MAXIMUM MATCHING AND RULE-BASED TECHNIQUES FOR KHMER WORD SEGMENTATION

PAKRIGNA LONG

A THESIS REPORT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ENGINEERING IN COMPUTING IN ENGINEERING SYSTEMS
INTERNATIONAL COLLEGE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2017

KMITL-2017-IC-M-11-08

MAXIMUM MATCHING AND RULE-BASED TECHNIQUES FOR KHMER WORD SEGMENTATION

PAKRIGNA LONG

A THESIS REPORT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ENGINEERING IN COMPUTING IN ENGINEERING SYSTEMS
INTERNATIONAL COLLEGE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2017
KMITL-2017-IC-M-11-08

THESIS TITLE	Maximum Matching and Rule-based Techniques For Khmer Word Segmentation
STUDENT NAME	Mr. Pakrigna Long
STUDENT ID	59610056
DEGREE	Master of Engineering
PROGRAME	Computing in Engineering Systems (International Program)
ADVISOR	Assoc.Prof.Dr.Veera Boonjing

ABSTRACT

In most Asian languages, boundaries of words inside a sentence or clause are usually omitted. These matters of unsegmented words have been causing huge problems for information retrieval, machine translation, text-to-speech, and many other natural language processing (NLP). Thus, word separation is the essential assignment to be done in NLP research. There are many studies involved with word segmentation of Khmer and Asian languages have been investigated. In Khmer Word Segmentation, several approaches related to segmenting words based on dictionary have been studied. There are only a few researches about solving unknown word problem. This matter is a quite challenge task in word separation. In this research, Maximum Matching algorithm (MMA) together with Rule-based technique has been proposed. First, MMA and a Khmer manual corpus were used to make word boundaries in each sentence. Then the unknown words were then defined and solved by using 21 grammar rules created. We tested the segmentation with 2018 sentences from agriculture, magazine, newspaper, technology, health and history. With Maximum Matching alone, we could achieve the accuracy of 88.10% and along with Rule-based, the accuracy increased to 92.02%.

ACKNOWLEDGEMENT

Without the contribution of many people, this thesis would not have been existed. It owes the existence to the supports and inspirations from a lot of people. First of all, I would like to thank to my parents and my relatives for always be there providing me with the support needed in order to continually push myself to succeed. Without their love and support, I would not be here today! Secondly, I would like to thank to my advisor **Assoc.Prof.Dr. Veera Boonjing** who has helped and guided me since the second semester. Without him, it could not be possible for me to complete thesis on time. Lastly, I would like to thank to all the lecturers, staffs and classmates in KMITL for teaching, encouraging and recommending me. I am so grateful for everything that you all have given me, and I could not have done it without you.

Love always!

Pakrigna Long

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Overview of Khmer Language

Khmer is one of the oldest language in Southeast Asia. It is the official and national language used by the people in Cambodia. It is also the language spoken by some individuals who live in the eastern of Thailand and southern of Vietnam. Most of the Khmer characters and words are taken from Pali-Sankrit language which shares the similarity with Thai and Lao. Based on the book written by Chhorn, C. (2002), there are 33 consonants, 23 dependent and 13 independent vowels, 10 digits of numeric, and approximately 27 diacritics/special characters as shown in **Table 1.1.**

Table 1.1 Types of Khmer characters

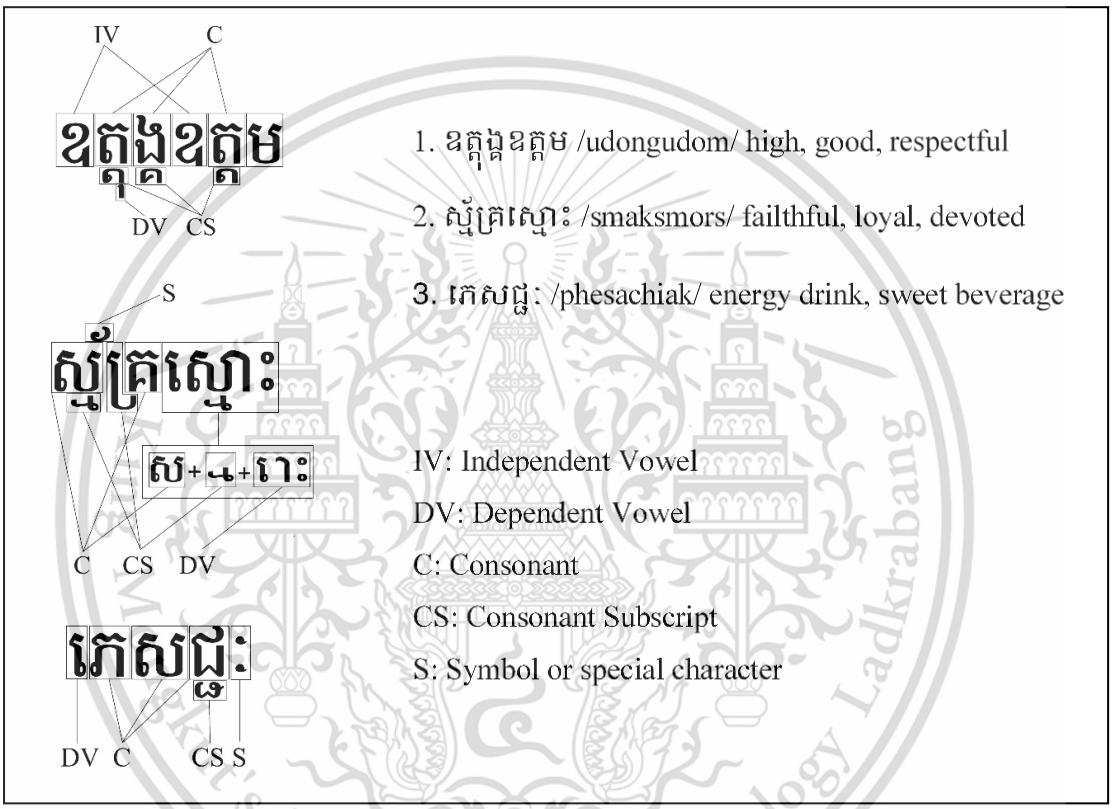
Consonants/ its subscripts	ក ខ គ ឃ ង ច ឆ ជ ឈ ញ ដ ឋ ឌ ឍ ណ ត ថ ទ ធ ន ប ផ ព ភ ម យ រ ល វ ស ហ ឡ អ
Dependent Vowels	◌ា ◌ិ ◌ី ◌ឹ ◌ឺ ◌ុ ◌ូ ◌ួ ◌ើ ◌ឿ ◌ៀ ◌េ ◌ែ ◌ៃ ◌ោ ◌ៅ ◌ុំ ◌ំ ◌ាំ ◌ះ ◌ុះ ◌េះ ◌ោះ
Independent Vowel	ឥ ឦ ឧ ឩ ឪ ឫ ឬ ឭ ឮ ឯ ឰ ឱ,ឲ ឳ
Numeric	០ ១ ២ ៣ ៤ ៥ ៦ ៧ ៨ ៩
Diacritics/special characters	◌ៈ ◌៉ ◌៊ ◌់ ◌៌ ◌៍ ◌៎ ◌៏ ◌័ ◌៑ ◌្ ◌៓ ។ ៗ ៘ () ◌៝ ? ! ៖ ៛ ។ល។ . , ◌ៗ " /

Source: Chhorn, C. (2002). Khmer Grammar for General Studies. Phnom Penh, Cambodia.

In Khmer language, a word is made up of one or more syllable(s). A syllable is created by using a combination of the characters. To produce a syllable, at least one starting consonant or independent is required. The syllable cannot start with a

dependent vowel or special character. In order to formulate a Khmer word, it can be complicated since the position of the characters after the first one can be on the left, right, above, middle-above, below, and double-below as shown in **Figure 1.1**. The consonant subscripts are often used and always in the below position.

Figure 1.1 The structure of Khmer words



Source: Chhorn, C. (2002). Khmer Grammar for General Studies. Phnom Penh, Cambodia.

1.2 Characteristics of Khmer Word Taken from Pali-Sankrit and Its Borivasap

Because of the influence of Hinduism and Buddhism, similar to the neighboring country languages like Thai and Lao, most of the Khmer words are borrowed from Pali-Sankrit, which is the mixture language of Pali and Sankrit. According to the book of Chhorn, C. (2002), we can recognize the Pali-Sankrit taken words based on their characteristics or rules as following:

1. No symbols such as "◌ំ ៗ" at the end of word: ជន ទារក មរតក ចៅក្រម នានា ភព ...
2. Symbols such as "◌ៈ ◌ៃ ◌ោ" at the end of word: ប្រយោជន៍ ស្រាពណ៌ អេសដ្ឋៈ ស្នេហ៍ ...
3. The word has independent vowel: ឯកតោ ឥដ្ឋា ឯកការ ឯកសារ ឫកពា ឫស្សី ឧទាន ...
4. The word has "◌ៈ" symbol: អាល័យ វិស័យ អនាម័យ និស្ស័យ ឧស្ម័ន ន័យ វ័យ រហ័ស ...
5. The first consonant of word has the sounds "អាក់ អ៊ាក់": កសិករ ករី គណិត ធនាគារ ...
6. The character "ណ" but sound "ន": ធរណី ប្រពៃណី មាណព បវេណី វេរមណី វិស្ណុ រមណី ...
7. The character "ប" but sound "ប៉": បរិមាណ បរិវេណ បាតុកម្ម បាតុភូត បច្ចេកទេស បាតុភូត ...
8. The character "ត" but sound "ដ": តំបន់ តារា បិតា តោ តំណាល ចេតិយ សេនាបតី សីតា ...
9. The character "◌ោ" but sound "◌ែ◌ោ": ខេត្ត ប្រេត ស្នេហ៍ កេណ្ឌ សង្ខេប សេដ្ឋី ព្រះសុមេរុ កេតុ ...
10. The character "◌ិ" but sound "◌ី": បិតា បិសាច ខន្តី ...
11. The character "◌ិ" but sound "◌េ◌": ស្ថិរភាព ចិរភាព សិរ ព្រះសិរ ...
12. The character "◌ំ" but sound "◌ង": សំវាស សំស្ក្រឹត សំយោគ សំវេគ សំណោហៈ សំសារៈ ...
13. The word has "ខ គ ឃ ច ទ ធ ភ" at the end: មុខ សុខ នាគ មេឃ ប្រមាថ វិតាទ ពិសោធ ...
14. The word has "ជ ដ ឧ ណ ព ឋ ឍ ព": មុជ រាជ កោដ្ឋ ទ្រុឌទ្រោម ប្រាណ លេណដ្ឋាន ភាព ...
15. The word has double consonant: ខេត្ត វគ្គ ប័ណ្ណ ពេទ្យ សមុទ្រ រាស្ត្រ វល្លិ រដ្ឋ សំបុត្រ សង្ឃ ...
16. A consonant and a vowel together act as a consonant: ភូមិ ជាតិ ប្រតិបត្តិ វល្លិ ត្រុដិ ...
17. The word has the same form as original Khmer word (Usually one or two syllable): គោ ស្រី ប្រុស កាល ពិការ វាចា គ្រូ សាលា សិត សុបិន កីឡា លីលា សុញ្ញ នារី ...

