STATISTICAL MACHINE TRANSLATION SYSTEM BETWEEN KAREN AND ENGLISH LANGUAGE USING PBSMT MODEL

SHARO PAW

M.C.Sc. **SEPTEMBER 2022**

STATISTICAL MACHINE TRANSLATION SYSTEM BETWEEN KAREN AND ENGLISH LANGUAGE USING PBSMT MODEL

By

SHARO PAW

B.C.Sc(Hons:)

A Dissertation Submitted in Partial Fulfillment of the Requirement for the Degree of

Master of Computer Science

M.C.Sc.

University of Computer Studies, Yangon
September 2022

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 Sharo Paw

ACKNOWLEDGEMENTS

To complete this thesis, many things are needed like my hard work and the supporting of many people who gave a lot of idea to me to complete this thesis.

Firstly, I am to express my appreciation and thanks to **Dr. Mie Mie Khin**, Rector of the University of Computer Studies, Yangon, for giving the opportunity to develop this thesis.

I greatly appreciate and thanks to **Dr. Tun Myat Aung**, Principal and Pro-rector of the University of Computer Studies, Hinthada, for his kind permission to submit this dissertation.

I am thankful to my thesis supervisors, **Dr. Hmway Hmway Tar,** Professor of Faculty of Computer Science, University of Information Technology, Yangon, for her close supervision, proper guidance, valuable suggestions, guidance, advice and encouragement during the course of this work.

I would like to express and appreciate my special thanks to my course coordinators, **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw,** Professor of Faculty of Computer Science Department of the University of Computer Studies, **Dr. Yuzana**, Pro- Rector of the University of Computer Studies, Pyay and **Dr. Yi Mon Shwe Sin**, Lecturer of the University of Computer Studies, Yangon for their superior suggestions and administrative supports during my academic study.

I also would like to express my appreciation to **Daw Win Lai Lai Bo,** Assistant Lecturer of Department of English of University of Computer studies, Yangon, for her advice, editing and suggestion from the language point of view.

I would also like to acknowledge my thanks to **all my teachers** who taught me throughout the master's degree course and wish to express my gratitude to **my beloved family** for their invaluable support and encouragement to fulfill my wish.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Nowadays, there are top performance of machine translation systems for some foreign languages (high resource languages). Machine Translation (MT) is the automatic translation mechanism from one natural language into another language by means of a computerized system. There are many researches using machine translation systems is not only foreign languages but also Myanmar Ethnic languages (lower source languages) such as English-Myanmar, Myanmar-Rakhine, Myanmar-Dawei and Kachin-Rawang and so on. In this system, over 10K Karen-English parallel sentences are collected from Karen-English published books via internet and other sources. And the phrase-based statistical machine translation system is proposed by using the Moses toolkit for Karen and English language pairs. The word segmented source language was aligned with the word segmented target language using GIZA++. The alignment was symmetrized by grow-diag-final and heuristic. The lexicalized reordering model was trained with the msd-bidirectional-fe option. We use KenLM and SRILM for training with 2-gram, 3-gram and 5-gram language models for both Karen to English and English to Karen langauage pairs. Minimum error rate training (MERT) was used to tune the decoder parameters and the decoding was done using the Moses decoder. Finally, the experimental results of the system are measured in terms of BLEU scores and compared them. For Karen to English PBSMT model, the expeimental result of KenLM with 5-gram languge model is the best. And the expeimental result of KenLM with 3-gram languge model is the best for English to Karen PBSMT model.

CHAPTER 1
INTRODUCTION

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. The Statistics' development, Statistical Machine Translation (SMT) is popular in research area in the late 1980s. SMT also translates from source to target language based on the parallel corpora. This method performed and produced the better result than other method such as RBMT (Rule-based Machine Translation). Word-based, phrase-based, syntax-based and hierarchical phrase-based are the approaches based on SMT. The most prevalent version of SMT is Phrase-based SMT (PBSMT), which in general includes pre-processing, sentence alignment, word alignment, phrase extraction, phrase feature preparation, and language model training. The key component of a PBSMT model is a parallel corpus.

For Myanmar language, the automatic machine translation systems began in 2010. Besides, there are many translation systems in Myanmar Ethnic languages such as English-Myanmar, Myanmar-Rakhine, Myanmar-Dawei and Kachin-Rawang and so on. The proposed system is for Karen language which is one of the eight Myanmar Ethnic languages. The Karen languages, members of the Tibeto-Burman group of the Sino-Tibetan language family, consist of three mutually unintelligible branches: Sgaw, Eastern Pwo, and Western Pwo. The ethnic groups that speak Sgaw karen languages include many groups such as Kayin, Pa-Le-Chi, Paku, Bwe, Monnepwa, Monpwa, Shu.

In the proposed system, phrase-based statistical machine translation (PBSMT) model is used to translate from the source language to target language (Karen-English or English-Karen). Firstly, over 10K parallel corpus is collected manually via the internet, English language books published by Cambridge University and Karen language books. As preprocessing step, Moses's tokenization scripts are used to segment for English sentences and Karen sentences are segmented manually. And Moses's cleaning scripts are used for both languages. To get the translation model for each language pair, the word from the source language is segmented and aligned with the word segmented target language using GIZA+. The alignment is symmetrized by

grow-diag-final and heuristic. The lexicalized reordering model was trained with the msd-bidirectional-fe option. For language model, KenLM and SRILM are used for each language pair. Finally, the decoder describes a simple phrase-based translation model consisting of phrase-pair probabilities of translation model and language model. There are six experiments for each language pairs and these experimental results are measured in terms of BLEU score. For Karen to English PBSMT model, the expeimental result of KenLM with 5-gram languge model is the best. And the expeimental result of KenLM with 3-gram languge model is the best for English to Karen PBSMT model.

1.1 Statistical Machine Translation (SMT)

The statistical approach breaks the source text down into segments then compares them to an aligned bilingual corpus, using statistical evidence and distortion probabilities to choose the most appropriate translation. Given one sentence, SMT divides it into several sub-sentences, then every part could be replaced by target word or phrase. The most prevalent version of SMT is Phrase-based SMT (PBSMT), which in general includes pre-processing, sentence alignment, word alignment, phrase extraction, phrase feature preparation, and language model training. The key component of a PBSMT model is a phrase-based lexicon, which pairs phrases in the source language with phrases in the target language. The lexicon is built from the training data set which is a bilingual corpus. By using phrases in this translation, the translation model could utilize the context information within phrases. Thus, PBSMT could outperform the simple word-to-word translation methods.

The statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The SMT is a corpus based approach, where a massive parallel corpus is required for training the SMT systems. The SMT systems are built based on two probabilistic models: language model and translation model. The advantage of SMT system is that linguistic knowledge is not required for building them. The difficulty in SMT system is creating massive parallel corpus. There are three different statistical approaches in machine translation. They are word-based translation, phrase-based translation, and hierarchical phrase-based model. There are also several types of statistical-based machine translation models.

They are: hierarchical phrase-based translation, Syntax-based translation, Phrase-based translation, and Word-based translation.

1.2 Motivation of the Thesis

The previous work of machine translation mostly learnt by using rule-based as well as statistically based approach. However, the machine translation of Karen language is lack and there is no parallel corpus. Therefore, a English-Karen parallel corpus are created and proposed a phrase-based statistical machine translation system between Karen and English language pairs. This system is able to translate for people who are not understand the Karen language and address the language barrier about Karen and English languages. This system has the ability which can translate faster than human translators and save costs is more effective for most users. And the system can translate more relevant and specific results. Due to these facts, using PBSMT is more convenient than human translators.

1.3 Objectives of the Thesis

The objectives of the thesis are as follows:

(i) To create Karen-English parallel corpus
(ii) To evaluate the quality of machine translation between Karen and English languages
(iii) To know about the machine translation (MT)
(iv) To learn phrase-based statistical machine translation (PBSMT) model
(v) To examine BLEU scores between Karen-English and English-Karen
(vi) To describe detail analysis on confusion pairs of Karen-English and English- Karen

1.4 Organization of the Thesis

This thesis is organized into abstract, acknowledgment, five chapters, and references. In chapter 1, the introduction of Machine Translation (MT), statistical Machine Translation (SMT) and Phrased-Based Statistical Machine Translation (PBSMT) are presented. And the proposed system of the thesis is introduced. Besides, the motivation, objectives and organization of this thesis are described. In chapter 2,

an introduction to natural language processing, statistical natural language processing, building Karen-English Parallel Corpus and Karen Language and the literature reviews of machine translation are explained. In chapter 3, the background theory of Phrase-based Statistical Machine Translation System is explained in detailed. In chapter 4, Design and implementation of Statistical Machine Translation System between Karen and English Language Using PBSMT are presented. First of all, the system flow diagram, the system's architecture, and structure are described. Finally, the experimental results are shown with charts and tables. The conclusion of thesis and further extensions made with some ideas are presented in chapter five. The limitations of the system are also described in this chapter.

CHAPTER 2
BUILDING KAREN-ENGLISH PARALLE CORPUS

Firstly, an introduction to natural language processing is described and statistical natural language processing. And then, about the building Karen-English Parallel Corpus, Karen Language and the literature reviews of machine translation are briefly explained.

2.1 Natural Language Processing

Natural language processing (NLP) is a form and 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 are to work with the natural input and output.

Language, a system to communicate with one to one and one to many communications in daily lives. It is sign of both verbal and written expression to help people for displaying thoughts, and feelings caused in their mind. Languages are a wide range of sounds, sings, and symbols to create words, sentences, paragraphs and other media and the mediums used for expression and organization wanted to know thoughts and feelings.

As the fact, most languages become and appear changes in environments, social and become popular as technologies changes. There are formal languages and they are artificial and contrived languages. These are developed to purpose specially. The formal language example is a computer language. This language is used and permitted as a limited form of communication between people and computers.

