) for Machine translation(MT) Evaluation".

[51] X. Li, J. Zhang, C. Zong, "Towards Zero Unknown Word in Neural Machine Translation", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).

[52] W. Ling, I. Trancoso, C. Dyer, A. Black, "Character-based Neural Machine Translation", Under review as a conference paper at ICLR 2016.

[53] V. Liu, J. R. Curran, "Web Text Corpus for Natural Language Processing."

[54] H. LIU, M. NUO, J. WU, Y. HE, "Building Large Scale Text Corpus for Tibetan Natural Language Processing by Extracting Text from Web Pages", Proceedings of the 10th Workshop on Asian Language Resources, pages 11–20, COLING 2012, Mumbai, December 2012.

[55] M. Luong, C. D. Manning, "Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1054–1063 ,2016.

[56] M. Luong, C. D. Manning, "Stanford Neural Machine Translation Systems for Spoken Language Domains."

[57] M. Luong, H. Pham, C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal, 17-21 September 2015.

[58] M. Luong, I. Sutskever, Q. V. Le, O. Vinyals, W. Zaremba, " Addressing the Rare Word Problem in Neural Machine Translation", Proceedings of the 53rd

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 11–19, Beijing, China, July 26-31, 2015.

[59] "Machine Translation Models based on Non-parallel Corpus and Semi-supervised Transductive Learning."

[60] C. D. Manning, P. Raghavan, H. Schutze, "An Introduction to Information Retrieval" Draft of April 1, 2009, Online edition 2009 Cambridge UP, Cambridge University Press Cambridge, England.

[61] Y. Matsumura, M. Komachi, "Tokyo Metropolitan University Neural Machine Translation System", Proceedings of the 4th Workshop on Asian Translation, pages 160–166, Taipei, Taiwan, November 27, 2017.

[62] T. Matsuzaki, A. Fujita, N. Todo, N. H. Arai, "Evaluating Machine Translation Systems with Second Language Proficiency Tests", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 145–149, Beijing, China, July 26-31, 2015.

[63] Z. M. Maung, Y. Mikami, " A Rule-based Syllable Segmentation of Myanmar Text", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 51-58, Hyderabad, India, January 2008.

[64] F. Meng, Z. Lu, H. Li, Q. Liu, "Interactive Attention for Neural Machine Translation", Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2174–2185, Osaka, Japan, December 11-17 2016.

[65] H. Mino, M. Utiyama, E. Sumita, T. Tokunaga, "Key-value Attention Mechanism for Neural Machine Translation", Proceedings of the 8th International Joint Conference on Natural Language Processing, pages 290–295, Taipei, Taiwan, November 27 – December 1, 2017.

[66] T. Mitamura, E. Nyberg and J. Carbonell, "An Efficient Interlingua Translation System for Multilingual-document Production," 1991.

[67] MLC. 2002, “Myanmar-English Dictionary”, Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[68] M. Morishita, Y. Odam, G. Neubig, "An Empirical Study of Mini-batch Creation Strategies for Neural Machine Translation", Proceedings of the First Workshop on Neural Machine Translation, pages 61–68, Vancouver, Canada, August 4, 2017.

[69] M. Morishita, J. Suzuki, M. Nagata, "NTT Neural Machine Translation Systems at WAT 2017", Proceedings of the 4th Workshop on Asian Translation, pages 89–94, Taipei, Taiwan, November 27, 2017.

[70] T. Q. Nguyen, D. Chiang, "Transfer Learning Across Low-Resource, Related Languages for Neural Machine Translation."

[71] K. T. Nwet, K. M. Soe, "Myanmar-English Machine Translation Model."

[72] R. Ostling, Y. Scherrer, J. Tiedemann, G. Tang, T. Nieminen, "The Helsinki Neural Machine Translation System", Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 338–347, Copenhagen, Denmark, September 711, 2017.

[73] W. P. Pa, N. L. Thein, "Myanmar Word Segmentation using Hybrid Approach", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170.

[74] W. P. Pa, Y. K. Thu, A. Finch, E. Sumita, "A Study of Statistical Machine Translation Methods for Under Resources Languages", 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, Yogyakaeta, Indonesia.

[75] S. Pal, P. Pakray, S. K. Naskar, "Automatic Building and Using Parallel Resources for SMT from Comparable Corpora", Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 48–57, Gothenburg, Sweden, April 27, 2014.

[76] K. Papineni, S. Roukos, T. Ward, W. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting

of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[77] A. K. Pathak, P. Acharya, D. Kaur, R. C. Balabantaray, "Automatic Parallel Corpus Creation for Hindi-English News Translation Task."