However, there are some exceptional cases for some words which are not borrowed from Pali-Sankrit but still has the same form as in the rule 2, 4, 17 and others.

Examples:

1. Khmer original words: ខ្ញុំរ ញ៉ែរ ជំរ ទំល ស្ពាន់ធ៉ែរ សង្ឃ៉ែរ សូរប៉ុងៗ ...
2. Foreign words: អេដស៍ ស៊ីម៉ង់ត៍ ដុកទ័រ កុងទ័រ កុំព្យូទ័រ វិចទ័រ ត្រាក់ទ័រ...

Borivasap is the term often used in daily Khmer word which one word is added to another word to combine as a new word but keep the same meaning. There are a few ways to recognize the Borivasap word but in this thesis, we show only how to identify them by their sound patterns. In this case, the Borivasap has 2 syllables and 4 syllables. For the 2 syllables case, the sound of the first syllable is similar to the second's because the same consonants are used. Not different from the previous case, the 4 syllables, the first and second syllable are alike to the third and fourth. In the Borivasap, there are usually 2 to 8 consonants in the word as you can see in **Figure 1.2** and the following examples:

1. Borivarsap with 2 syllables: យំប៉ែក ឡូឡា ងក់ងរ លូកលាន់ ល្អិតល្អន់ ង៉ៃចរេ សមស្ងួន ហួងហែង គិតគូរ គ្រាំគ្រា ឃោរឃៅ ខិតខំ ខ្វល់ខ្វាយ ចែចង់ ឆ្អែឆ្អា ឆោមឆាយ ជាតិជៅ តាក់តែង ...
2. Borivarsap with 4 syllables: កំព្រីកំព្រា បន្ថែបន្ថុក គំហែះគំហើយ រវើរវាយ ឧដុង្គឧត្តម សង្កត់សង្កិន សម្ងួសម្ងុក អណ្តែតអណ្តូង រង្វេងរង្វាន់ ងងឹតងងុល ដំណែដំណើង ញាញីមញាញែម ចង្អៀតចង្អល់ ...

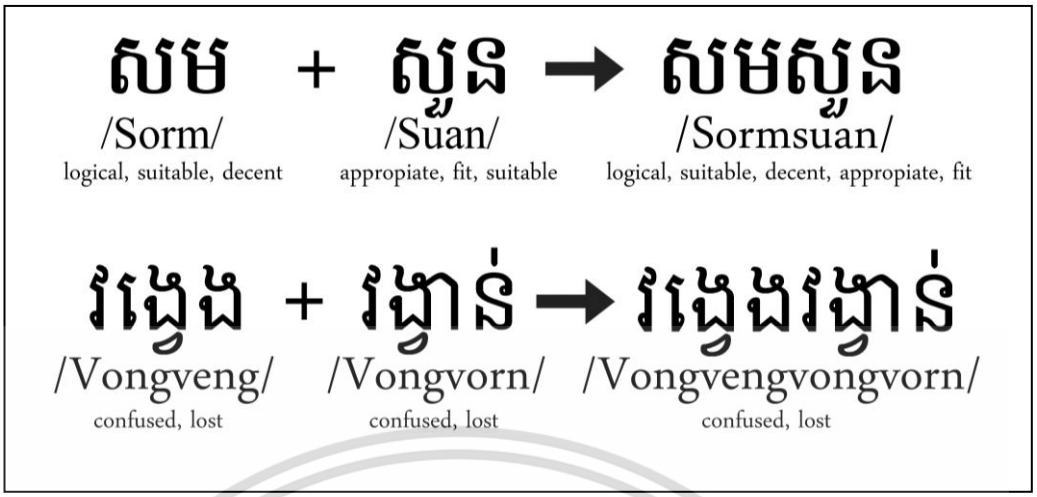


Figure 1.2 Examples of Khmer Borivasap words

1.3 Research Problem

Not like English and other Latin languages, in the Khmer standard of writing, boundaries of words inside a sentence or clause are usually omitted. There are no specific grammatical rules for word separation. These matters of unsegmented words have been causing huge problems for information retrieval, machine translation, text-to-speech, and many other natural language processing (NLP) applications. The reason is that we cannot solve these NLP problems without having done the word separation. In other word, word segmentation is the first assignment we need to do, and it is a great significant to solve the ambitious problems.

Several approaches which are related to segmenting words based on dictionary have been studied. The methods cannot solve the problem when it come to the out-of-vocabulary word. There are only a few researches about solving unknown word problem and they are not effective enough. This matter is a quite challenge task in Khmer word segmentation.

1.4 Research Objectives

Based on the problem mentioned above, in this thesis, we aim to propose a Khmer word segmentation technique which has the capacity to solve the problem of unsegmented sentences and to deal with the out-of-vocabulary words of the Khmer language. The approach is consisting of 2 steps. The first step is to do the word segmentation based on the Khmer corpus by using Maximum Matching approach. If every word inside the sentence matches with the dictionary, we can obtain the result of segmenting word in this step. The second step is related to solving the problem of word which does not exist in the dictionary by using rule-based techniques. To deal with the problem, we have used 21 rules which were created based on the principle of Khmer grammar book.

1.5 Research Scope

There are several approaches used for word segmentation and they are often studied in Asian languages, i.e. Japanese, Thai, Lao, Khmer, etc. A long with word segmentation task, a few assignments may also be done. This is related to identifying part-of-speech (POS) and name entity recognition (NER) for the segmented words. In this study, we focus on Maximum Matching algorithm for Khmer word segmentation based on the dictionary. We also pay attention to solving the problem of unknown words which do not exist in the dictionary by applying with Rule-based techniques. There are 21 rules for solving the problem and they were created based on the principle of Khmer grammar.

1.6 Research Method

To implement Khmer word segmentation, it is involving with three modules such as datasets, algorithm and programming. The datasets are consisting of a Khmer corpus and a collection of testing sentences. The corpus and the testing sentences were created

based on the data from the Internet by implying with 3 steps detailed in Section 4.1. The datasets are stored in Notepad.

The unsegmented sentence problem was solved by applying Maximum Matching algorithm along with Rule-based techniques. The Maximum Matching made the boundaries of each word inside the sentences based on the Khmer corpus, and the Rule-based solve the problem of unknown by using 21 rules.

To apply the algorithm for doing word segmentation testing, programming task are to be done. In this study, we used python programming language along with the PyCharm integrated development environment (IDE). Before the coding, a Khmer Unicode software (KhmerUnicode2.0.1.exe) along with Khmer fonts were needed to install and insert to the computer. And then UTF-8 encoding was used to recognize the Khmer Unicode.

1.7 Structure of the Thesis

Chapter 1: Introduction

This chapter provides an overview of Khmer language, the problem and the objective of the research (motivation), and the thesis structure.

Chapter 2: Literature Review

Some existing solutions such as Bigram, Bi-directional Maximum Matching, Constrained Conditional Random Fields, and Rule-based which are already used to make boundaries inside the Khmer phrases or sentences are mentioned in here.

Chapter 3: Proposed Solution

This chapter describes about Maximum Matching and Rule-based techniques which are the methodologies used to do the Khmer word segmentation based on dictionary and solve the unknown word problem.

Chapter 4: Experimental Results and Discussion

This chapter mentions about the dataset which are used to create the Khmer corpus and sentence testing. It also shows about the result of the Khmer word segmentation in each article and in average. Furthermore, it discusses about the affective of corpus for the baseline method, the reason of choosing Rule-based technique for unknown word problem, the result of word segmentation and the limitation of the Rule-based, and the problem with the previous works.

Chapter 5: Conclusions and Recommendations

This chapter concludes and summarizes all of contents in this thesis. It also mentions about the recommendations which should be done in the future for Khmer Word Segmentation.

CHAPTER 2
LITERATURE REVIEW

A great deal of studies related to word segmentation has been researched and they are often done in most Asian languages such as Chinese, Thai, Burmese, Javanese, Vietnamese, Lao and Khmer.

In this study, 4 existing solutions related to Khmer word segmentation have been investigated. They are the word segmentation based on Bi-gram model, Bi-directional Maximum Matching, Constrained Conditional Random Fields, and Linguistic and Rule-based.

2.1 Bi-gram Model

Word Segmentation of Khmer written text based on Bi-gram model was presented by Chea, S. H. et al. (2006) to segment the Khmer sentences based on corpus. This model carried out two extended methods which are Orthographic Syllable and Word bi-gram. In order to do word segmentation, first the characters in each sentence were merged into combination of characters called Khmer Character Cluster (KCC). Then, KCC matching module read KCCs one by one from left to right and matched them. And then, the KCCs were converted into KCE string. The KCE string was used to look up whether the words exist in the dictionary or not. After finishing the lookup task, multiple possible segmentation was obtained. Finally, bi-gram model was used to select an appropriate segmentation. This model used disambiguation model to choose the best segmentation among those candidates. The process of the segmentation is illustrated in **Figure 2.1**.

For doing the experiment, this research used a corpus with the size of 10.6 MB which contained 673295 words for training and the 13025 words for testing. The result of word segmentation is obtained as **Table 2.1**.

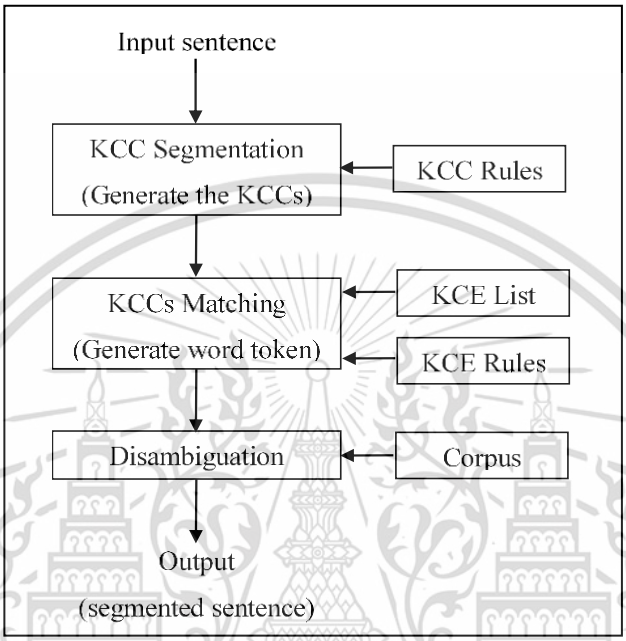


Figure 2.1 Flowchart of the system using Bi-gram model

Source: Chea, S. H., et al. (2006). Word Bigram VS Orthographic Syllable Bigram in Khmer Word Segmentation. PAN localization, Cambodia.

Table 2.2 Comparison of result between Word Bigram and KCC Bigram

	Precision	**Recall**	**F-measure**
Word Bigram	91.562	92.138	91.849
KCC Bigram	72.327	72.438	72.382

Source: Chea, S. H, et al. (2006). Word Bigram VS Orthographic Syllable Bigram in Khmer Word Segmentation: PAN localization, Cambodia.