According to the rules for developing the system, structuring words or phrases in the language sentences together to produce a form of complete sentence and thoughts is known as grammar. Structure of grammar consists of two basic parts to

produce a right, good and complete sentence, syntax and semantics and then orders subjects, verbs, objects, places, manners, time and reasons. Syntax is the arranging of words and phrases to form the sentences. It is a subset of grammar's structure concerned with the putting and ordering of various words or phrases 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 means a function of syntax. It refers to the meaning of language. It is studied the relationships between words and the way they are assembled to present a specific event. Semantics provides the analyzed and interpreted way which is being said. The way words are used and ordered very much determined the meaning of the combination.

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

Context refers to the complete idea or thought 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. One thing but mean another are often said and written. For example, in the question "what time it is", but really meaning for this question is "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.

As the theory, natural language processing is a emulated form and method to between human and computer. natural language processing does 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. So, the main goal of natural language processing is to understand how exactly the human beings understand, generate and learn languages.

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 Building Karen-English Parallel 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.

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. It is also the first step in building a translation model for low resource languages where there is no pre-created parallel corpus. Because there is no Karen-English parallel corpus also, this system proposed to build Karen-English parallel corpus. Karen language is regarded as a low resource language, so there are some difficulties to build a parallel corpus. Building a Karen-English parallel corpus for Karen language is conditioned by various factors like the availability of the texts in that language.

Some parallel sentences are collected from Karen-English published books via internet. Some parallel sentences are collected by translating manually from English to Karen language or Karen to English languages. Therefore, Karen-English parallel corpus is a general domain. The corpus consists of over 10K parallel sentences collected from different domains. Example of Karen-English parallel corpus is shown in Figure 2.1.

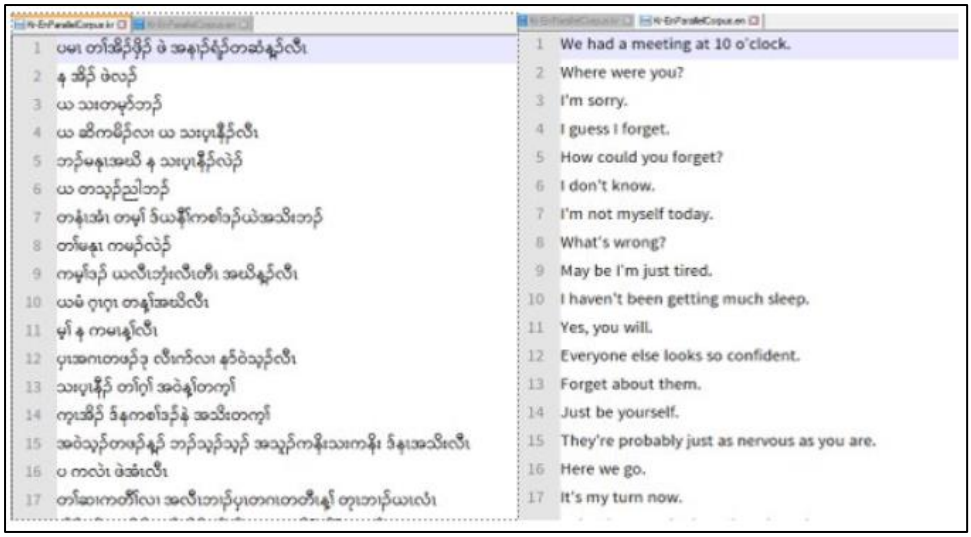

Figure 2.1 Karen-English parallel corpus

2.3 Karen Language

Myanmar is an ethnically diverse nation with 135 distinct ethnic groups officially recognized by the Burmese Government. These are grouped into eight "major national ethnic races" are Kachin, Kayah, Karen, Chin, Bamar, Mon, Rakhine and Shan.

Among them, Karen People did not originate from Burma. The Karen are believed to originate from Mongolia. Their southward migration would take them across the Gobi Desert on the Chinese-Mongolian border. The term "Kayin" translates as "flowing sands", supporting the Gobi hypothesis. They now live throughout much of lower Burma, with the main populations in the Irrawaddy Delta and Thai borderlands. The long migration resulted in the Karen reaching modern Myanmar between the third to eighth centuries.

Even though Karen and Burmese are written in the same script and derive from the same Sino-Tibetan language family, they are different languages and the speakers of one cannot understand the other. The Karen alphabet was derived from the Burmese script as created by the help of the **American missionary Jonathan Wade** around the 1830s. The Karen alphabet was created for the purpose of

translating the Bible into the Karen language. They are unusual among the Sino-Tibetan languages in having a subject–verb–object word order.

There are 11 karen ethanic groups. They are:

1. Kayin
2. Kayinpyu
3. Pa-Le-Chi
4. Mon Kayin
5. Sgaw
6. Ta-Lay-Pwa
7. Paku
8. Bwe
9. Monnepwa
10. Monpwa
11. Shu

The Karen languages, members of the Tibeto-Burman group of the Sino-Tibetan language family, consist of three mutually unintelligible branches: Sgaw, Eastern Pwo, and Western Pwo. The ethnic groups that speak Sgaw karen languages include many groups such as Kayin, Pa-Le-Chi, Paku, Bwe, Monnepwa, Monpwa, Shu.

Karen is a Sino-Tibetan language spoken by the S'gaw Karen people of Myanmar and Thailand. A Karenic branch of the Sino-Tibetan language family, S'gaw Karen is spoken by over 2 million people in Tanintharyi Region, Ayeyarwady Region, Yangon Region, and Bago Region in Myanmar, and about 1 million in northern and western Thailand along the border near Kayin State. A few Karen have settled in the Andaman and Nicobar Islands, India, and other Southeast Asian and East Asian countries.

There are main eight ethnic groups with respective languages in Myanmar Nations: Kachin, Kayah, Kayin, Chin, Mon, Burma, Rakhine and Shan.

The Karen alphabet consist of 25 consonants, 9 vowels, 5 tones and 5 medials. Grouped consonants are show in Figure 2.2.

က k (kaˀ)	ခ kh ($k^haˀ$)	ဂ gh (ɣ)	ဃ x (x)	င ng (ŋ)
စ s (s)	ဆ hs (s^h)	ၐ sh (ʃ)	ည ny (ɲ)	တ t (t)
ထ hṭ (t^h)	ဒ d (d)	န n (n)	ပ p (p)	ဖ hp (p^h)
ဘ b (b)	မ m (m)	ယ y (j)	ရ r (r)	လ l (l)
ဝ w (w)	သ th (θ)	ဟ h (h)	အ vowel holder (ʔ)	ဧ ahh (ɦ)

Figure 2.2 Grouped consonants of Karen Language

In grouped consonants,

- **က** has a sound intermediate between **k** and **g**; as in g for **g**ood
- **ခ** is the aspirate of **က**. It is pronounced like **kh** as heard in the word **c**amp.
- **ဂ** has no analogue in the European languages.
- **ဃ** is pronounced like **ch** in the German ba**ch**, or the Scottish lo**ch**.
- **င** is pronounced like **ng** as heard in si**ng**
- **စ** has a sound intermediate between s and z.
- **ဆ** is the aspirate of **စ**. It has the sound of **ssh**, as heard in the phrase hi**ss** him.
- **ၐ** is pronounced like **sh** as heard in **sh**ell
- **ည** is pronounced like **ny** as heard in ca**ny**on
- **တ** has a sound intermediate between **t** and **d**; say **t** without air coming out
- **ထ** is the aspirate of **တ**. It is pronounced like **ht** as heard in the word ho**t**
- **ဒ** is pronounced like **d** as heard in **d**ay
- **န** is pronounced like **n** as heard in **n**et
- **ပ** has a sound intermediate between **b** and **p**; say **p** without air coming out
- **ဖ** is pronounced like **p** as heard in **p**ool
- **ဘ** is pronounced like **b** in **b**all
- **မ** is pronounced like **m** as heard in **m**all

- ꤃ is pronounced like **y** as heard in back**y**ard
- ꤮ is pronounced like **r** as heard in **r**oom
- ꤤ is pronounced like **l** as heard in schoo**l**
- ꤪ is pronounced like **w** as heard in **w**onderful
- ꤫ is pronounced like **th** as heard in **th**in
- ꤬ is pronounced like **h** as heard in **h**ouse
- ꤀ as a consonant, has no sound of its own; it is a mere stem to which vowel signs are attached. Vowel carrier
- ꤭ is pronounced as a **ɦ** sound.

Vowels can never stand alone and if a word starts with a vowel syllable, use the vowel carrier "꤀" which is silent in order to write words that start with vowel.

◌ꤰ	◌ꤵ	◌ꤱ	◌ꤶ	◌ꤷ	◌ꤸ	◌ꤹ	◌ꤺ	◌ꤻ
ah (a)	ee (i)	uh (ɨ)	u (ɯ)	oo (u)	ae or ay (e)	eh (ɛ)	oh (o)	aw (ɔ)

Figure 2.3 Vowels of Karen Language

Vowels are showed in Figure 2.3 and their explanations are as follows:

- ꤀ – a in quota
- ꤀ꤰ – a in bad
- ꤀ꤵ – i in mean
- ꤀ꤱ – German ö in Göthe
- ꤀ꤶ – German ü in Glück and Korean Hangul character "ㅡ"
- ꤀ꤷ – u in rule, oo in moon
- ꤀ꤸ – a in rate
- ꤀ꤹ – e in met
- ꤀ꤺ – o in note
- ꤀ꤻ – aw in raw

In Sgaw Karen, every syllable consists of a vowel, either alone, or preceded by a single or double consonant. A syllable always ends in a vowel. Every syllable may be pronounced in six different tones of voice, the meaning varying according to the tone in which it is pronounced. Tones and their descriptions are show in figure 2.4.

Tones	Description
◌၏(အၢသံ)	is pronounced with a heavy falling inflection
◌ၵ်(အးသံ)	is pronounced abruptly, at a low pitch
◌း(ဖျၢဉ်ဆံး)	is pronounced abruptly at an ordinary pitch
◌ၣ်(ဟးသံ)	is pronounced with a falling circumflex inflection
◌ၤ(ကွၣ်ဖိ)	is pronounced with a prolonged even tone

Figure 2.4 Tones and their description of Karen Language

When one consonant follows another with no vowel sound intervening, the second consonant is represented by a symbol, which is joined to the character representing the first consonant. Medials are shown in figure 2.5.