[78] M. T. Pilevar, H. Faili, A. H. Pilevar, "TEP Tehran English-Persian Parallel Corpus", A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 68–79, 2011. © Springer-Verlag Berlin Heidelberg 2011.

[79] M. Post, C. Callison-Burch, M. Osborne, "Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing", Proceedings of the 7th Workshop on Statistical Machine Translation, pages 401–409, Montre´al, Canada, June 7-8, 2012.

[80] Pytorch-OpenNMT. http://github.com/OpenNMT/OpenNMT-py

[81] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, R. Sun, V. Chea, K. M. Soe, K. T. Nwet, M. Utiyama, C. Ding, "Introduction of Asian Language Treebank with a Suvery of Asian NLP Resources".

[82] K. P. Scannell, "The Crúbadán Project Corpus Building for Under-resourced Languages", Cahiers du Cental, 5(2007).

[83] Seljan, T. Vicic, M. Brkic, "BLEU Evaluation of Machine-Translated English-Croatian Legislation."

[84] R. Sennrich, B. Haddow, A. Birch, "Improving Neural Machine Translation Models with Monolingual Data."

[85] R. Sennrich, B. Haddow, A. Birch, "Neural Machine Translation of Rare Words with Subword Units", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1715–1725, Berlin, , August 7-12, 2016.

[86] S. Singh, R. Panjwani, A. Kunchukuttan, P. Bhattacharyya, "Comparing Recurrent and Convolutional Architectures for English-Hindi Neural Machine

Translation", Proceedings of the 4th Workshop on Asian Translation, pages 167–170, Taipei, Taiwan, November 27, 2017.

[87] H. Shimanaka, T. Kajiwara, M. Komachi, "Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations", Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 106–111, New Orleans, Louisiana, June 2 - 4, 2018.

[88] D. Shterionov, P. Nagle, L. Casanellas, R. Superbo, T. O'Dowd, "Empirical Evaluation of NMT and PBSMT Quality for Large-scale Translation Production."

[89] T. T. Soe, A. Thida, "Applying Rule-Based Maximum Matching Approach for Verb Phrase Identification and Translation(Myanmar to English)", International Journal of Science and Research(IJSR), India Online ISSN: 2319-7064, Volume 2 Issue 9, September 2013.

[90] L. Tang, J. Xu, X. Hu, Q. Wei, H. Xu, "Building a Biomedical Chinese-English Parallel Corpus from MEDLINE".

[91] T. T. Thet, J. Na, W. K. Ko, "Word Segmentation for the Myanmar Language", Journal of Information Science, 34 (5) 2008, pp 688-704, first published on April 3, 2008.

[92] Y. K. Thu. 2017. Syllable segmentation tool for Myanmar language (Burmese). https://github.com/ye-kyaw-thu/sylbreak.

[93] Y. K. Thu, V. Chea, A. Finch, M. Utiyama, E. Sumita, "A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language", 29th Pacific Asia Conference on Language, Information and Computation pages 259-269, Shanghai, China, October 30 – November 1, 2015.

[94] Y. K. Thu, W. P. P., Y. Sagisaka, N. Iwahashi, "Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar pronunciation Dictionary", Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, pages 11–22, Osaka, Japan, December 11-17 2016.

[95] N. Thinnyunypu, "Myanmar Script Learning Guide."

[96] L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, Y. Lu, S. Li, Y. Wang, L. Wang , "UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation."

[97] H. Uchida, "Fujitsu Machine Translation System ATLAS". In proceeding of International Sytmposition M, 1985T.

[98] Unicode.https://en.wikipedia.org/wiki/Unicode

[99] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, J. Uszkoreit, "Tensor2Tensor for Neural Machine Translation."

[100] A. Waibel, A. Lavie, and L. S. Levin, "Janus: A System for Translation of Conversational Speech". Kunstliche Intelligenz, 1997.

[101] B. Wang, Z. Tan, J. Hu, Y. Chen, X. Shi, "XMU Neural Machine Translation Systems for WAT 2017", Proceedings of the 4th Workshop on Asian Translation, pages 95–98, Taipei, Taiwan, November 27, 2017.

[102] S. Wang, F. Bond, "Building The Sense-Tagged Multilingual Parallel Corpus."

[103] M. T. Win, M. M. Win, M. M. Than, "Burmese Phrase Segmentation", Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

[104] S. Wu, M. Zhou, D. Zhang, "Improved Neural Machine Translation with Source Syntax", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

[105] J. Wu, H. Hou, Z. Shen, J. Du, J. Li, "Adapting Attention-based Neural Network to Low-resource Mongolian-Chinese Machine Translation", (1777)©Springer-Verlag Berlin Heidelberg 2011.