2.2 Bi-directional Maximum Matching

Word segmentation proposed by Bi, N. et al. (2014), the word segmentation was done based the dictionary by using Bi-directional Maximum Matching and some linguistic rules such as Khmer Character Clusters (KCC) and Khmer Unicode Error Correction (KUEC).

First, each character inside the input sentence were merged into Khmer Character Cluster (KCC). Then, the characters position inside the cluster which are not correct was modified. Mainly, only the characters in the first and second position after the main characters are needed to interchange each other. And then, Bi-directional Maximum Matching algorithm (BiMM) was applied to do the segmentation as shown in Figure **2.2**. BiMM is the combination of Forward and Backward Maximum Matching. This method made the boundaries inside the sentence based on the Khmer corpus by scanning and segmenting text from the beginning to the end and starting from the end to the beginning. By combining this two together, it could solve the problem of ambiguity in Khmer Word Segmentation.

To do the experiment, 85,000 of unique words were used as the corpus list. Document for experiment were randomly selected from general administration letter, books, and Khmer news website. With the proposed method and the dataset, they could achieve the accuracy of 98.13% and spent 2.581 seconds for 160 000 of Khmer words.

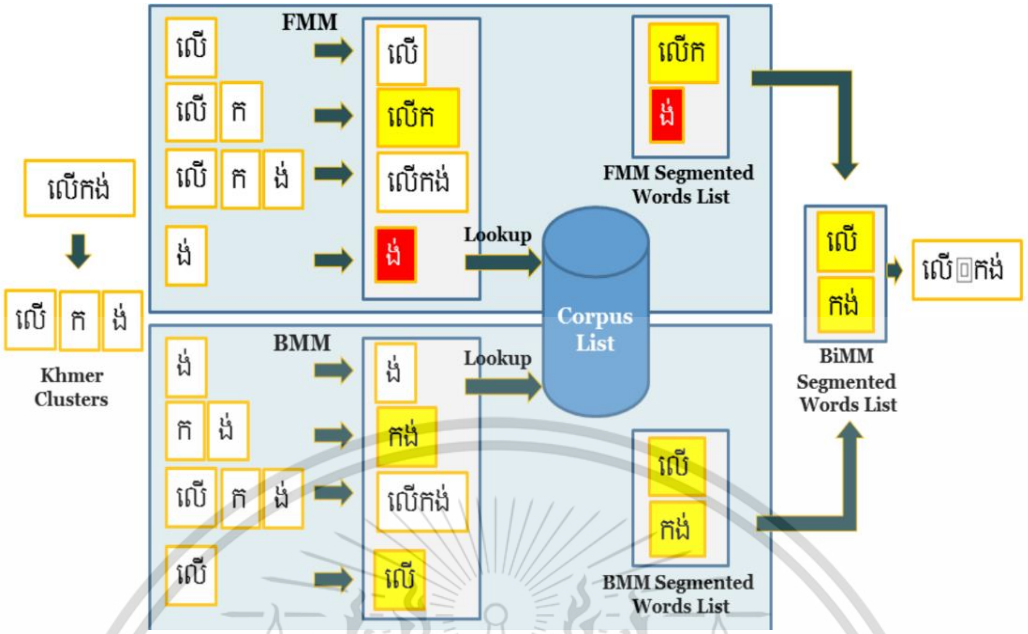


Figure 2.2 Process flow of Bi-directional Maximum Matching

Source: Bi, N. et al. (2014). Khmer Word Segmentation based on Bi-Directional Maximum Matching for Plaintext and Microsoft Word Document: Royal University of Phnom Penh, Cambodia.

2.3 Constrained Conditional Random Fields

Huang, S. et al. (2016) prososed a technique called "Constrained Conditional Random Fields (CCRF)". The study mainly focused on solving name entity recognition (NER) task. NER concerns with the identification of the individual entity in text, basically including names of people, locations and organizations. To achieve the assignment, 3 implementation steps were done. First, the added OOV (out-of-vocabulary) entity words in user dictionary were defined as constrained. Then, CCRF model was applied to segment and tag the part of speech of words. In this case, the constrains of proper noun were employed. And finally, NER-CCRF was used to recognize named entity. The recognized entity names were saved to the Khmer Entity Corpus. The process of CCRF is shown in **Figure 2.3**. For doing experiment, a large quantity of corpus from PAN Localization Cambodia and web documents was obtained. With this method, Khmer word segmentation achieved precision of 90.78% and recall

of 91.341%, and name entity recognition with 86.07% on precision, 86.54% on recall, and 86.30% on F-measure.

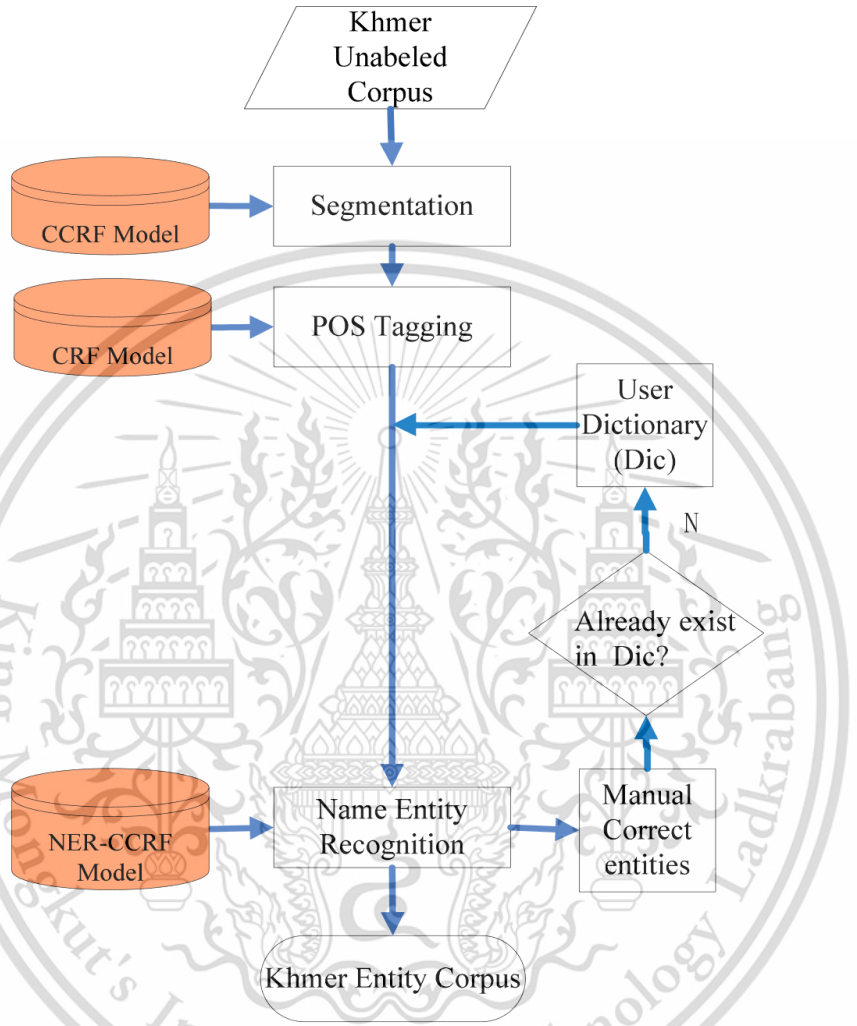


Figure 2.3 Flowchart of CCRF process

Source: Huang, S., et al. (2016). Construction of Khmer Entity Annotation Corpus Based on Constrained Conditional Random Fields: University of Science and Technology, Kunming, China.

2.4 Collocation of Repeated Characters Subsequences

Van, C. et al. (2013) presented a rule-based technique obtained by statistical analysis as well as specific linguistic rules of Khmer to tackle the issue of OOV word, compound words, proper name, derivative words and new words. Using these techniques, Khmer word segmentation could achieve 77.70% on precision, 75.55% on recall, and 76.50% on F-measure.

First, a large collection texts along with a rule-learning algorithm called SEQUITUR algorithm were used. SEQUITUR is an algorithm used to create the rules of repeated character subsequences. There are two steps in this rule learning process: subsequence extraction and rule extraction as shown in **Figure 2.4**. In the subsequence step, Longest Word Matching was used to segment the texts by using a Khmer word list. The outcome was defined as an array of extracted term. And the second step was applied to discover the rules of the subsequences that appear more than once from the array of extracted terms. A rule set was obtained based on the training corpus.

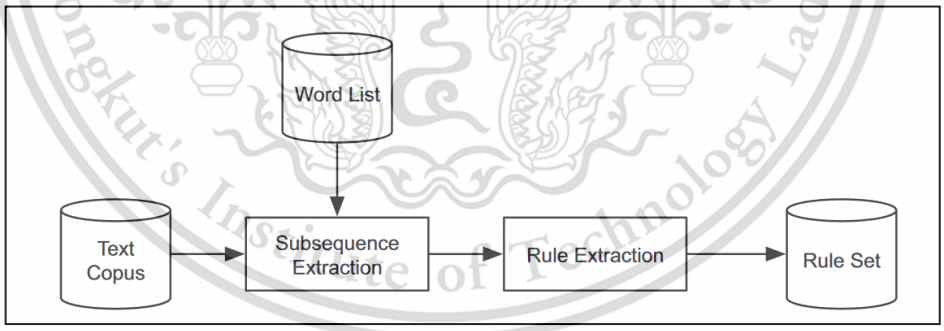


Figure 2.4 The Rule-learning Process

Source: Van, C. et al. (2013). Khmer Word Segmentation and Out-of-Vocabulary Words Detections using Collocation Measurement of Repeated Characters Subsequences: Waseda University, Japan.

The statistical measurements were used in order to measure the frequency of the collocation of the rules. After completing the rule learning, the rule tagging step and the word extraction step were carried out. The rule tagging was done based on different kind of statistical measurement that are entropy, mutual information, mutual

dependency, log-frequency biased mutual dependency, and chi-square test. These statistical measurements were applied to weight the strength of each rule to be a word. The process of word segmentation using collocation of repeated characters subsequences is shown in **Figure 2.5**.