◌ၟ (hg)	◌ၢ (y)	◌ၠ (r)	◌ၡ (l)	◌ၥ (w)
ဂ	ယ	ရ	လ	ဝ

Figure 2.5 Medials of Karen Language

The examples of writing the Karen alphabet are:

- ခ + ◌ံ → ခံ, pronounced /ki/
- လ + ◌ံ + ◌း → လံး, pronounced /li/
- က + ◌ၢ +◌ိ → ကိၢ, pronounced /kjo/
- က + ◌ၡ +◌ိ + ◌ၣ် → ကျိၣ်, pronounced /klo/

2.4 Literature Reviews of Machine Translation

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. Also, Myanmar ethnic languages such as Myanmar-Rakhine, Myanmar-Dawei and Kachine-Rawang and so on begins aroung 2015. Most of all previous research is applied to the rule-based and the statistical based approaches.

Firstly, the author presented the first large scale evaluation of statistical machine 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 on 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: syllable segmentation, maximum matching word segmentation with dictionary, CRF word segmentation. The results show that the highest quality machine translation was attained with supervised word segmentation in all of the experiments. Furthermore, for almost all language pairs the HPBSMT approach gave the highest translation quality when measured in terms of both the BLEU and RIBES scores.

Secondly, the author presented the five-state-of-the-art statistical machine translation methods for English and the under resourced languages in both directions. 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 trained these statistical machine translation systems using the ASEAN-MT parallel corpus for each language pair. Moses toolkit are used for training PBSMT, HPBSMT, OSM, S2T and T2S systems. The Barkeley Parser was used for tree annotation of English for S2T and T2S experiments. In their experiments, the phrase-based statistical machine translation method generally gave the highest BLEU scores.

Thirdly, the author investigated SMT performance for Myanmar (Burmese) and Myeik language pair . It is developed a Myanmar-Myeik parallel corpus (around 10K sentences) based on the Myanmar language of ASEAN MT corpus, which is a parallel corpus in the travel domain. There are two types of segmentation were studied: word segmentation and syllable segmentation. Besides, it uses three different statistical machine translation approaches provided by the Moses toolkit: phrase-based, hierarchical phrase-based, and operation sequence model (OSM). The results show that all three statistical machine translation approaches give higher and comparable BLEU and RIBES scores for both Myanmar to Myeik and Myeik to

Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three approaches.

Fourthly, the author contributed the first evaluation of the quality of machine translation between Myanmar (Burmese) and Rakhine (Arakanese). Myanmar-Rakhine parallel corpus (around 18K sentences) are created and the 10 folds cross-validation experiments were carried out using three different statistical machine translation approaches: phrase-based, hierarchical phrase-based, and the operation sequence model (OSM). The results show that all three statistical machine translation approaches give higher and comparable BLEU and RIBES scores for both Myanmar to Rakhine and Rakhine to Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three approaches.

At last, the author studied the key technologies of phrase-based Tibetan-Chinese statistical machine translation, including phrase-translation models and reordering models, and proposes a phrase-based Tibetan-Chinese statistical machine translation prototype system. The method proposed in this paper has better accuracy than Moses, the current mainstream model, in the CWMT 2013 development set, and shows great performance improvement.

CHAPTER 3
PHRASE-BASED STATISTICAL MACHINE TRANSLATION AND THEORY BACKGROUND

In this chapter, what is machine translation and the background theory of Phrase-based Statistical Machine Translation System are explained in detailed.

3.1 Machine Translation

Machine translation is a computerized automated translation between the languages. Machine translation is an area of research that combines ideas and techniques from Linguistics, Computer Science, Artificial Intelligence, Translation theory and Statistics for automating the process of translation from one language to another.

The new field of "machine translation" appears in Warren Weaver's Memorandum on Translation (1949), and the first researcher in the field, Yehosha Bar-Hillel, begins his research at MIT (1951). A Georgetown MT research team follows (1951) with a public demonstration of its system in 1954. MT is touted as a solution to help the U.S. keep tabs on Russian. It's also one of the first non-numerical applications for computers. MT research programs pop up in Japan and Russia (1955), and the first MT conference is held in London (1956). Researchers continue to join the field as the Association for Machine Translation and Computational Linguistics is formed in the U.S. (1962) and the National Academy of Sciences forms a committee (ALPAC) to study MT (1964).

ALPAC's report states MT cannot compete with human translation quality, and suggests funding for MT research should be stopped. But the research continues. MT is also put to work: the French Textile Institute to translate abstracts from and into French, English, German and Spanish (1970); Brigham Young University starts a project to translate Mormon texts by automated translation (1971); and Xerox uses Systran to translate technical manuals (1978). Various MT companies are launched, including Trados (1984), which is the first to develop and market translation memory technology (1989). The first commercial MT system for Russian/English/German-Ukrainian is developed at Kharkov State University (1991).

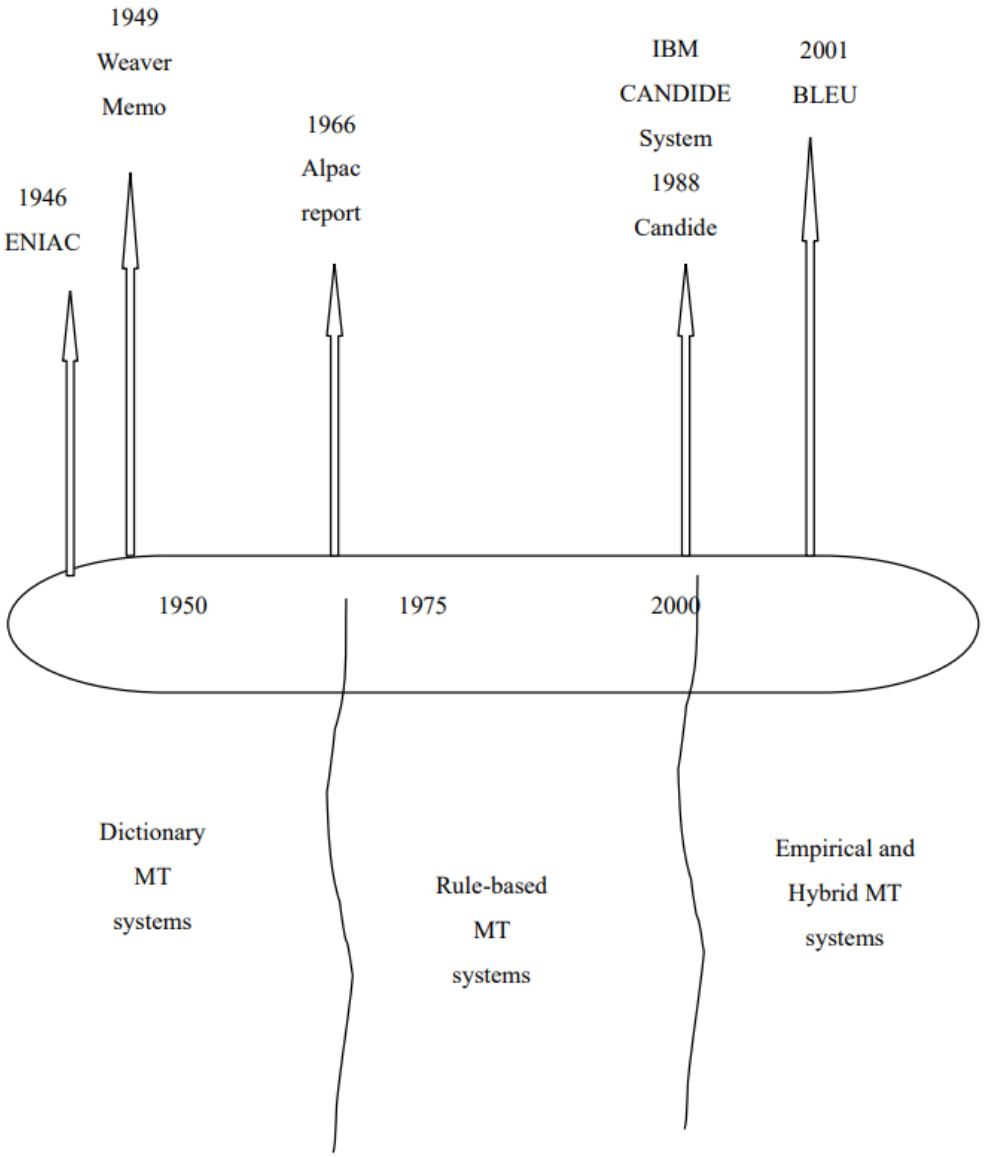

Figure 3.1 History of Machine Translation

MT on the web starts with Systran offering free translation of small texts (1996), followed by AltaVista Babelfish, which racked up 500,000 requests a day (1997). Franz-Josef Och (the future head of Translation Development at Google) wins DARPA's speed MT competition (2003). More innovations during this time include MOSES, the open-source statistical MT engine (2007), a text/SMS translation service for mobiles in Japan (2008), and a mobile phone with built-in speech-to-speech translation functionality for English, Japanese and Chinese (2009). Recently, Google announced that Google Translate translates roughly enough text to fill 1 million books

in one day (2012). Whew! That's a lot, and we didn't cover 90% of the history of machine translation! All the negative talk about MT seems to forget it's an incredible, advanced technology. Its quality is lower than human translation but that doesn't mean it doesn't have good, practical uses—like translating old press releases from 5 years ago.