[106] H. Xiao, X. Wang, "Constructing Parallel Corpus from Movie Subtitles", ICCPOL 2009, LNAI 5459, pp. 329–336, 2009. © Springer-Verlag Berlin

Heidelberg 2009.

[107] J. Xu, G. hen, "Phrase Based Language Model for Statistical Machine Translation."

[108] M. Yang, H. Jiang, T. Zhao, S. Li, "Construct Trilingual Parallel Corpus on Demand."

[109] Z. Yang, W. Chen, F. Wang, B. Xu, "A Character-Aware Encoder for Neural Machine Translation", Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3063–3070, Osaka, Japan, December 11-17 2016.

[110] E. Yildiz, A. C. Tantug, B. Diri, "The Effect of Parallel Corpus Quality Vs Size in English-to-Turkish SMT", Natarajan Meghanathan et al. (Eds) : ICCSEA, SPPR, VLSI, WiMoA, SCAI, CNSA, WeST – 2014, pp. 21–30, 2014.

[111] H. Yu, X. Zhu, "Recurrent Neural Networks Based Rule Sequence Model for Statistical Machine Translation", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 132–138, Beijing, China, July 26-31, 2015.

[112] R. Zens, H. Ney, "Improvements in Phrase-Based Statistical Machine Translation.

[113] R. Zens, F. J. Och, H. Ney, "Phrase-Based Statistical Machine Translation", M. Jarke et al. (Eds.): KI 2002, LNAI 2479, pp. 18–32, 2002. ©Springer-Verlag Berlin Heidelberg 2002.

[114] J. Zhang, C. Zong, "Briding Neural Machine Translation and Bilingual Dictionaries."

[115] J. Zhang, C. Zong, "Exploiting Source-side Monolingual Data in Neural Machine Translation", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Austin, Texas,

November 1-5,2016.

[116] S. Zhao, Z. Zhang, "An Efficient Character-Level Neural Machine Translation", Under review as a conference paper at ICLR 2017.

[117] S. Zhao, Z. Zhang, "Deep Character-level Neural Machine Translation by Learning Morphology", Under review as a conference paper at ICLR 2017.

[118] Z. Zhu, "Evaluating Neural Machine Translation in English-Japanese Task."

[119] B. Zoph, D. Yuret, J. May, K. Knight, "Transfer Learning for Low-Resource Neural Machine Translation", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas, November 1-5, 2016.

[120] T. T. Zin, K. M. Soe, N. L. Thein, "Improving Phrase-based Statistical Myanmar to English Machine Translation with Morphological Analysis", International Journal of Computer Applications(0975-8887) Volume 28-No. 1,August 2011.

[121] T. T. Zin, K. M. Soe, N. L. Thein, "Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System", 25th Pacific Asia Conference on Language, Information and Computation, pages 130–139.

[122] မြန်မာစာလုံးပေါင်းသတ်ပုံကျမ်း။

APPENDICES

Pytorch OpenNMT toolkit is used to implement the Myanmar-English Neural Machine Translation models in both directions. In this section, we will present how to implement the Myanmar-English neural translation models including the Experimental setup for Neural Machine Translation System, creation of Myanmar-English parallel corpus and Myanmar monolingual corpus, data preprocessing step, training models and translation steps.

Appendix A: Experimental Setup for Neural Machine Translation Systems

All experiments were conducted on Nvidia Tesla K80 GPU. As requirements, Python 2 or 3 must be installed. To download and install python, follow the installation note at python official website: https://www.python.org/.

Before the installation pytorch, it is needed to use at least one of two supported package managers: Anaconda and pip. Commands to install from binaries via Conda or pip wheel are on the PyTorch official website: https://pytorch.org. To do so, as prerequisites, Anaconda was firstly installed. To install Anaconda, Anaconda installation guide documentation can be referenced at https://docs.anaconda.com/anaconda/. Then we install the PyTorch.

Pytorch, port of OpenNMT, is an open-source neural machine translation system. It is designed to be research friendly to try out new ideas in translation, summary, image-to-text, morphology, and many other domains. Some companies have proven the code to be production ready. To use the pytorch OpenNMT, the followings steps are needed to install.

1. Install PyTorch

To install Pytorch, operating system, package, language and CUDA version are selected. According to the selected item, the official website gives the install command to run on terminal.

Command:

```
conda install pytorch torchvision cudatoolkit=10.0 -c pytorch
```

Now, PyTorch have finished the installations successfully. It's time to test PyTorch by executing torch program.