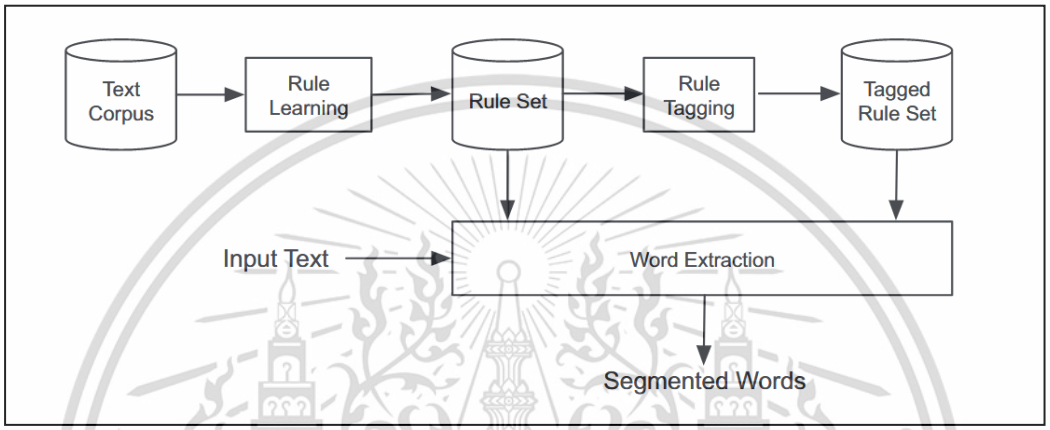


Figure 2.5 The Process of Khmer Word Segmentation using Rule-based

Source: Van, C. et al. (2013). Khmer Word Segmentation and Out-of-Vocabulary Words Detections using Collocation Measurement of Repeated Characters Subsequences: Waseda University, Japan.

2.5 Conclusion of the Related Studies

The word segmentation techniques mentioned in the 4 sections above are the approaches used for solving the problem of unsegmented words, recognizing name entity and solving the problem of unknown words in Khmer sentences.

Chea, S. H. et al. (2006) and Bi, N. et al. (2014) proposed the Bi-gram and Bi-Directional Maximum Matching which are the methods that can only deal with Khmer word segmentation based on the dictionary. On the other hand, when it comes up with the unknown word problem, it will obtain the negative result.

The Khmer word segmentation technique proposed by Huang, S. et al. (2016) could handle with the task of Khmer word segmentation based on the corpus. It also could

deal with the problem of unknown words, but it focused mainly on the problem of name entity recognition (NER).

Van, C. et al. (2013) presented a technique which could handle with both Khmer word segmentation based on dictionary and unknown word problem. To solve the task of unknown word, a rule learning and a rule tagging task were needed. Concerning with both of the rules, there are several smaller tasks need to be done. It took long times to obtain the result of solving the unknown words.

In conclusion, only two of the methods mentioned above could solve the problem of out-of-vocabulary word. However, the work proposed by Chea, S. H. et al. (2006) should be able to deal with the general unknown words. In addition, the work proposed by Van, C. et al. (2013) should take a shorter steps and time performances for solving the issue.

For this reason, we would like to propose a new technique which can handle with the unknown word problem of Khmer word segmentation in a better way. To process with this task, we first need to do a word segmentation task based on the dictionary. Based on the result by Huang, S. et al. (2016), we chose Maximum Matching algorithm. After solving the dictionary-based segmentation, we used the Rule-based techniques with 21 rules to solve this matter. The Rule-based techniques were chosen based on majorities of Khmer words taken from Pali-Sankrit and the popularity of using Borivasap words as mentioned in section 1.2.

CHAPTER 3
PROPOSED SOLUTION

This study proposes to improve Maximum Matching algorithm in Khmer Word Segmentation with Rule-based method. The process of segmentation is shown in **Figure 3.1**. First of all, whitespace and invisible space which can cause problem to word segmentation in each sentence were eliminated. There are no specific rules for using the whitespace and it is impossible to see the invisible space which is provided by Khmer Unicode. We did that to prevent from the error space due to most of our contents were collected from Internet. Then word segmentation was done based on the Khmer corpus. This step made the boundaries of words in each sentence into single words based on the Longest Matching approach. And then, the unknown words which do not exist in the dictionary were then defined and solved by using 21 grammar rules created.

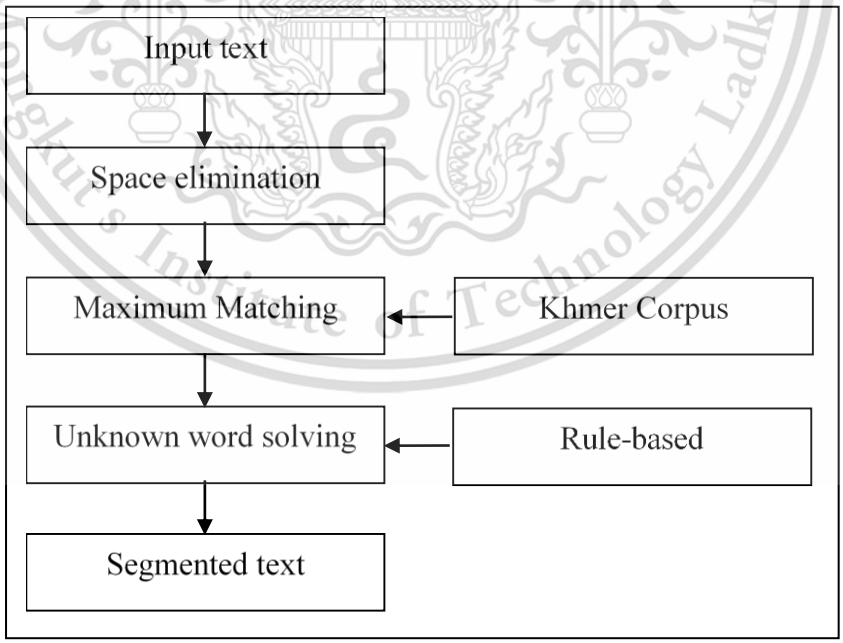


Figure 3.1 Khmer word segmentation flowchart

3.1 Maximum Matching Algorithm

Maximum Matching algorithm is one of the most popular and powerful word segmentation technique that use a baseline method in word segmentation [Liu, T. Y. et al., 1994]. In order to do word separation, this method compares and finds the longest matched word from the dictionary. If all the words in a sentence matches with the dictionary, the result is obtained as result 1, nor is shown as result 2. If the words in each sentence do not exist in the dictionary, they are defined as the unknown word and will be solved by using our created rules.

Algorithm:
1. *Start at the first character and try to find the longest matched word from the dictionary.* 2. *When the word is found, set the next character as the new starting point.* 3. *The algorithm ends when it reaches the end of the sequence* 4. *Otherwise, go to (1).*

Figure 3.2 Maximum Matching algorithm for Word Segmentation

Sentence 1: ដោយសារតែក្តីអាណិតនិងស្រលាញ់គាត់បានផ្តល់កន្លែងស្នាក់នៅព្រមទាំងលុយកាក់សម្រាប់នាងបន្តការសិក្សានៅភ្នំពេញ

Result 1: ដោយសារតែ/ក្តីអាណិត/និង/ស្រលាញ់/គាត់/បានផ្តល់/កន្លែងស្នាក់នៅ/ព្រមទាំង/លុយកាក់/សម្រាប់/នាង/បន្ត/ការសិក្សា/នៅ/ភ្នំពេញ

(Because of pity and love he has supported her accommodation and money for her study in Phnom Penh)

Sentence 2: នាយករដ្ឋមន្ត្រីអង់គ្លេសទិទៀនបក្សប្រជាជាតិថាចង់បំបែកស្កត់លែនចេញពីអង់គ្លេស

Result 2: នាយករដ្ឋមន្ត្រីអង់គ្លេស/[ទិទៀន]/បក្សប្រជាជាតិ/ថា/ចង់បំបែក/ស្កត់លែន/ចេញពី/អង់គ្លេស
(The England Prime Minister criticized that the Nation Party want England and Scotland break up)

In result 2, there is an unknown word defined in [] symbol ([ទិទៀន]). This unknown word will be solved in the next step.

3.2 Rule-based Techniques

Most common Khmer words are borrowed from Pali-Sankrit. We can recognize them based on their characteristics or grammar rules. A Khmer grammar book written by Chhorn, C. (2002) describes about how to identify the original Khmer words and the Khmer words which taken from Pali-Sankrit. Based on the grammar rules, by investigating the pattern of those Pali-Sankrit and Khmer words, to solve the unknown word (UW) problem, we have created 21 rules. We built the rules by counting and comparing the characters and the order of dependent vowel (DV), independent vowel (IV), consonant (C), and special character and symbol (S). **Figure 3.3** shows the structure of word combination of Rule 6 and Rule 9. These rules are called character combination and divided into 3 groups. Group 1 (Rule 3-5) is the group of no DV, Group 2 (Rule 6-11) is the group of one DV, Group 3 (12-21) is the group of two or more DV. When an unknown word matches with the rules, the new word will be created, and the problem of out-of-vocabulary are solved. In the symbol case, we only focused on some special characters such as "◌ ◌ ◌ ◌ ◌ ◌ ◌ ◌ ◌:".

Abbreviations:

1. **UW**: an unknown word/out-of-vocabulary word which is cannot be solved in the Maximum Matching step.
2. **C**: stands for consonant, including its capital scripts and subscripts.
3. **S**: refers to diacritics/special characters.
4. **DV**: represents dependent vowel, the vowel cannot stand without a consonant.
5. **IV**: denotes independent vowel, the vowel that can act as the consonant.

Figure 3.3 Description of the abbreviations

Rule 1: *No English or unknown characters*

Rule 2: *No numbers*

Rule 3: *If UW contains some C (2 to 8) with zero, one or two S (္), these characters will be combined as a new word.*

Example: ភព កក ផល ជន មរតក កង ពល កញ្ចត កម្ម ពលករ ក្រហម គណបក្ស ជប ជនក សម្ពព ជប ក្អក ផ្ទះ ផ្នក ល្អ អក្សរ ក្រម គ្រប កករ ក្រគរ បក ដប ចប នរក សកល ខ្សង សម្ព វត្ត ហង្ស ស្កក ប្រឈម មណ្ឌល ពលរដ្ឋ សម្ផស្ស កម្មករ បង្កក បង្កង ទ្រព្យធន ប្រភព ប្រកប ក្រចក បបរ ក្អម ក្រ ថ្ម ...

Rule 4: *If UW contains one S and same rule as Rule 3, these characters will be combined as a new word.*

Example: កក់ កង់ កង្វល់ កញ្ចក់ កត់ រស់ ទន់ អន់ ឡប់ ទល់ ដប់ សក់ បក់ លក់ ពណ៌ មត៌ក គត៌ បត៌ត ខ្ទះ ជ័រ ក្ប័យ ន័យ សម័យ ជ័យ ព័ន្ធ ព័ទ្ធ រហ័ស ភ័យ មធ្យ័ត ប្រស័យ រ៉ុន វ័យ សង្ឃ័យ រយៈ អក្ខរ ខណៈ វណ្ណៈ មរណៈ ទស្សន៍ ពលៈ វចនៈ របស់ ខ្សត់ ឆន្ទៈ ជ្រលក់ វប្បធម៌ កខ្វក់ កញ្ចក់ កញ្ចប់ ទ័ព ...