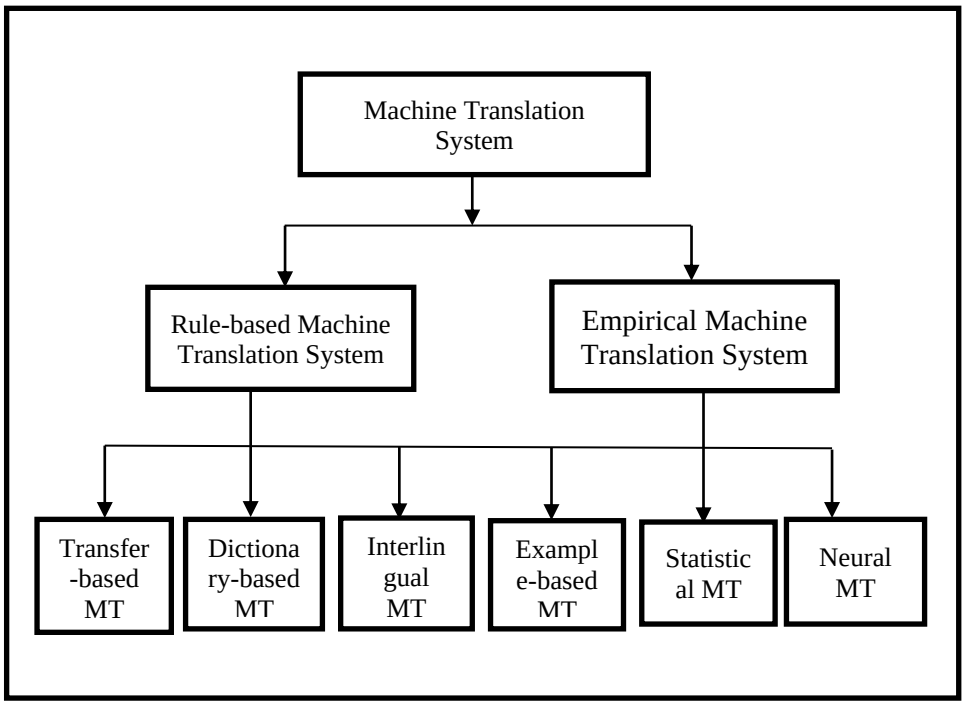

Figure 3.2 Heterotical types of Machine Translation

Recent years have seen significant step advancements in machine translation technology with Google's research on Neural Machine Translation implying an optimistic future for the industry. It has become clear that machine translation is While producing content at low cost and as quickly as possible continues to have an adverse effect on Quality, machine translation offers many translating organizations an edge at achieving the holy grail of translation outcome- a balance of cost, quality, and time-to-market. There are many ongoing attempts to develop MT systems for regional languages using various approaches. The approaches to machine translation are categorized as, Rule Based or Knowledge Driven approaches (RBMT) and Corpus Based or Data-Driven approaches. The RBMT approaches are further classified into Interlingua MT, transfer based MT, and Dictionary based MT. The Corpus Based approaches are classified into Example Based MT, Statistical Machine Translation and Neural Machine Translation.

Among them, the major machine translation techniques are:

1) Rule Based Machine Translation (RBMT)

2) Example Based Machine Translation (EBMT)

3) Statistical Machine Translation (SMT)

4) Neural Machine Translation (NMT)

3.1.1 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 rule. 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, wordby-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.

3.1.2 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. Example of transfer-based approach are showed in Figure 3.3.

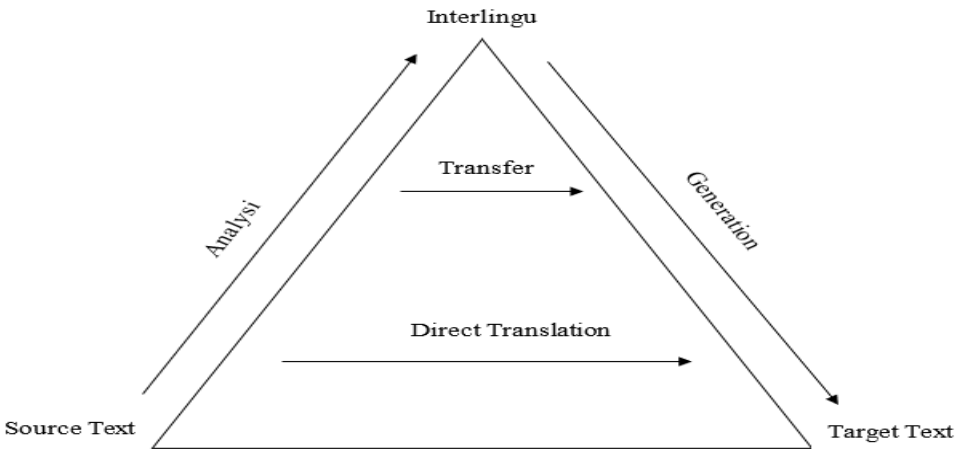

Figure 3.3 Transfer-based approach

3.1.3 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.

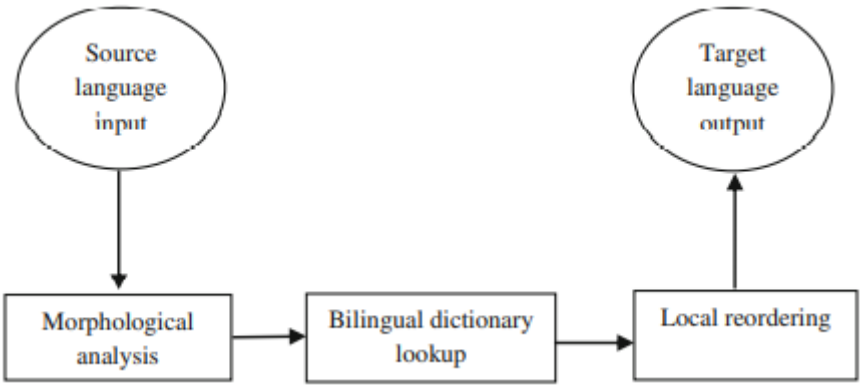

Figure 3.4 Direct MT approach

3.1.4 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. Example of Interlingual MT approach is showed in Figure 3.5

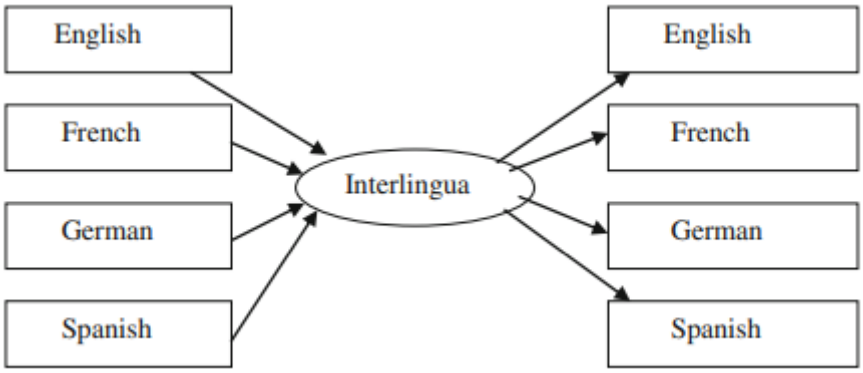

Figure 3.5 Example of Interlingual MT approach

3.1.5 Example-based Machine Translation System

In 1994, Nagao Makoto firstly proposed the example-based machine translation system. Exampled-based approach is often using 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. Example-based MT are showed in Figure 3.6.

3.2 Statistical Machine Translation (SMT) 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, this 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.

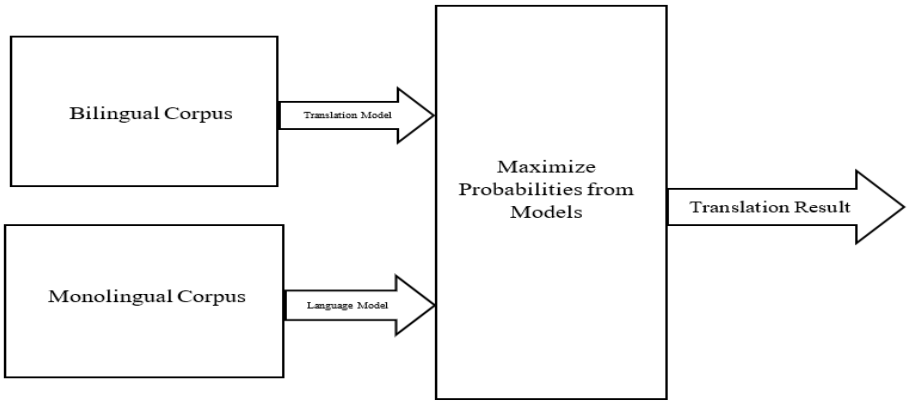

Figure 3.6 Example-based MT approach

Statistical Machine Translation as a research area started in the late 1980s with the Candide project at IBM. IBM's original approach maps individual words to words and allows for deletion and insertion of words. Lately, various researchers have shown better translation quality with the use of phrase translation. Phrase-based MT can be traced back to Och's alignment template model, which can be re-framed as a phrase translation system. Other researchers used augmented their systems with phrase translation, such as Yamada, who use phrase translation in a syntax-based model.

Marcu introduced a joint-probability model for phrase translation. At this point, most competitive statistical machine translation systems use phrase translation, such as the CMU, IBM, ISI, and Google systems, to name just a few. Phrase-based systems came out ahead at a recent international machine translation competition (DARPA TIDES Machine Translation Evaluation 2003-2006 on Chinese-English and Arabic-English). Of course, there are other ways to do machine translation. Most commercial systems use transfer rules and a rich translation lexicon. Machine translation research was focused on transfer-based systems in the 1980s and on knowledge based systems that use an interlingua representation as an intermediate step between input and output in the 1990s.

There are also other ways to do statistical machine translation. There is some effort in building syntax-based models that either use real syntax trees generated by syntactic parsers, or tree transfer methods motivated by syntactic reordering patterns. The phrase-based statistical machine translation model we present here was defined by Koehn et al. (2003).

3.2.1 PBSMT Model

The figure below illustrates the process of phrase-based translation. The input is segmented into a number of sequences of consecutive words (so-called phrases). Each phrase is translated into an English phrase, and English phrases in the output may be reordered.