2. Get OpenNMT-py

Clone the OpenNMT-py git repository on Github into a local folder. The commands are as follows:

```
git clone https://github.com/OpenNMT/OpenNMT-py
cd OpenNMT-py
```

3. Install required libraries

```
pip install -r requirements.txt
```

To check whether PyTorch was successfully installed or not, run the following commands.

```
Python
>>import torch
>>torch.cuda.is_available()
True
```

If True is returned, it is ready to use the pytorch on the machine. And then, neural machine translation models for Myanmar-English language pair was trained by configuring different hyper-parameter settings.

Appendix B: Creation of UCSY-corpus and Myanmar Monolingual Corpus

In this section, data collection and data preprocessing step for Myanmar-English neural machine translation models are presented.

1. Data Collection

Data collection is a essential step to develop Myanmar-English NMT models. The performance of machine translations is depends on the availability of parallel corpus. In our experiments, we use two parallel corpora: the ALT corpus and the UCSY corpus. ALT corpus consists of twenty thousand Myanmar-English parallel sentences from Wiki news articles. The UCSY corpus consists of 200K Myanmar-English parallel sentences collected from different domains, including Local news, Travel Domain, school text books and spoken text. Local News is from the government official websites such as Myanma Alin, The Global New Light of Myanmar. It starts to collect from Janay 2016 to December 2017. Travel Domain contains about 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). School text books and spoken text mainly contain widely used spoken English. There are 220K parallel sentences in total. And Myanmar monolingual corpus is constructed to improve the performance of machine translation systems. The collected sentences of Myanmar monolingual corpus is local news data and 170 thousand Myanmar sentences.

The collected sentences include sometimes noise data such as HTML tags, the color, font, style attribute and the extra spaces. These noise is to remove in the collected sentences. And the duplicate sentences are also removed. Another problem is that this dataset contains spelling errors. Thus, we are manually modified the spelling error in order to clean noisy data. All Myanmar-English parallel sentences have been converted into UTF-8 text file format and Myanmar3 font. The corpus preparation is a very important task to train the NLP tasks and it takes for one year and six months.

2. Data Preprocessing

After cleaning the corpus, this corpus is ready to use in the research of the Natural Language Processing tasks. The corpus is randomly divided into training data, development data and test data. Therefore, the data for Myanmar-English and English-Myanmar translation tasks is a mix domain data collected from different sources.

The collected raw sentences are not segmented correctly and some do not have almost no segmentation is essential for the quality improvement of Machine Translation. As the preprocessing step, we use syllable segmentation that is the python script described at: https://github.com/ye-kyaw-thu/sylbreak/tree/master/python is applied. "sylbreak" [91] is segment the Myanmar sentence into syllable level. After that, the data preparation have finished and ready to train the NMT models. The implementation of Myanmar-English neural machine translation model are described in Appendix C.

Appendix C: Implementation of Myanmar-English Neural Machine Translation Models

In this work, Preprocess the data, Train the model and Translate the test sentences are presented.

1. Preprocess the data

The preprocessing script takes training file and validation file in both of source and target sides as input. In our experiment, training files has 2200K sentences. Validation files have 2200 sentences and testing files have 1270 sentences. This preprocessing script is done by using the following command:

```
python preprocess.py
-train_src data/src-train.txt
 -train_tgt data/tgt-train.txt
-valid_src data/src-val.txt
-valid_tgt data/tgt-val.txt
-save_data data/demo
```

Here parallel data set is labeled as to source(src) and target(tgt). The dataset has one sentence per line and every Myanmar syllables token and English words token are separated by a space. And we have source(src) and target(tgt) training, test and validation files in it. This script will yield the three output files. They are