Rule 5: *If UW contains a/two IV (s) and same rule as Rule 3 or 4, these characters will be combined as a new word.*

Example: ឧត្តម ឥស្សរជន ឥត ឯក ឯកជន ឧក្កដ្ឋ ឯកវចនៈ ឧត្តរ ឧទរ ឱ្យ ឧស្ម័ន ឧប ឧបករណ៍ ឧក ឧក្កដ្ឋ ឧច្ច ឧទ្ធរណ៍ ឥដ្ឋ បឋម ឱសថ ឧបត្ថម្ភ ឫក ឫស ហឫទ័យ អមឫត ឱន ឧបក្រម ឧបសគ្គ ឧបស្ស័យ រឭក ឧបសម្ព័ន ឧម្មង្គ ឯកឧត្តម ពលឯក ឯតទគ្គកម្ម ឦសធរ ឯកវចនៈ ឫក្សណៈ ឫសភ ឦស ឧណ្ហ ...

Rule 6: *If UW contains a DV with zero, one or two S (s), starts with a C and ends with a C, these characters will be combined as a new word.*

Example: សមាជ សុខ នាគ ឆាគ ប្រមុខ ប្រយោគ សម្ភោជ ប្រាកដ ចៅក្រម ប្រភេទ អាល័យ វិស័យ អនាម័យ និស្ស័យ ប៊ិច ប៉ូវ នាគ សុក្រ ពុធ មាន កន្លែង សប្បាយ ឃ្លាំង បាំង តែប កូន នុយ ក្នុង ឆ្នុក អភិជន មនុស្ស កន្ទុយ កណ្តុរ កន្ត្រែង គុក គាប ខួប ខូច ឃើញ ឃ្លាត ងងឹត ងូត ចន្ទាល់ ចម្ការ បង្អែក ...

Rule 7: *If UW contains one/two DV(s) with one, two or three S (s), starts with a C and ends with a S, these characters will be combined as a new word.*

Example: ប្រយោជន៍ និរន្តរ៍ ឆាចន៍ ព្រាហ្មណ៍ ប្រសាសន៍ ព្រាហ្មណ៍ អភិវឌ្ឍន៍ ស្រាពណ៌ ព្យាករណ៍ ស្នេហ៍ សៅរ៍ អាចន៍ មគ្គុទេសក៍ អក្សរសិល្ប៍ អរជន៍ ភារៈ សុខៈ មោហៈ លោភៈ តេជៈ វិរិយៈ ជីវៈ សីហៈ ភាវៈ តេសជ្ជៈ ឃោសៈ ធុរៈ គារវៈ ថាវរៈ កុម្ភៈ សក្ការៈ ខេមរៈ សួគ៌ អាថ៌ អរិយធម៌ បរិបូណ៌ គុណធម៌ ...

Rule 8: *If UW contains a DV with zero, one or two S (s), starts with a C and ends with a DV, these characters will be combined as a new word.*

Example: សេះ ហោះ គោ កោះ ទា លឺ ទឹ ជ្រៅ ញើ រញឹ ស្វា ធំ មុំ ទេ ភេ ក្រពា ក្រពើ ត្រសេះ តឿ ដីផ្ទឹ ពន្លៃ ប្រេះ ប៉ះ ខ្ទុំ ស៊ី ញុាំ ជ្រោះ កោះ ទា លាភ ដី ស្រី សុាំ ចក្រី ករី រសជាតិ ផ្សុំ រន្ទះ កង្រឹ ត្រូ ជលធី លតា សភា ខន្តិ វល្លិ ឃ្លៅ ខ្ចី ឃ្លី គោះ ចោះ ចោះ កា សី មន្ទីរ ត្រីៈ ចង្កឹៈ មេ ស្តី ចន្ទី ក្របី ...

***Rule 9**: If UW contains a DV with zero, one or two S (s), starts with a IV and ends with a C, these characters will be combined as a new word.*

Example: ឥស្សរ ឥសាន ឧទ្ទិស ឧក្រិដ្ឋ ឧត្តាត ឧទ្ទាម ឧទ្ទេស ឧបាសក ឫសដូង ឥណាទាន ឱនភាព ឪពុក ឧទាន ឯក ឯកការ ឯកសារ ឯកភាព ឯកទេស ឯកសេសន័យ ឯកអគ្គរដ្ឋទូត ឧបាយកល ឧទ្ទេស ឧកហ្លង ឧដុង្គ ឧបកិច្ច ឧបទ្វីប ឧបនិស្ស័យ ឧស្សាហកម្ម ឱកាស ឪម៉ាល់ ឫក្សពារ ឫក្សៃស ឥឡូវ ឪឡឹក ...

***Rule 10**: If UW contains a DV with zero, one or two S (s), starts with a IV and ends with a DV, these characters will be combined as a new word.*

Example: ឧបមា ឧស្សា ឯកា ឯកោ ឯតទគ្គ ឯថា ឧតុ ឧដុ ឧណ្ណា ឧភតោ ឯណា ឧកញ្ចា ឯកតោ ឥច្ឆា ឥសី ឫសី ឥន្ទន្ទ ឫកពា ឫស្សី ឱរា ឩកា ឪម៉ែ ឯកតា ឯកតោ ឫទ្ធិ ឫសដី ឫស្សា ឪជំ ឥច្ឆា ឥណ្ណា ឥន្ទ្រា ឥណ្ឌ ឥត្ថី ឥន្ទណំ ឥរា ឫតិយា ឫទ្ធា ឫទ្ធិ ឫសដី ព្រជា ឯតើ ឯថា ឱង្កា ឱរ៉ែ ឥស្សា ឧត្តមោ ឧត្តរិ ...

***Rule 11**: If UW contains a DV with one or two or three S (s), starts with a IV and ends with a S, these characters will be combined as a new word.*

Example: ឧទាហរណ៍ ឧទ្ធរណ៍ ឧស្សាហ៍ ឧស្សាហៈ ឧបករណ៍ ឋានៈ ឱត្តប្បៈ ឥន្ធនៈ ឧត្តមៈ ឧទរិយៈ ឧបនាហៈ ឧបនិមន្តន៍ ឧបវសន៍ ឧបាទរ៍ ឧដ្ឋានៈ ឧត្រាសៈ ឧត្សាសៈ ឧបជ្ឈាយ៍ ឧបវសន៍ ឧបាទរ៍ ឧល្ហង្ឃនៈ ឯកវចនៈ ឧបក្ការៈ ឱសធៈ ឱសថៈ ឱត្តប្បៈ ឱកាកៈ ឯកទគ្គៈ ឧល្លោកនៈ ឧភយៈ ឧបោសថ៍ ...

Rule 12: If UW contains two DVs, two Cs with zero, one or two S (s), and C1 = C2, these characters will be combined as a new word.

Example: រុះរើ ឡូឡា មួម៉ៅ នានា ឆាឆៅ ញ៉ោះញ៉ោះ រ៉ាំរ៉ៃ ឡេះឡោះ ទូទៅ អេះអុះ ឆាំឆា ដោះដៃ ងេងើ ខះខំ ដុំដី ដុំដៃ ដាំដែង តេះតោះ តែតោ ទីទែ ណេះណោះ ទូទៅ បំបៅ បំបះ បំប៉ះ ណោណា សំសែ ហាហា ហេហា ហែហែ កិកិ ទីទែ កាកី ជើជើ យាយី យ៉ោះយ៉ុ អរា ង៉ៀរ៉ៃ សិសុ សោះសា សំសែ ហេហា ...

Rule 13: If UW contains two DVs, three Cs with zero, one or two S (s), and C1 = C2 or C1= C3, these characters will be combined as a new word.

Example: ដុះដាល អ៊ូអរ ទូទាត់ រារាំង រាម៉ែក ជុំជិត ចែចូវ មាំមួន ធំធេង យំយែក ទីទុយ ទីទើរ បេះបួយ បេះបិទ បោកបោះ បំបិទ បំបួស បំបែក រីងរៃ រួញរា លែងលះ ឆិនឆៃ សោកសៅ ហៀវហោះ កោះកុង រិវាទ ខាន់ខៅ ខាបខា ខិតខំ ដុំដែក ដឹកជឿ ថាថាង ទាមទា ទូទាត់ ទីទើរ ធំធាត់ បំបាក់ បំបាត់ សាក្សី ...

Rule 14: If UW contains two or four DVs, four Cs with zero, one or two S (s), and C1= C3, these characters will be combined as a new word.

Example: ចាកចេញ តិចតួច គិតគូរ ស៊ើបសួរ ណែនណាន់ ដុនដាប តាក់តែង ថ្លៃថ្លា យឺតយ៉ាវ លូកលាន់ រៀចរេរ ហួងហែង កៀកកើយ មាយីមាយា ហែកហួរ ទៀងទាត់ នាងនួន ពេមពើម រីករាយ រីងរួស រុងរឿង រួសរាន់ រឿងរ៉ាវ រៀបរាប់ លុកលុយ ហេលហាល ឡុកឡាក់ អួតអាង ឆែកឆេរ ជួសជុល ដុនដាប ...

Rule 15: If UW contains two or four DVs, five Cs with zero, one or two S (s), and C1= C3 or C1 & C2 = C3 & C4 or C1 & C2 = C4 & C5, these characters will be combined as a new word.

Example: ស្តុកស្តំ គំរោះគំរើយ តំណិះដំណៀល ដំណែដំណឹង ត្រៀបត្រា ធ្លុះធ្លាយ ឈ្លេចឈ្លី ក្រៀមក្រំ ខ្ពុះខ្ពាយ ស្ងុះស្ងែង ឆ្មើងឆ្មៃ ញាញីញាញ័រ ធ្លុះធ្លាយ ឈ្លុះឈ្លាន នែបនិត្យ សាបសូន្យ ស្លឹមស្ដែ ផ្តេកផ្តោះ ម្នាម្នេញ ឆ្កាំឆ្កង ឃ្លាតឃ្លា ត្រួតត្រា ខ្មេចខ្មឹ ផ្គូផ្គង ម្នាម្នេញ ប្រែប្រួល ឆ្ពិនឆ្ពៅ ចាប់ចិត្ត ថ្កុំថ្កើង ខ្ទះខ្ទែង ...

Rule 16: If UW contains two or four DVs, six Cs with zero, one two or three S (s), and C1 & C2 = C4 & C5, these characters will be combined as a new word.

Example: ខ្សឹបខ្សៀវ អាពាហ៍ពិពាហ៍ ទំនៀមទំលាប់ ទទឹងទទែង ជ្រោមជ្រែង ញញឹមញញែម ចំតិតចំតួង កំប៊ុកកំប៉ុក ទទឹងទទែង ផ្តេសផ្តាស រប៉ាតរប៉ាយ ផ្ទាញ់ផ្ទាល រខេករខាក ស្តើតស្តើន លលីមលលាម កកិចកកុច ក្តែងក្តាយ ខ្ជាប់ខ្ជួន ខ្ញិកខ្ញុក ឃ្លេងឃ្លោង ច្រិមច្រុម ផ្ទៀវផ្ទាត ត្រាណត្រើយ ថ្នាំងថ្នាក់ ខ្វាត់ខ្វែង ថ្កើងថ្កាន ...

Rule 17: If UW contains two three or four DVs, seven Cs with zero, one, two or three S (s), and C1 & C2 = C4 & C5 or C1 & C2 = C5 & C6, these characters will be combined as a new word.