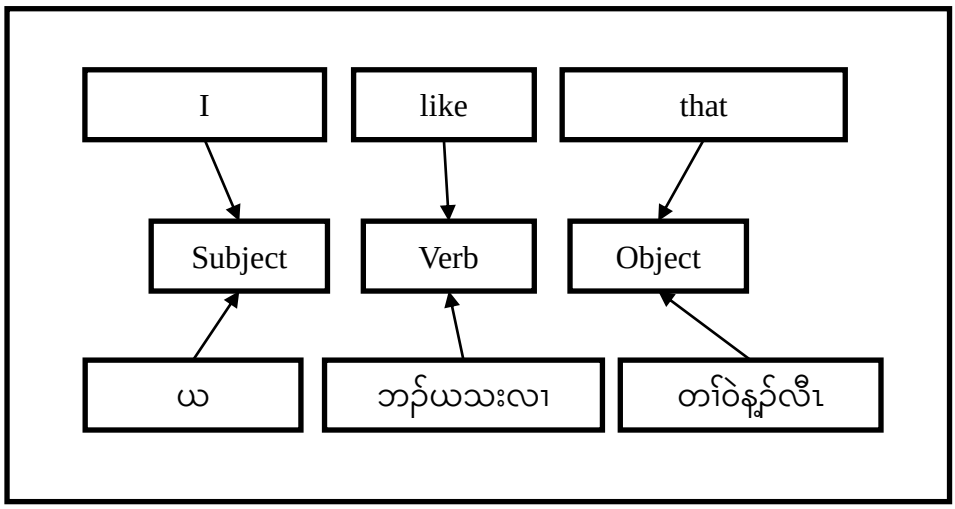

Figure 3.7 Phrase-based Translation

The phrase-based translation system is based on noisy channel model and it apply Bayes rule to reformulate the translation probability for translating a foreign sentence **f** into English **e** as:

$$\text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})\ p(\mathbf{e}) \tag{3.1}$$

where s= source sentence, t=target sentence, p(s|t) is translation model and p(t) is language model. Decompoistion of the translation model:

$$\text{p}(s_1^i|\ t_1^i) = \prod_{i=1}^{i} \emptyset\ (s_i|t_i)\ \text{d}(start_i - end_{i\ \_\ 1} - 1) \tag{3.2}$$

where (si|ti) is the table with phrase translations and their probabilities (phrase tabl) and phrase translation is modeled by a probability distribution. Reordering of the source output phrases is modeled by a relative distoration probability distribution. And strati denotes the start position of the source phrase that was translated into the ith target phrase and endi-1 denotes the end position of the source phrase that was translated into the (I-1)t target phrase.

In order to calibrate the output length, it introduce a factor (called word cost) and language model pLM. Usually, this factor is larger than 1, biasing toward longer output. In summary, the best output sentence ebest given an input sentence f according to our model is:

$$e_{best} = \text{argmax}_t\ \text{p(t|s)} = \text{argmax}_t\ \text{p}_{\text{TM}}\text{(s|t)}\ \text{p}_{\text{LM}}\text{(t)}\ \omega^{\text{length(e)}}$$

$$= \text{argmax}_t \prod_{i=1}^{i} \emptyset\ (s_i|t_i)\ \text{d}(start_i - end_{i\ \_\ 1} - 1)\ \prod_{i=1}^{i} P_{LM}(t)\ \omega^{\text{length(e)}} \tag{3.3}$$

3.2.2 Learning a Phrase Translation System

To get a Phrase Translation Table, its task learns the model from a parallel corpus with three stages. They are:

- Word alignment
- Extraction of phrase pairs
- Scoring of phrase pairs

3.2.2.1 Word Alignment

The most common tool to establish a word alignment is to use the toolkit GIZA++. This toolkit is an implementation of the original IBM models that started statistical machine translation research. However, these models have some serious draw-backs. Most importantly, they only allow at most one English word to be aligned with each foreign word. To resolve this, some transformations are applied.

First, the parallel corpus is aligned bidirectionally, e.g., Spanish to English and English to Spanish. This generates two-word alignments that have to be reconciled. If two alignments interact each other, a high-precision alignment of high-confidence alignment points is gotten. If we take the union of the two alignments, a high-recall alignment with additional alignment points is gotten.

3.2.2.2 Scoring phrase translation

After extraction of phrase pairs, it is need to assign the probability of phrase pairs. To get Phrase pair extraction, it is collected all phrase pairs from the corpus and assign probabilities to phrase translations (Phrase pair scoring). The probability is calculated using relative frequency count:

$$\Phi\ (\bar{t}|\bar{s}) = \frac{count\ (\bar{s},\bar{t})}{\sum_{\bar{t}i} count\ (\bar{s},\bar{t})} \tag{3.4}$$

Example calculation for den Vorschlag learned from the Europarl corpus is like below:

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Figure 3.8 Probability values of Phrase-based Translation

The second part is a distance-based reordering model. $start_i$ is the start position of the source phrase which is the translation of the target phrase i, and end_{i-1} is the last word in the previous phrase. Hence reordering distance is calculated as ($start_i - end_{i-1} - 1$). Equation 3.2 is considered to be the calculation of the translation model for standard phrase-based SMT. However, phrase-based translation system usually uses log-linear model, since it allows using more features instead of just using translation model and language model probabilities as in noisy-channel model. And, Log-linear model are discussed in more details in Section 3.2.2.3.

3.2.2.3 Log-linear models

As we saw before, the standard phrase-based model has two components, the translation model and the language model. However, the translation model actually can be split into two models, the phrase-translation model and the distortion or reordering model. Using the noisy-channel model Equation 3.1 and the reverse translation model *P(s|t)* Equation 3.2, we can get the translation output as follows:

$$t_{best} = arg_t max \prod_{i=1}^{I} \emptyset\ (\bar{s}_i|\bar{t}_i)\ d\ (a_i - b_{i-1} - 1) P(t) \tag{3.5}$$

This equation is actually a multiplication of the phrase translation model, the reordering model and the language model, all getting the same uniform weight which is 1. It would be better to give different weight for each model as in the following equation and then find a way to calculate the best weights.

$$t_{best} = arg_t max \prod_{i=1}^{I} \emptyset\ (\bar{S}_i|\bar{t}_i)^{\lambda_\Phi}\ d\ (a_i - b_{i-1} - 1)\ P(t)^{\lambda_d} \prod_{i=1}^{|t|} P(t_i \backslash t_1 \dots t_{i-1})^{\lambda_{LM}} \tag{3.6}$$

Where $(\lambda_{\Phi}, \lambda_{d,}\lambda_{lm})$ are the weights that can be chosen for the contribution of each model.

If $h_1 = \log \prod_{i=1}^{I} \emptyset(\overline{s_i} \mid \overline{t_i}) = \sum_{i=1}^{I} log\emptyset(\bar{s}_i \mid \bar{t}_i)$,

and $h_2 = \log \prod_{i=1}^{I} d\ (a_i - b_{i-1} - 1) = \sum_{i=1}^{I} log\ d\ (a_i - b_{i-1} - 1)$,

and $h_3 = \log \prod_{i=1}^{|t|} P(t_i|t_1 \dots t_{i-1}) = \sum_{i=1}^{|t|} \log P(t_i|t_1 \dots t_{i-1})$

we will get

$$t_{best} = arg_t \max exp\ (\lambda_{\Phi} h_1 + \lambda_d h_2 + \lambda_{LM} h_3) \tag{3.7}$$

Assume that $n = 3$, $\lambda_1 = \lambda_{\Phi}, \lambda_2 = \lambda_d, \lambda_3 = \lambda_{LM}$, in Equation 3.7 we will get the following:

$$t_{best} = arg_t \max exp \sum_{i=1}^{n} \lambda_i h_i\ (s, t, a, b) \tag{3.8}$$

Which is using the basic form of a log-linear model:

$$p(x) = exp \sum_{i=1}^{n} \lambda_i h_i\ (x) \tag{3.9}$$

Using a log-linear model gives us two advantages over the noisy-channel model. First, different weights to each component model can be given. The second advantage is that one can add more component models, also called feature functions. Usually the weights in a log-linear model are optimized using Minimum Error Rate Training (Mert) to maximize the overall system translation quality using a translation evaluation metric [Och, 2003]. The following are the commonly used feature functions in the state-of-the-art phrase-based systems:

• LM probability

• Bidirectional (i.e. source to target and target to source) phrase translation 27 probabilities.

• Bidirectional lexical probabilities.

• Phrase reordering model.

• Word/phrase penalty.

• Operation Sequence Model features.

3.2.2.3.1 Language model

An LM is an important component in many natural language processing tasks. In SMT, the LM is responsible of the fluency of the translation output as a feature function in the log-linear model in Equation 3.8. The LM is trained on a monolingual corpus in order to be able to estimate the probability of a sequence of words. In the next sections, I will cover the n-gram LM, neural network LM and the evaluation of LMs using perplexity.

3.2.2.3.2 N-gram language models

The joint probability a P($\omega_1,\ldots,\omega_m$) of a sequence of words $\omega_1,\ldots,\omega_m$ is computed using the chain rule as a multiplication of the conditional probabilities of each word ω_i as shown in Equation 3.10

$$P(\omega_1, \ldots, \omega_m) = \prod_{i=1}^{m} P(\omega_i | \omega_1, \ldots, \omega_{i-1}) \tag{3.10}$$

$$P(\omega_1, \ldots, \omega_m) \approx \prod_{i=1}^{m} P(\omega_i | \omega_{i-(n-1)}, \ldots, \omega_{i-1}) \tag{3.11}$$

This is called n-gram LM with order *n*. An n-gram LM estimates the conditional probability for a word given the previous $n - 1$ words. The words' conditional probabilities are multiplied to estimate the joint probability of the whole sentence.

If $n = 1$, the n-gram is called a unigram, if $n = 2$, the n-gram is called a bigram and if $n = 3$ the n-gram is called trigram.