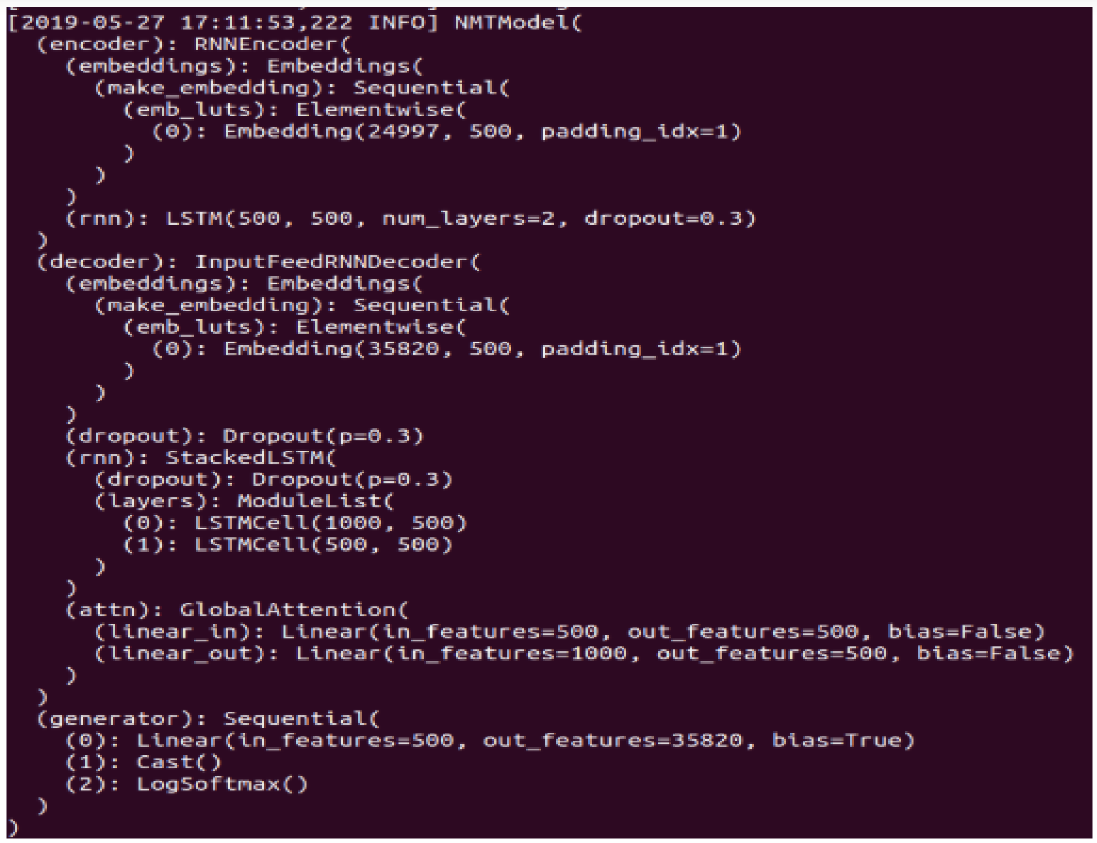
1. demo.train.pt: serialized PyTorch file containing training data
2. demo.valid.pt: serialized PyTorch file containing validation data
3. demo.vocab.pt: serialized PyTorch file containing vocabulary data

2. Train the model

The main train command is quite simple. The training script takes data files that is the output of the preprocessing step. And it outputs the model files. Typically, the default model goes on till 100000 epochs, such that a check-point is saved after every 5000 epochs. This will run the default model, which consists of a 2-layer LSTM with 500 hidden units on both the encoder/decoder. This training script is done by using the following command:

```
python train.py
-data data/demo
-save_model demo-model
```

This generates the 20 model files as default model. And it also generates the accuracy value and perplexity value of each model files. The following figure shows the screen output of the model when we run the above command.



```
[2019-05-27 17:11:53,222 INFO] NMTModel(
  (encoder): RNNEncoder(
    (embeddings): Embeddings(
      (make_embedding): Sequential(
        (emb_luts): Elementwise(
          (0): Embedding(24997, 500, padding_idx=1)
        )
      )
    )
    (rnn): LSTM(500, 500, num_layers=2, dropout=0.3)
  )
  (decoder): InputFeedRNNDecoder(
    (embeddings): Embeddings(
      (make_embedding): Sequential(
        (emb_luts): Elementwise(
          (0): Embedding(35820, 500, padding_idx=1)
        )
      )
    )
    (dropout): Dropout(p=0.3)
    (rnn): StackedLSTM(
      (dropout): Dropout(p=0.3)
      (layers): ModuleList(
        (0): LSTMCell(1000, 500)
        (1): LSTMCell(500, 500)
      )
    )
    (attn): GlobalAttention(
      (linear_in): Linear(in_features=500, out_features=500, bias=False)
      (linear_out): Linear(in_features=1000, out_features=500, bias=False)
    )
  )
  (generator): Sequential(
    (0): Linear(in_features=500, out_features=35820, bias=True)
    (1): Cast()
    (2): LogSoftmax()
  )
)
```

3. Translate the model

The translation script use the model file and test file as the input. The model file is the output of the training script and the test file is that wants to translate.

```
python translate.py
-model demo-model.pt
-src data/src-test.txt
-output pred.txt
```

The model file is replaced with the own model. The translation script will generate the translated output and store the predictions into a file named pred.txt.