Example: បង្ខិតបង្ខំ អន្ទះអន្ទែង ស្រមោលស្រមៃ បញ្ចើចបញ្ចើ បន្ទាប់បន្សំ ស្រងេះស្រងោច បន្ថែបន្ថុក គគ្រែងគគ្រាំ ប្រឡិមប្រឡុំ អណ្តាប់អណ្តា អង្គាប់អង្គើ កញ្ជះកញ្ជាយ បាយឡុកបាយឡ សិស្សានុសិស្ស កម្ទេចកម្ទី ពន្លាក់ពន្លើ កម្រៀមកម្រោះ តម្កុំតម្កើង សន្ធឹកសន្ធៃ សម្រុះសម្រួល អង្កែលអង្កៃ ក្បែមក្បាន្ត...

Rule 18: *If UW contains two or four DVs, eight Cs up with zero, one, two or three S (s), and C1 & C2 & C3 = C5 & C6 & C7, these characters will be combined as a new word*.

Example: បន្តិចបន្តួច បណ្តើតបណ្តោយ ប្រកៀកប្រកិត ប្រញាប់ប្រញាល់ បំផ្លិចបំផ្លាញ ចម្រុងចម្រើន ចម្រុងចម្រាស ច្រងេងច្រងាង ច្រងាប់ច្រងិល ក្រញ៊ិកក្រញុក ត្រដាបត្រដួស កណ្តោចកណ្តែង កន្ត្រែកកន្ត្រោក គគ្រឹកគគ្រែង ក្រមេក្រវាម ស្រពេចស្រពើល ស្រវាក់ស្រវាន់ គ្រហឹកគ្រហុក ក្រឡេចក្រឡុច អង្គុកអង្គុល ...

Rule 19: *If UW contains two DVs (◌ា), zero or one S with two, three or four Cs, these characters will be combined as a new word.*

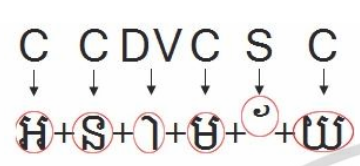
Example: មាគ៌ា មាតា សាវតា អាហារ ភាសា សាសនា អាត្មា វាចា យាត្រា ពាលា សាវតា ការពារ ទាយាទ ទាហាន សារ៉ា ការងារ នាដកា ជានា ឆាយា ជាតា សាលា កាឡា ខាន់ស្លា សាខា គាថា រាបសា នារ៉ា ចាម្ប៉ា រាចា តារា ទាមទា នាសា អាសា ពានា ភាវនា មាយា មាលា យាត្រា សាត្រា រាមា ធានា ...

Rule 20: *If UW contains two or three DVs (◌ា ◌ិ ◌ី ◌ុ), zero or one S with two, three or four Cs and ends with a DV, these characters will be combined as a new word.*

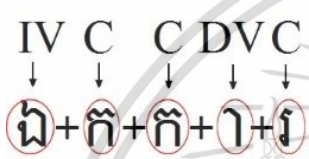
Example: នាទី សមាធិ កុមា តុលា មុតិតា សុរា គុហា មិថុនា ចាលី មិនា គុហា សីហា សុនីតា មុសា ភាគី នារី មាលតី នាឡិកា សាលាក្តី កុមារី សុជាតា បរិញ្ញា ថវិកា កិរិយា សីលា ឱរតី អធិបតី ពិធី កីឡា បុត្រាវីសា ដុល្លា បិតុលា កិរិយា សុធានី កល្យាណី កុលធីតា រាសី អធិបតី ចំប៉ី វិថី ទាសី ចរិយា ចិន្តា ...

Rule 21: *If UW contains two DVs (◌េ ◌ា ◌ី), zero or one S with two, three or four Cs, starts with (◌េ or ◌ោ) and ends with a DV (◌ា or ◌ី), these characters will be combined as a new word*

Example: មេបា មេសា សេនា ចេនឡា ទេសនា ផលា មេឃា ទេវតា ចេតនា ឃោសនា ផាហា ហោរា មេត្តា ជេស្តា ផហា យោធា កេសា ចេស្តា ទេសនា បេឡា សេដ្ឋា ជេដ្ឋា លេខា ប្តេជ្ញា ប្រាជ្ញា សេវនា ហេរិកា អចេតនា ខេមរា នេត្រា ស្នេហា សេរី សេដ្ឋី បេតី មេត្រី សេចក្តី មហេសី ផេទនា ទេសា សេវា ...



1. The word "អនាម័យ" starts with the consonant "អ", ends with consonant "យ", has a dependent vowel "◌ា" and a symbol "◌ៈ". It is in case of Rule 6.



2. The word "ឯកការ" starts with the independent vowel "ឯ", ends with consonant "រ" and has a dependent vowel "◌ា". It is in case of Rule 9.

Figure 3.4 Character combination in Rule 6 and 9.

CHAPTER 4
EXPERIMENT RESULTS AND DISSCUSSION

4.1 Experiment Setup

Various online contents and books from agriculture, magazine, newspaper, health, technology and history as shown in **Table 4.1** were used to create a manually Khmer corpus and a collection of testing sentences. With randomly 395 articles selected, we could build an annotated corpus with 19,500 unique words and 2018 testing sentences. The corpus and the collection of sentences were then used to do the experiment with the algorithms proposed.

Table 4.1 Types and Number of Articles

No	Article Type	#Articles
1	Agriculture	38
2	Magazine	20
3	Newspaper	220
4	Technology	40
5	Health	25
6	History	52
	Total	**395**

The raw contents which have been collected from the Internet cannot directly be used for the experiment. To get a ready dataset for testing, an annotated dictionary and a testing sentences are needed. This is involving with 3 processing steps such as Data Collection, Data Cleaning and Data Annotation. The corpus and testing are stored in Notepad files.

1. **Data Collection:** Since there is no standard Khmer corpus and because of the lack of Khmer optical character recognition to detect the texts from the electronic books, to build the dictionary, we have collected a variety of online

articles from different websites. Most of the contents that we have chosen are taken from popular and government websites which can be considered as reliable sources. We also have used few contents from some books and magazines.

2. **Data Cleaning:** The data collected from the Internet were because they have some error spelling and unwanted contents. To have a good dataset, we eliminated the unstructured data such as images, links, symbols and unnecessary texts from the contents. We also deleted spaces in the sentences including normal and invisible spaces. Moreover, we checked and corrected the spelling of each words based on Chuon Nat dictionary which recognized by the Ministry of Education.
3. **Data Annotation:** To reduce the redundancy of the contents, words inside the dictionary and the sentences inside the testing corpus must be unique. We did the annotation by deleting every word and sentence which existed more than once. This is the final datasets which are ready to apply with the algorithm to evaluate the result.

4.2 Results

To obtain the result of Khmer word segmentation, we applied Maximum Matching (MM) and Maximum Matching with Rule-based Technique (MM with Rule-based) with the created corpus and the testing sentences shown in **Table 4.2**. We calculated the accuracy of each articles by counting words which are correct segment comparing to their total words, see section 4.1. We got the average accuracy of 88.10 % in MM and increase to 92.02 in MM with Rule-based as shown in Table 4.3.

$$Accuracy = \frac{Number\ of\ correct\ instances}{\text{Number of total words}} \quad (4.1)$$

- Number of correct instances: total number of words in each article which are correct segment
- Number of total words: total number of words in each article

Table 4.2 Number of sentences, total words and unknown words of testing corpus

Article	# Sentences	# Total words	# Unknown Words	% Unknown Words
Agriculture	168	5125	242	4.72
Magazine	117	3653	215	5.88
Newspaper	1113	32438	1682	5.18
Technology	166	4789	214	4.46
Health	127	4124	197	4.77
History	327	13881	963	6.93
Total	**2018**	**64010**	**3513**	**5.48**

Table 4.3 Accuracy in each article using MM and MM with Rule-based

Article	MM		MM with Rule-baed	
	# correct instances	Accuracy (%)	# correct instances	Accuracy (%)
Agriculture	4601	89.77	4763	92.93
Magazine	3176	86.94	3329	91.13
Newspaper	28785	88.73	29865	92.06
Technology	4316	90.12	4458	93.08
Health	3754	91.02	3863	93.67
History	11763	84.74	12625	90.95
Total	**56395**	**88.10**	**58903**	**92.02**

After getting the result, we compared the **MM with Rule-based** with the **MM** to see the accuracy increasement for each article types and it is shown in **Table 4.4**.

Table 4.4 Increasing accuracy and correct instances of MM and MM with Rule-based

Article Type	**# Increasing correct instances**	**% Accuracy increase**
Agriculture	162	3.16
Magazine	153	3.77
Newspaper	1080	3.33
Technology	142	2.96
Health	109	2.65
History	862	6.21
Total	**2498**	**3.92**

4.3 Discussion

In this thesis, regarding to Khmer word segmentation, we would discuss about 4 issues: *the affective of corpus for the baseline method*, *the reason of choosing Rule-based technique for unknown word problem*, *the result of word segmentation and the limitation of the Rule-based*, and *the problem with the previous works*.

As mentioned in the previous chapter, baseline method is the approach for segmenting word based on corpus. Maximum Matching algorithm which have been applied is a kind of the baseline method. We cannot do the segmentation without having a dictionary. The size and the contents used to create the corpus are also important. It can affect to the accuracy of word segmentation. The bigger is size of the dictionary, the better is accuracy of word segmentation. In this thesis, we have used a corpus of

19500 words which can be efficient to deal with the popular articles such as agriculture, magazine, newspaper, health, technology and history.

There are many techniques for solving out-of-dictionary word problem when it comes to word segmentation. We have chosen Rule-based as the technique for solving unknown words which do not exist in the corpus. The reason is that many Khmer words are taken from Pali-Sankrit, the popularity of using Borivasap in Khmer sentences and some original Khmer and foreign words have the Pali-Sankrit form. In both Pali-Sankrit and Borisvasap forms, by seeing their pattern, as mentioned in section 1.3, we could convert and build the word structure rules based on their characteristics.

Based on the result in section 4.2, we can see that Rule-based cannot deal with every unknown word. There are still incorrect cases in word segmentation. Although we have built 21 rules to solve the problem, but some words in the testing sentence are out of rules due to some words are taken from foreign language and technical term. In **Table 4.4**, the accuracy of article in history increases so high comparing to others. In the history, we input the highest number of unknown words in term of percentage which make the baseline word segmentation got low accuracy and the accuracy got high when the Rule-based can solve many of them.