The n-gram conditional probability is estimated using Maximum Likelihood Estimation (MLE) by collecting frequency counts as follows:

$$P\left(\omega_i \middle| \omega_{i-(n-1)}, \ldots, \omega_{i-1}\right) = \frac{count(\omega_{i-(n-1)}, \ldots, \omega_{i-1}, \omega_i)}{count(\omega_{i-(n-1)}, \ldots, \omega_{i-1})} \tag{3.12}$$

One major problem in estimating the n-gram model using MLE is the fact that many possible n-grams are not observed in the training data. This can lead to zero probability (numerator is zero) or an undefined value (denominator is zero). Many smoothing techniques have been proposed in the literature (e.g. add-one smoothing,

Laplace Smoothing, Good-Turing Discounting or KneyserNey smoothing). A good overview of n−gram smoothing techniques is presented in [Chen and Goodman, 1996]. In the following sections, LM interpolation and back-off techniques will be covered.

Interpolation:

Interpolation is a linear composition of lower and higher order n-gram LMs. It is motivated by the idea that lower order n-gram models are less sparse than higher order n-gram models. Each n-gram model contributes with a specific weight λi to the total probability estimation as follows:

$$P_{intr}(\omega_n|\omega_1,\dots,\omega_{n-1}) = \lambda_1 p_1(\omega_n) + \lambda_2 p_2(\omega_n|\omega_{n-1}) + \cdots + \lambda_n p_n(\omega_n|\omega_1,\dots,\omega_{n-1}) \quad (3.13)$$

where P_i is an i-gram language model and $0 \le \lambda i \le 1$. $\sum_i P\ i\ \lambda i = 1$ to ensure that P_{intr} is a proper probability distribution. One way to find the best weights is using the EM algorithm on a held-out set. It converges on locally optimal weights.

Back-off LM:

Like interpolation, back-off is used to address the problem of unseen n-grams. The difference is that in a back-off model, we only use the higher order n-gram probability if it is available, otherwise we back off to a lower order LM to get the probability as follows:

$$P_n^{BO}\left(\omega_i\middle|\omega_{i-(n-1)},\dots,\omega_{i-1}\right) = \begin{cases} d_n\left(\omega_{i-(n-1)},\dots,\omega_{i-1}\right)P_n\left(\omega_i\middle|\omega_{i-(n-1)},\dots,\omega_{i-1}\right) \\ \text{if } count_n(\omega_{i-(n-1)},\dots,\omega_{i-1}) > 0 \\ \alpha_n\left(\omega_{i-(n-1)},\dots,\omega_{i-1}\right)P_n^{BO}\left(\omega_i\middle|\omega_{i-(n+2)},\dots,\omega_{i-1}\right) \\ \text{otherwise} \end{cases} \quad (3.14)$$

A discounting function d is used to make sure that all probabilities add up to 1. The lower order probabilities are multiplied by a discounting factor α between 0 and 1 in order to ensure that only the probability mass set aside by the discounting step is distributed to the lower-order n-grams. More details on back off LM can be found in [Katz, 1987]

LM Evaluation and perplexity:

We can measure the LM quality using two ways. The first way is an end-toend evaluation. In this method, the performance of different LMs is evaluated in the framework of the full system (i.e. a MT system in our case). This is the best evaluation but it is more expensive. The second way is to calculate an independent LM quality measure on an development set. The standard metric is the perplexity (PP). Perplexity is based on the concept of entropy *H*(p), which measures uncertainty in a probability distribution as defined below:

$$H(\mathrm{p}) = -\sum_x p(x) \log_2 p(x) \tag{3.15}$$

The perplexity is a simple transformation of cross-entropy. Given an evaluation set (ω_1, ω_2 ..., ω_m), the language model P_{LM}, the cross-entropy $H(P_{LM})$ is defined as follows:

$$H(P_{LM}) = -\frac{1}{m}\sum_{i=1}^{m} \log_2 P_{LM}(\omega_i|\omega_1, \dots, \omega_{i-1}) \tag{3.16}$$

and the perplexity is defined as follows:

$$PP = 2^{H(P_{LM})} \tag{3.17}$$

The PP is a positive number. The smaller the value, the better the language model is. It is important to note that the PP of two LMs are only directly comparable if they use the same vocabulary.

3.2.2.4 Decoding in SMT

The goal of the decoder is to find the best target sentence that maximize the translation probability *P(t|s)* as expressed in the log-linear Equation 3.8. Several decoders are publicly available like Jane [Freitag et al., 2014], Cdec [Dyer et al., 2010] and Moses [Koehn et al., 2007b]. Moses is an open source SMT toolkit and implements a beam search decoder.

SMT decoding is NP-complete [Knight, 1999], however heuristic techniques work well. Decoding for word-based SMT had a higher complexity because of the possible reordering of individual words compared to phrase-based SMT which use

larger translation units (i.e. phrases). The decoding algorithm for word-based SMT could be implemented using optimal A* search [Och et al., 2001], integer programming [Germann et al., 2001] or greedy search algorithms [Wang and Waibel, 1998].

In phrase-based SMT, the most commonly used decoding algorithm is beam-search stack decoding, other algorithms like Beam search based on converge stacks, A* search, Greedy Hill-Climbing decoding and Finite state transducer decoding which have been proposed in the literature.

In beam search decoding, the decoder starts by looking for all possible translations in the phrase table. This includes the possible translations of all possible phrases of a given source sentence as shown in the upper part of Figure 3.6.

Decoding of a source sentence starts with an initial empty hypothesis, then the translation output hypotheses are constructed from left to right. The hypotheses are expanded by picking the available translation options as shown in the lower part of Figure 3.6. The decoder then updates the source translation coverage vector for these new expanded hypotheses. It incrementally computes the translation probability of each of them. Several techniques are used to limit the exponential explosion of the search space. These techniques include hypotheses recombination (i.e. combine similar hypotheses which cover the same source translation but have different scores), pruning out bad hypotheses with worse scores from the hypotheses stack, estimating hypotheses future cost to prevent pruning out good future hypotheses. The expansion process of each remaining hypothesis continues until all source words are covered. These hypotheses are called completed hypotheses. If there are no more incompleted hypotheses, the decoder selects the hypothesis with the highest probability from the completed hypotheses as the most likely translation t_{best}.

3.2.2.5 Minimum error rate training

The log-linear model gives us two advantages over the noisy-channel model: the first one is that we can give different weights to different component models. The second advantage is the possibility to easily add new components (also called feature functions). Usually the weights λi in the log-linear model (Equation 3.8) are optimized using the Mert algorithm proposed by [Och, 2003]. Mert is an efficient

supervised algorithm used to maximize the translation quality on a held-out set as measured by an automatic metric.

Mert works as follows:

• Initialization: initialize λi randomly or based on some heuristics.

• Translation: n-best translation of the development set with current λi

• Comparison: compare the objective score (such as Bleu) of the n-best translation with previous run

• Re-estimation: Re-estimate the weights λi

• Iterate: Iterate until weights have converge

Mert does not scale well to large number of feature functions [Ittycheriah et al., 2007], so other tuning algorithms have been proposed to overcome this issue like MIRA tuning algorithm [Chiang, 2012; Hasler et al., 2011] and the pairwise ranked optimization (PRO) [Hopkins and May, 2011].

CHAPTER 4
IMPLEMENTATION

This chapter describes about system implementation of phrase-bases statistical machine translation system between Karen and English Languages. And the dataset, preprocessing steps and Karen-English PBSMT are explained. Moreover, the results and evaluation details of the systems are described.

4.1 Experimental Setting

This section describes the dataset, preprocessing steps and the Phrase-based Statistical machine translation models.

4.1.1 Dataset

Karen language is regarded as a low resource language, so there are no many Karen-English parallel sentences. Therefore, Karen-English parallel corpus is created from Karen-English published books via internet and other sources. There are over 11K parallel sentences in total. The corpus is randomly divided into training data, development data and test data. This corpus is general corpus covering difference domains. Table 4.1 shows data statistics used for the experiments.

Table 4.1 Data Statistics of the Corpus

	Parallel Sentences
Total Number of Sentences	11500 (over 11k)
Training File	10000
Development File	1000
Testing File	500

There are some examples of Karen and English.

Kr: စံးဘျုး နၤ ဒိၣ်မးလီၤ

En : Thank you a lot.

Kr: ယမိၢ်နှၣ် မ့ၢ်ဟ်မုၣ်လၢ အဃံၤလၢတဂၤလီၤ

En : My mother is beautiful lady.

Kr: ဟ်ခွါဖိသ့ၣ်တဖၣ်နၤ် လိာ်ကွဲလ၊ သ့ၣ်မု၊် အဖီလာ်နၤ်လီ၊

En : The boys are playing under the tree.

Kr: မု၊်နၤ် ဟဲထီၣ်လ၊ မု၊်ထီၣ်တခီလီ၊

En : The sun rises in the east.

Kr: ဘၣ်မနု၊အဃိ အဝဲ အိၣ် ဖဲအံ၊လဲၣ်

En : Why is he here.?

4.1.2 Preprocessing Step

Like Myanmar Language, Karen is an unsegmented language and there is no clear definition of word boundaries. It does not contain white space to delimit the words like English. Tokenization, called word segmentation, is not a trivial task for Karen 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 Karen language, the proper text segmentation is lack. Therefore, Karen sentences are manually segmented in the system. For English language, Moses's tokenization script is used to segment English sentence. Figure 4.1 and 4.2 show the segmentation process for both languages. In the preprocessing step, Moses's clean-corpus script is also used for both languages.

The clean-corpus script is small script that cleans up a parallel corpus, so it works well with the training script. It performs the following steps:

- removes empty lines
- removes redundant space characters
- drops lines (and their corresponding lines), that are empty, too short, too long or violate the 9-1 sentence ratio limit of GIZA++.

Before Segmentation - ယဒ့ကထါကန့်၎ါ
After Segmentation - ယဒ့ ကထါ ကန့်၎ါ

Figure 4.1 Example segmentation of Karen sentence

Before Segmentation - Let’s hope for the best.
After Segmentation - Let ’s hope for the best.

Figure 4.2 Example segmentation of English sentence

4.2 Model

Karen-English Phrase-based Statistical Machine Translation system is implemented by using Moses's Statistical Machine Translation System. Moses is a statistical machine translation system that allows automatically train translation models for any language pair.

The word segmented source language was aligned with the word segmented target language using GIZA++. The alignment was symmetrized by grow-diag-final and heuristic. GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). GIZA++ includes a lot of additional features. The extensions of GIZA++ were designed and written by Franz Josef Och. GIZA++ is used to train IBM Models 1-5 and an HMM word alignment model. Alignment models depending on word classes. Example are shown in Figure 4.3.

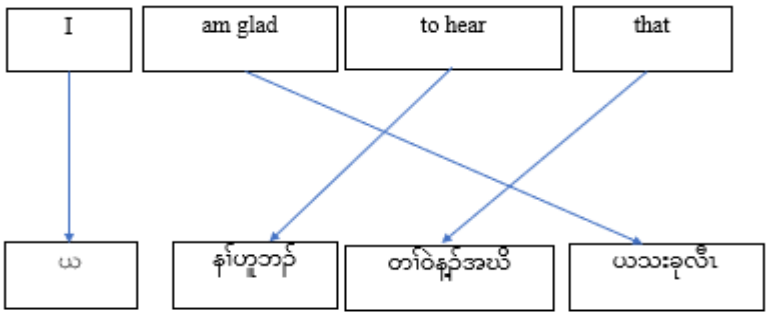

Figure 4.3 Example of Karen and English alignment

The lexicalized reordering model was trained with the msd-bidirectional-fe option. Lexicalized reordering models play a crucial role in phrase-based translation systems. They are usually learned from the word-aligned bilingual corpus by examining the reordering relations of adjacent phrases. The system experiments with

KenLM and SRILM for training the 2-gram, 3-gram and 5-gram language models. Therefore. there are six model for each direction, namely, 2gramKenLM, 2gramSRILM, 3gramKenLM, 3gramSRILM, 5gramKenLM and 5gramSRILM. Minimum error rate training (MERT) was used to tune the decoder parameters and the decoding was done using the Moses decoder.

4.3 Experimental Results and Discussion

For the evaluation result of the translation output, the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) is used. In Karen to English phrase-based statistical machine translation, it is clear that the evaluation results of 5gramKenLM model are much better than those of the others. Table 4.2 shows the evaluation results between Karen and English phrase-based statistical machine translation models.

It is observed that 5gramKenLM model is effective when translating into English, obtaining a BLEU score of 22.50. In English to Karen phrase-based statistical machine translation system, 3gramKenLM model is the best. It is compared with other models and these are shown with Figure 4.4 and 4.5 in BLEU for both directions.

Table 4.2 Evaluation Results (BLEU) of Karen-English SMT system

Model	Kr-En (BLEU)	En-Kr (BLEU)
2gramKenLM	21.96	19.77
2gramSRILM	21.47	18.58
3gramKenLM	22.18	20.12
3gramSRILM	21.53	19.68
5gramKenLM	22.50	20.05
5gramSRILM	21.64	19.45

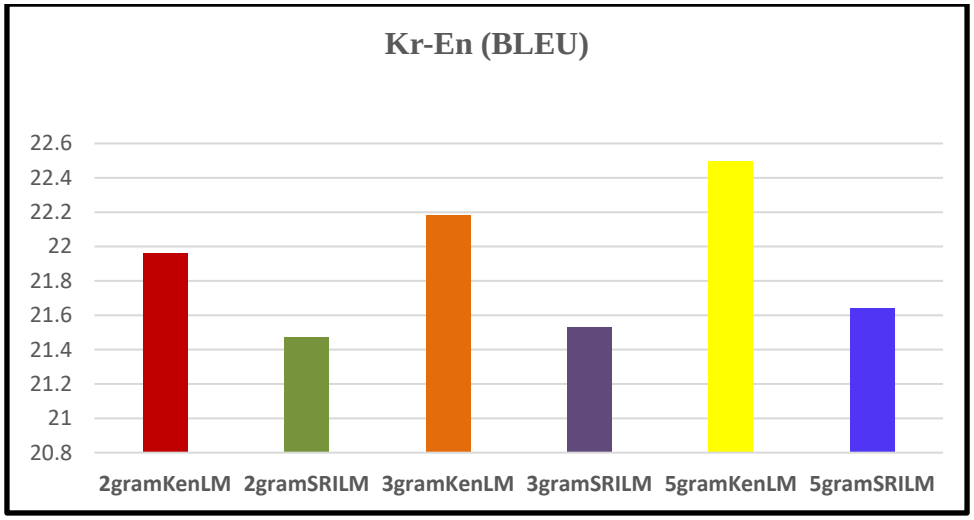

Figure 4.4 Comparison of Karen to English Evaluation Results

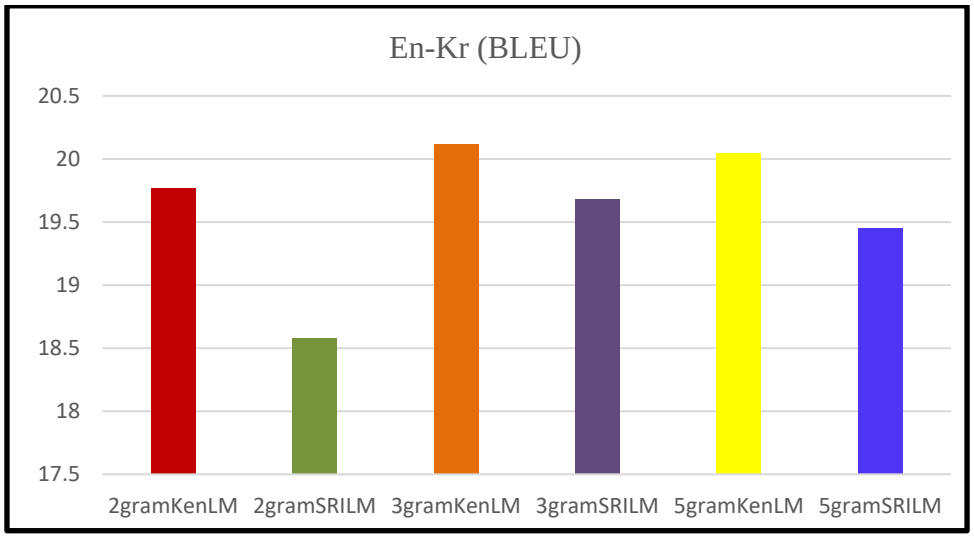

Figure 4.5 Comparison of English to Karen Evaluation Results

4.4 Example of Karen-English PBSMT on Terminal

This section describes the example translation of Karen to English phrase-based statistical machine translation system on terminal. The following commands is to translate Karen language to English language in the system.

```
$ ~/tools/mosesdecoder/bin/moses -f ~/tools/work/KrEn/filtered/moses.ini
```

```
$ ~/tools/mosesdecoder/bin/moses -f ~/tools/work/KrEn/filtered/moses.ini < in.txt > out.txt
```

After running the commands on the terminal, the user can enter the input Karen sentence. Example are shown in Figure 4.6.

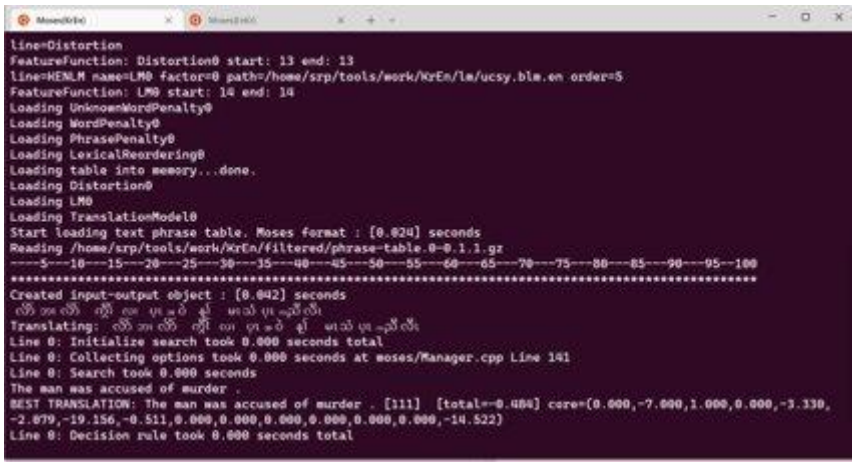

Figure 4.6 Example translation of Karen to English language

4.5 Example of English-Karen PBSMT on Terminal

This section describes the example translation of English to Karen phrase-based statistical machine translation system on terminal. The following commands is to translate English language to Karen language in the system.

```
$ ~/tools/mosesdecoder/bin/moses -f ~/tools/work/EnKr/filtered/moses.ini
```

```
$ ~/tools/mosesdecoder/bin/moses -f ~/tools/work/EnKr/filtered/moses.ini < in.txt > out.txt
```

After running the commands on the terminal, the user can enter the input English sentence. Example are shown in Figure 4.7.

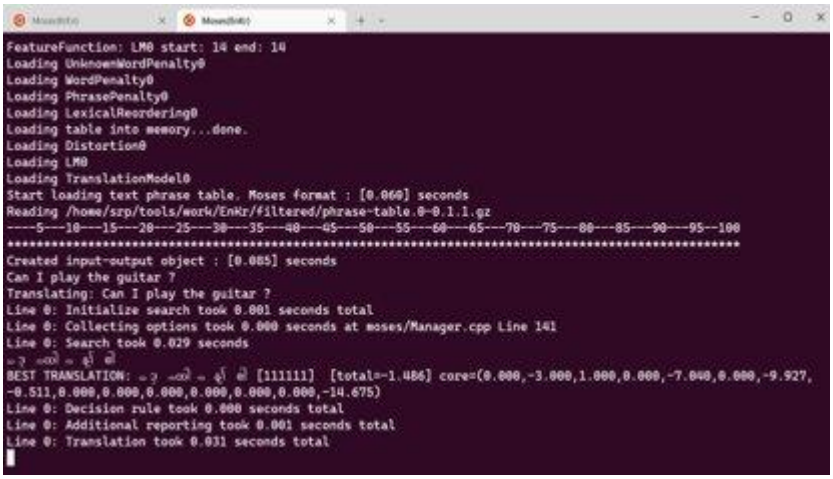

Figure 4.7 Example translation of English to Karen language

CHAPTER 5
CONCLUSION AND FURTHER EXTENSIONS

This thesis intends to develop to translate from Karen to English and English to Karen language by using PBSMT model. The various components, environments of this system are investigated and their contributions to the overall performance of the system are analyzed. In this chapter, the main contents of the thesis are concluded, advantages and limitations of the system, and future work are suggested.