As presented in Chapter 2, there are 4 existing Khmer word segmentation techniques have been investigated. They are Bi-gram model, Bi-directional Maximum Matching, Constrained Conditional Random Fields, and Linguistic and Rule-based. Bigram is the model for segmenting word with a requirement of a corpus. It is high ambiguity and high computation in Bi-gram since the words is formed by the combination of one more KCCs [Chea, S. H. et al., 2006]. For the Bi-Directional Maximal Matching, it is a bit better in accuracy but big performance if comparing to

the traditional Maximum Matching [Bi, N., 2014]. It performs the searching task in both direction; start from the left to right and then from the right to the left. Constrained Conditional Random Fields presented by Huang, S. et al. (2016) focused on part of speech tagging (POS) and solving problem of unknown words in word segmentation but only name entity type. In Linguistic and Rule-based [Van, C. et al, 2013], SEQUITUR algorithm was used to solve out-of-vocabulary of word. It is an algorithm used to create the rules of repeated character subsequences. According to Chapter 2, this algorithm uses many steps just to solve the unknown word problem.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

In this thesis, we presented a Khmer word segmentation technique which have the capacity to segment words of the Khmer unsegmented sentences based on the corpus. This also described about how to solve the problem of unknown words which do not exist in the dictionary. We used Maximum Matching algorithm and Rule-based technique along with a Khmer corpus and a set of testing sentences. First, the segmentation based on the dictionary by using Maximum Matching algorithm was done. In this case, if all the words in a sentence matches with the dictionary, the result is obtained. If the words in each sentence do not exist in the dictionary, they are defined as the unknown word and will be solved by Rule-based Technique. The Rule-based is a technique for combine characters to become a word by using rules. We have created 21 rules and they were created based on the principle of Khmer grammar books [Chhorn, C., 2002]. With the proposed solution, we got an average accuracy of 88.10% in word segmentation based on the corpus and it increased to 92.02 % when the rules were applied.

We believe that our Rule-based method can solve many Khmer unknown words due to the effective of Pali-Sankrit on Khmer Language and the popularity of using Borivasap. The technique is also expected to use for Pali-Sankrit word segmentation.

Due to the limitation of the Rule-based, more rules should be created. There may be some other Khmer grammar structures rather than the Pali-Sankrit and Borivasap which we can also learn from their patterns and extract more rules.

REFERENCES

Alivanh, I. (2017). Word Segmentation and Part-of-Speech Tagging for Lao Language. A Master Thesis of Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

Bi, N. and Taing, N., "Khmer word segmentation based on Bi-directional Maximal Matching for Plaintext and Microsoft Word document," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Siem Reap, 2014, pp. 1-9. doi: 10.1109/APSIPA.2014.7041822

Chea, S. H., Top R., Ros, P. H. & Vann, N. (2006). Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation. PAN Localization Cambodia.

Chhorn, C. (2002). Khmer Grammar for General Studies. Phnom Peng, Cambodia.

Huang, S., Yan, X., Yu, Z. and Lei, Q., "Construction of Khmer entity annotation corpus based on constrained conditional random fields," *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, 2016, pp. 2246-2251. doi: 10.1109/FSKD.2016.7603531

Liu, T. Y. & Shen, K. X. (1994). The Word Segmentation Methods for Chinese Information Processing. Hua University Press and Guang Xi Science and Technology Press, China.

Long, P. and Boonjing, V."Longest Matching and Rule-based Techniques for Khmer Word Segmentation", The 2018 - 10th International Conference on Knowledge and Smart Technology, Jan 31- Feb 3, 2018,Chiang Mai, Thailand, 978-1-5386-4015-9/18/$31.00 @2018 IEEE.

Mahatthanachai, C., Malaivongs, K., Tantranont, N. and Boonchieng, E., "Development of thai word segmentation technique for solving problems with unknown words," *2015 International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, 2015, pp. 1-6. doi: 10.1109/ICSEC.2015.7401423

Thorn, H. (2011). Khmer Grammar for Phumseuksa. Phnom Penh, Cambodia.

Tran Van, Nam. (2017). Building a Syllable Database to Solve the Problem of Khmer Word Segmentation. International Journal on Natural Language Computing. 6. 10.5121/ijnlc.2017.6101.

Van, C. & Kameyama, W. (2010). Query Expansion for Khmer Information Retrieval. Proceedings of the 8th Workshop on Asian Language Resource, Beijing, China.

Van, C. & Kameyama, W. (2013). Khmer Word Segmentation and Out-of-Vocabulary Words Detection Using Collocation Measurement of Repeated Characters Subsequences. Graduate School of Global Information and Telecommunication Studies, Waseda University.

APPENDIX A

Longest Matching and Rule-based Techniques for Khmer Word Segmentation

Pakrigna Long
International College
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
longpakrinha@gmail.com

Veera Boonjing
International College
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kbveera@kmitl.ac.th

Abstract—**Word boundaries are the essential assignment to be done in natural language processing research. In most Asian languages, as well as Khmer language, many studies involved with word segmentation have been investigated. In Khmer Word Segmentation, several approaches related to segmenting words based on dictionary have been studied. There are only few researches about solving unknown word problem. This matter is a quite challenge task in word separation. In this research, Maximum Matching algorithm (MMA) together with Rule-based technique has been proposed. First, MMA and a Khmer manual corpus were used to make word boundaries in each sentence. Then the unknown words were then defined and solved by using 21 grammar rules created. We tested the segmentation with 2018 sentences from agriculture, magazine, newspaper, technology, health and history. With Maximum Matching alone, we could achieve the accuracy of 88.55% and along with Rule-based, the accuracy increased to 92.81%.**

Keywords- Khmer Word Segmentation, Maximum Matching, Rule-based, unknown word

I. Introduction

Khmer is the official and national language used by the people in Cambodia. There are 33 consonants, 23 dependent and 13 independent vowels, 10 digits of numeric, and approximately 27 diacritics and special characters as shown in TABLE I.

TABLE I. Khmer Characters

Consonants/ its subscripts	ក្ក ខ្ខ គ្គ ឃ្ឃ ង្ង ច្ច ឆ្ឆ ជ្ជ ឈ្ឈ ញ្ញ ដ្ដ ឋ្ឋ ឌ្ឌ ឍ្ឍ ណ្ណ ត្ត ថ្ថ ទ្ទ ធ្ធ ន្ន ប្ប ផ្ផ ព្ព ភ្ភ ម្ម យ្យ រ្រ ល្ល វ្វ ស្ស ហ្ហ ឡ អ្អ
Dependent Vowels	ា ិ ី ឹ ឺ ុ ូ ួ ើ ឿ ៀ េ ែ ៃ ោ ៅ ុំ ំ ាំ ះ ុះ េះ ោះ
Independent Vowel	ឥ ឦ ឧ ឩ ឪ ឫ ឬ ឭ ឮ ឯ ឰ ឱ, ឲ ឳ
Numeric	០ ១ ២ ៣ ៤ ៥ ៦ ៧ ៨ ៩
Diacritics/special characters	ៈ ៉ ៊ ់ ៌ ៍ ៎ ៗ ៏ ័ ៜ ។ ៕ () ៖ ? ! ៛ ៘ ៚ . , ៙ " /

In the Khmer standard of writing, boundaries of words inside a sentence or clause are usually omitted. There are no specific grammatical rules for word separation. These matters of unsegmented words have been causing huge problems for information retrieval, machine translation, text-to-speech, and many other natural language processing applications. Consequently, word segmentation is a great significant to solve ambiguity problems. Existing methods for Khmer Word Segmentation such as Word Bi-gram and Orthographic Syllable Bi-gram in [4], Statistical Analysis with Linguistic Rules in [1], Bi-Directional Maximal Matching in [2], and Constrained Conditional Random Fields in [3] have been investigated. However, most of the approaches above cannot efficiently deal with the out-of-vocabulary words. The purpose of this study is to increase an accuracy of Khmer Word Segmentation by using Maximum Matching Algorithm along with grammatical based rules.

II. Related Works

A great deal of studies related to word segmentation has been investigated and they are often done in most Asian languages such as Chinese, Thai, Burmese, Javanese, Vietnamese, Lao and Khmer.

Word Segmentation of Khmer written text based on Bi-gram model was presented in [4]. The characters in each sentence were merged into combination of characters called Khmer Character Cluster (KCC). Then, KCC matching module read each KCC one by one from left to right and matched them. Then, it converted the KCCs into KCE string. The KCE string was used to look up whether it existed in the dictionary. Multiple possible segmentations of the input text were generated. Finally, disambiguation module was used to decide the most appropriate segmentation among the list of candidates.

Word segmentation proposed in [2] was done based on the dictionary by using Bi-Directional Maximal Matching and some rules such as Khmer Character Clusters (KCC) and Khmer Unicode Error Correction (KUEC). With these dictionary-based segmentation approaches, they achieved the accuracy of 98.13% and spent 2.581 seconds for 160,000 0f Khmer words.

Constrained Conditional Random Fields (CCRF), the model to improve the performance of word segmentation, part of speech tagging, and name entity recognition (NER) was presented in [3]. First, the added OOV entity words in user

978-1-5386-4015-9/18/$31.00 @2018 IEEE

dictionary were defined as constrained. Then, CCRF model was applied to segment and tag the part of speech. And then, NER-CCRF was used to recognize named entity. With this method, Khmer Word Segmentation achieved precision of 90.78% and recall of 91.341%, and name entity recognition with 86.07% on precision, 86.54% on recall, and 86.30% on F-measure.

[1] proposed a rule-based technique obtained by statistical analysis as well as specific linguistic rules of Khmer to tackle the issues of OOV words, compound words, proper name, derivative words and new words. Using this model, Khmer Word Segmentation could achieve 77.70% on precision, 75.55% on recall, and 76.50% on F-measure.

III. PROPOSED SOLUTION

This study proposes to improve Maximum Matching algorithm in Khmer Word Segmentation with Rule-based method. The process of segmentation is shown in Fig. 1. First of all, whitespace, invisible space which can cause problem to word separation in each sentence were eliminated. There are no specific rules for using whitespace and it is impossible to see invisible space which is provided by Khmer Unicode. Then word segmentation was done based on the Khmer Corpus. This step made the boundary of words in each sentence into single words based on the Longest Matching approach. And then, the unknown words were then defined and solved by using 21 grammar rules created.

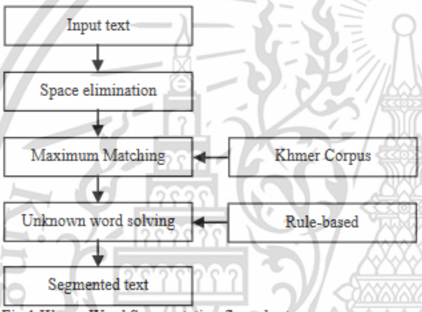


Fig.1 Khmer Word Segmentation flow chart

A. *Maximum Matching*

Maximum Matching algorithm is one of the most popular and a powerful word segmentation technique used as a baseline method in word segmentation [5]. In order to do word separation, this method compares and finds the longest matched word from the dictionary. If all the words in a sentence matched with the dictionary, the result is obtained as shown in result 1 and result 2 as examples. If the words in each sentence do not exist in the dictionary, they are defined as the unknown word.