5.1 Conclusion

As the conclusion, the proposed system firstly presents the implementation of translation of Karen-English and English-Karen pairs by using the Moses toolkit. Secondly, the word segmented of source language is aligned with segmented target language by using GIZA++. Thirdly, the lexicalized reordering model was trained with the msd-bidirectional-fe option. And then, for training with 2-gram, 3-gram and 5-gram language models for Karen to English and English to Karen language pairs, KenLM and SRILM are used. By using Moses decoder, the decoded parameter and the decoding process are done. And the performance of the system is measured in terms of BLEU scores and compared them. Finally, the system is proved that KenLM with 3-gram language model is the best result for English to Karen PBSMT model.

5.2 Advantages

The proposed system serves user-friendly, high-performance, and helps the learners and researchers to almost translate from source language they wanted to target languages for short time without human translators. So, the system saves costs: human translators, time consuming, and quick response translation. As system's proof, results with BLEU scores are evaluated by using two language models: KenLM and SRILM, results are compared, and the best result is produced by the system.

5.3 Further Extensions

In the proposed system, Karen-English parallel corpus is a general domain, therefore, translations focus mainly on informal translation. As a further extension,

the experiments will be done focusing on the identified domain. The quality of machine translation performance depends on the size of the parallel corpus. Therefore, more parallel sentences will be collected. In the future, we hope to extend our system to cover Karen and other languages with even more features.

REFERENCE

[1] Ye Kyaw 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. https://www.google.com/search?sxsrf=ALiCzsZJLUVUtA_RpWpKwFbhVDHYjXa0Cg:1664304045144&source=univ&tbm

[2] Win Pa Pa, Ye Kyaw 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. https://www.google.com/search?q=%22A+Study+of+Statistical+Machine+Translation+Methods+for+Under+Resources+Languages%22&client

[3] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, Thepchai Supnithi, "Statistical Machine Translation between Myanmar and Myeik", Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020) Yangon (Rangoon), Myanmar (Burma), 2020, pp. 36-45.Statistical Machine Translation between Myanmar and Myeik - WCSE 2020 Spring – WCSE. https://docplayer.net/228052898-Statistical-machine-translation-between-myanmar-burmese-and-rakhine-arakanese.html

[4] Thazin Myint Oo, Ye Kyaw Thu,K hin Mar Soe, "Statistical Machine Translation between Myanmar (Burmese) and Rakhine", https://docplayer.net/228052898-Statistical-machine-translation-between-myanmar-burmese-and-rakhine-arakanese.html

[5] Wei yang, Hanfei Shen, Y Ves Lepage, "Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese- Japanese Machine Translation", Journal of Information Processing, Vol.25, pp. 88-99, 2017. (PDF) Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese-Japanese Machine Translation (researchgate.net)

[6] Dojun Park, Youngjin Jang, Harksoo Kim, “Korean-English Machine Translation with Multiple Tokenization Strategy”, English translation of the original Korean thesis in KCC2021 Undergraduate/Junior Thesis Competition, 2021.
https://www.researchgate.net/publication/221487850_Improving_phrase based_Korean-English_statistical_machine_translation

[7] Aye Thida Win, “Phrase Reordering Translation System in Myanmar-English”, May 2011.
Phrase Reordering Translation System in Myanmar-English - CORE

[8] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, “Statistical Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)”.
https://onlinesource.ucsy.edu.mm/bitstream/handle/123456789/359/304-311.pdf

[9] Zar Zar Linn, Ye Kyaw Thu, Pushpa B. Patil, “Statistical Machine Translation between Myanmar (Burmese) and Kayah”, Journal of Intelligent Informatics and Smart Technology, Vol. 4, April 2020, pp.62-68.
http://docslib.org/doc/4010071/statistical-machine-translation-between-myanmar-burmese-and-kayah

[10] Zun Hlaing Moe, Thida San, Ei Thandar Phyu, Hlaing Myat Nwe, Hnin Aye Thant, Naw Naw, Htet Ne Oo, Thepchai Supnithi and Ye Kyaw Thu, “Myanmar Text (Burmese) and Braille (Mu-Thit) Machine Translation Applying IBM Model 1 and 2”, Journal of Intelligent Informatics and Smart Technology, Vol. 5, April 2021, pp. 18-26.
https://jiist.aiat.or.th/assets/uploads/16195380940305LtWrJIIST-44-FinalVersion.pdf

[11] 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, 2019.
Yi Mon Shwe Sin.pdf (ucsy.edu.mm)

[12] Dekang Lin, “Automatic Identification of Non-compositional Phrases”, pp. 317-324.
https://arxiv.org/ftp/arxiv/papers/1402/1402.0563.pdf

[13] Thet Thet Zin [a] , Khin Mar Soe [b] , and Ni Lar Thein [a] , “Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System”, 25th Pacific Asia Conference on Language, Information and Computation, pp.130–139.
https://link.springer.com/chapter/10.1007/978-3-642-00382-0_26

[14] Marta R. Costa-juss`, Carlos A. Henr´ıquez Q., Rafael E. Banchs, “Evaluating Indirect Strategies for Chinese–Spanish Statistical Machine Translation”, Journal of Artificial Intelligence Research 45 (2012), pp.761-78.
(PDF) Evaluating Indirect Strategies for Chinese-Spanish Statistical Machine Translation (researchgate.net)

[15] Jia Xu and Geliang Chen*, “Phrase Based Language Model for Statistical Machine Translation”, Working Paper — July 1, 2021.
https://arxiv.org/abs/1501.04324

[16] icola Bertoldi, Madalina Barbaiani†, Marcello Federico, Roldano Cattoni, “Phrase-Based Statistical Machine Translation with Pivot Languages”, Proceedings of IWSLT 2008, Hawaii - U.S.A.
https://www.researchgate.net/publication/221562550_Phrase-Based_Statistical_Machine_Translation

[17] Christophe SERVAN Simon PET I TRENAU D, “Calculation of phrase probabilities for Statistical Machine Translation by using belief functions”, Proceedings of COLING 2012: Posters, pages 1101–1110, COLING 2012, Mumbai, December 2012.
https://aclanthology.org/C12-2107.pdf

[18] Michael Collins, “Phrase-Based Translation Models”, April 10, 2013.
http://www.cs.columbia.edu/~mcollins/pb.pdf

[19] Thet Thet Zin, Khin Mar Soe, Ni Lar 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.
(PDF) Improving Phrase-based Statistical Myanmar to English Machine Translation with Morphological Analysis (researchgate.net)

[20] Sreelekha S, “Statistical Vs Rule Based Machine Translation; A Case Study on India Language Perspective”, August 2017.

https://www.researchgate.net/publication/319135296_Statistical_Vs_Rule_Based_Machine_Translation_A_Case_Study_on_Indian_Language_Perspective

[21] Naroa Zubillaga, Zuriñe Sanz and Ibon Uribarri, "Building a trilingual parallel corpus to to analyse literary translations fromGerman into Basque", Naroa Zubillaga, Zuriñe Sanz & Ibon Uribarri. 2015. https://langsci-press.org/catalog/view/76/70/280-1

[22] Richard Zens, Franz Josef Och, and Hermann Ne, "Phrase-Based Statistical Machine Translation", All content following this page was uploaded by Hermann Ney on 19 August 2014. https://www.researchgate.net/publication/2887790_Improvements_in_Phrase-Based_Statistical_Machine_Translation

[23] Daniel Marcu, William Wong, "A Phrase-Based, Joint Probability Model for Statistical Machine Translation", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 133-139. Association for Computational Linguistics. https://aclanthology.org/W02-1018.pdf

[24] Zin Thet Thet; Soe, Khin Mar; Thein, Ni Lar, " Myanmar Phrase Translation for Myanmar-English Machine Translation System", ICCA 2012. http:// onlineresource .ucsy.edu.mm/handle/123456789/2263

[25] Aye Thida Nway Nway Han Sheinn Thawtar Oo, "Statistical Machine Translation Using 5-grams Word Segmentation in Decoding", 32nd Pacific Asia Conference on Language, Information and Computation. The 5th Workshop on Asian Translation Hong Kong, 1-3December 2018. https://aclanthology.org/Y18-3009/

[26] Rrun Babhulgaonkar, Shefali Sonavane, "Empirical Analysis of Phrase-Based Statistical Machine Translation System for English to Hindi Language", Vietnam Journal of Computer Science, 2022. Empirical Analysis of Phrase-Based Statistical Machine Translation System for English to Hindi Language | Vietnam Journal of Computer Science (worldscientific.com)

[27] Honey Htun, Ye Kyaw Thu, Nyein Nyein Oo, Thepchai Supnithi, "English-Myanmar (Burmese) Phrase-Based SMT with One-to-One and One-to-Multiple Translations Corpora", 2020.

(PDF) English-Myanmar (Burmese) Phrase-Based SMT with One-to-One and One-to-Multiple Translations Corpora | Honey Htun - Academia.edu

Statistical Machine Translation between Myanmar and Myeik - WCSE 2020 Spring - WCSE

[28] Krzysztof Wok, Krzysztof Marasek, "Enhanced Bilingual Evaluation Understudy", 2014, Lecture Notes on Information Theory. https://www.academia.edu/25522151/Enhanced_Bilingual_Evaluation_Understudy

[29] Liling Tan, Jon Dehdari, Josef van Genabith, "An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation", Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 74–81, Kyoto, Japan, 16th October 2015. 2015 Copyright is held by the author(s). https://aircconline.com/ijnlc/V8N2/8219ijnlc01.pdf

[30] Richard Zens and Hermann Ney, "Improvements in Phrase-Based Statistical Machine Translation", June, 2004. https://www.researchgate.net/publication/2887790_Improvements_in_Phrase-Based_Statistical_Machine_Translation