Sentence 1: ដោយសារតែក្ដីអាណិតនិងស្រលាញ់គាត់បានផ្ដល់កន្លែងស្នាក់នៅព្រមទាំងលុយកាក់សម្រាប់នាងបន្តការសិក្សានៅភ្នំពេញ

Result 1: ដោយសារតែ/ក្ដីអាណិត/និង/ស្រលាញ់/គាត់/បានផ្ដល់/កន្លែងស្នាក់នៅ/ ព្រមទាំង/លុយកាក់/សម្រាប់/នាង/បន្ត/ការសិក្សា/នៅ/ភ្នំពេញ

(Because of pity and love he has supported her accommodation and money for her study in Phnom Penh)

Sentence 2: នាយករដ្ឋមន្ត្រីអង់គ្លេសទិទៀនបក្សប្រជាជាតិថាចង់បំបែកស្កត់លែន ចេញពីអង់គ្លេស

Result 2: នាយករដ្ឋមន្ត្រីអង់គ្លេស/[ទិទៀន]/បក្សប្រជាជាតិ/ថា/ចង់បំបែក/ស្កត់លែន/ចេញពី/អង់គ្លេស

(The England Prime Minister criticized that the Nation Party want England and Scotland break up)

In result 2, there is an unknown word defined in [] symbol ([ទិទៀន]). This unknown word will be solved in the next step.

B. *Solving the Out-of-vocabulary Problem*

Most common Khmer words are borrowed from Pali-Sankrit. We can recognize them based on their characteristics or grammar rules.

To solve the unknown word (UW) problem, 21 rules were created based on the principle of Khmer grammar. However, the words which are taken from other languages, it may be out of rules, most of the 21 rules are not efficient to deal with them. We built the rules by counting and comparing the characters and the order of dependent vowel (DV), independent vowel (IV), consonant (C), and special character and symbol (S). These rules are called character combination and divided into 3 groups. Group 1 (Rule 3-5) is the group of no DV, Group 2 (Rule 6-11) is the group of one DV, Group 3 (12-21) is the group of two or more DV. When an unknown word matches with the rules, the new word will be created and the problem of out-of-vocabulary is solved. In the S case, we only focused on some special characters such as "់ ៉ ្ ៊ ៍ ៌ ៏ ៑ ះ".

Rule 1: No English or unknown characters

Rule 2: No numbers

Rule 3: If UW contains some C (2 to 8) with zero, one or two S (្), these characters will be combined as a new word. Examples: ភព កក បក ដប ចប នរក សកល ក្រ ខ្យង សម្ល វត្ត ហង្សស្អក ប្រឈម មណ្ឌល ពលរដ្ឋ សម្ពស្ស កម្មករ បង្កក បង្កង ទ្រព្យធន ប្រភព ប្រកប ក្រចក...

Rule 4: If UW contains one S at the end and same rule as Rule 3, these characters will be combined as a new word. Examples: កក់ រស់ ទន់ អន់ ឡប់ ទល់ ដប់ សក់ បក់ លក់ ពណ៌ ដ៏ រយៈ ខណៈ មរណៈ វចនៈ របស់ ខ្សត់ ឆន្ទៈ ផ្តលក់ វេប្បធម៌...

Rule 5: If UW contains a/two IV (s) and same rule as Rule 3 or 4, these characters will be combined as a new word. Examples: ឧត្តម ឥស្សរជន ឯកជន ឯករចនៈ ឋានៈ ឧស្ម័ន ឧបករណ៍ ឧទ្ទេសណ៍ ...

Rule 6: If UW contains a DV with zero, one or two S (s), starts with a C and ends with a C, these characters will be combined as a new word. Examples: សមាជ សុខ នាគ ពោគ ប្រមុខ ប្រយោគ សម្ភោធ ប្រាកដ ថៅក្រម ប្រភេទ អាល័យ វិស័យ អនាម័យ និស្ស័យ ចិច ប៉ូវ...

Rule 7: If UW contains a DV with one, two or three S (s), starts with a C and ends with a S, these characters will be combined as a new word. Examples: ប្រយោជន៍ ប្រសាសន៍ ព្រាហ្មណ៍ អភិវឌ្ឍន៍ ព្យាករណ៍ ស្នេហ៍ ភេសជ្ជៈ កុម្ភៈ សក្ការៈ ខេមរៈ...

Rule 8: If UW contains a DV with zero, one or two S (s), starts with a C and ends with a DV, these characters will be combined as a new word. Examples: សេះ ហោះ គោ កោះ ទា លាភ ដី ប៉ះ ខ្ញុំ ស៊ី ញ៉ាំ ជ្រោះ កោះ ទា លាភ ដី ស្រី ស៊ាំ ចក្រី កវី សេដ្ឋកិច្ច ផ្សំ ខ្នុះ ...

Rule 9: If UW contains a DV with zero, one or two S (s), starts with a IV and ends with a C, these characters will be combined as a new word. Examples: ឥស្សរ ឧបាសក ឥណទាន ឥឡូវ ឱនភាព ឪពុក ឧទាន ឯកការ ឯកសារ ឧបាយកល ឧទ្ទេស ឧដុង្គ ឧបកិច្ច ...

Rule 10: If UW contains a DV with zero, one or two S (s), starts with a IV and ends with a DV, these characters will be combined as a new word. Examples: ឧបមា ឯណា ឧកញ៉ា ឯកតោ ឥច្ឆា ឥសី ឫកពា ឫស្សី ឧស្សា...

Rule 11: If UW contains a DV with one or two or three S (s), starts with a IV and ends with a DV, these characters will be combined as a new word. Examples: ឧទាហរណ៍ ឧទ្ធរណ៍ ឧស្សាហ៍ ឧបករណ៍ ឋានៈ...

Rule 12: If UW contains two DVs, two Cs with zero, one or two S (s), and C1 = C2, these characters will be combined as a new word. Examples: រុះរើ ឡូឡា មូម៉ៅ នានា ឆាឆៅ ញ៉េះញ៉ោះ ពុំអ៊ែ ឡេះឡោះ ទូទៅ អេះអុះ ឆាំឆា ដោះដៃ ងេងើ...

Rule 13: If UW contains two DVs, three Cs with zero, one or two S (s), and C1 = C2 or C1= C3, these characters will be combined as a new word. Examples: ដុះដាល ឆ្អឹរ ទូទាត់ អែអង់ ពាំង ជុំជិត ចែចូវ មាំមួន ធំធេង យំយែក ហែកហួរ...

Rule 14: If UW contains two or four DVs, four Cs with zero, one or two S (s), and C1= C3, these characters will be combined as a new word. Examples: ចាកចេញ តិចតួច គិតគូរ ស៊ើបសួរ ដុនដាប តាក់តែង ថ្លៃថ្លា យឺតយ៉ាវ លូកលាន់ វៀចវេរ ហុងហែង កៀកកើយ មាយីមាយា...

Rule 15: If UW contains two or four DVs, five Cs with zero, one or two S (s), and C1= C3 or C1 & C2 = C3 & C4 or C1 & C2 = C4 & C5, these characters will be combined as a new word. Examples: ស្អុកស្អំ គំរោះគំរើយ ត្រៀបត្រា ខ្នះខ្នែង ផ្លុះផ្លាយ ក្រៀមក្រំ ខ្ទះខ្ទាយ ស្វះស្វែង ឆ្មើងឆ្មៃ ឆ្គាំឆ្គង ញញើញញ៉ីរ ផ្លុះផ្លាយ នៃបនិត្យ សាបសូន្យ ខ្ទេចខ្ទី ផ្តុំផ្តើង ស្តីមស្តៃ...

Rule 16: If UW contains two or four DVs, six Cs with zero, one two or three S (s), and C1 & C2 = C4 & C5, these characters will be combined as a new word. Examples: ក្លែងក្លាយ ឡើបឡៀវ ខ្ទាត់ខ្ទែង ជ្រោមជ្រែង ញញឹមញញែម ចំតិតចំតូង ផ្ដើងផ្ដាន ទទឹងទទែង ផ្ដេសផ្ដាស របៅតរប៉ាយ ផ្កាញ់ផ្កាល រនេករនាក...

Rule 17: If UW contains two or four DVs, seven Cs with zero, one, two or three S (s), and C1 & C2 = C4 & C5 or C1 & C2 = C5 & C6, these characters will be combined as a new word. Examples: បង្គិតបង្គុំ អន្ទះអន្ទែង ស្រមោលស្រមៃ បញ្ចើចបញ្ចើ ក្រមុកក្រាន្ត បន្ទាប់បន្សំ ស្រងេះស្រងោច បន្លែបន្លុក...

Rule 18: If UW contains two or four DVs, seven Cs with zero, one, two or three S (s), and C1 & C2 & C3 = C5 & C6 & C7, these characters will be combined as a new word. Examples: បន្តិចបន្តួច ប្រកៀកប្រកិត ប្រញាប់ប្រញាល់ បំផ្លិចបំផ្លាញ ចម្រុងចម្រើន ចម្រូងចម្រាស ប្រងេងប្រងាង...

Rule 19: If UW contains two DVs (ា), zero or one S with two, three or four Cs, these characters will be combined as a new word. Examples: មាគ៌ា មាតា សាវតា អាហារ ភាសា សាសនា អាត្មា វាចា យាត្រា ពាលា សាវតា ការពារ ទាយាទ ទាហាន...

Rule 20: If UW contains two or three DVs (ា ិ ី ុ), zero or one S with two, three or four Cs and ends with a DV, these characters will be combined as a new word. Examples: នាទី សមាធិ ភាគី មាលតី នាឡិកា សាលាក្តី កុមារី សុជាតា បរិញ្ញា ថវិកា កិរិយា សីលា និរតី អធិបតី ពិធី កីឡា បុត្រា មាតុភូមិ ...

Rule 21: If UW contains two DVs (េ & ា or េា & ា), zero or one S with two or three Cs and ends with a DV (ា), these characters will be combined as a new word. Examples: មេធា សេនា ចេនឡា ទេសនា វេលា មេឃា ទេវតា ចេតនា ជោហា ហោរា មេត្តា ជេស្ឋា យោធា…

IV. EXPERIMENTS AND RESULTS

A. Experiment Setup

Various online contents and books from agriculture, magazine, newspaper, health, technology and history as shown in Table II were used to create a manually corpus and a collection of sentences. With randomly 395 articles selected, we could build an annotated corpus with 19500 unique words and 2018 testing sentences (64010 words). The known words and unknown words inside the testing sentences are approximately 95.48% and 4.52%.The corpus and the collections of sentences were then used to do the experiment with the algorithms proposed.

TABLE II. TYPE AND NUMBERS OF ARTICLES

No	Article Type	# Articles
1	Agriculture	38
2	Magazine	20
3	Newspaper	220
4	Technology	40
5	Health	25
6	History	52
	Total	395

B. Results

The results of word segmentation were counted by each sentence and were separated into two cases. The first case is

the result for word segmentation based on dictionary and the second case is for solving the unknown known words.

Each article has its different accuracy as shown in TABLE IV. Using Maximum Matching algorithm alone, we got the accuracy of 88.34%, 89.28%, 88.17%, 90.36%, 92.12% and 87.38% on agriculture, magazine, newspaper, technology, health and history respectively. Then we compared the results of the Maximum Matching with Rule-based by using the same datasets. As shown in TABLE IV, the accuracy in every article-tested increased from 2.68% to 4.76%.

TABLE III. COMPARION OF PRECISION, RECALL AND F-SCORE

Article	Maximum Matching (MM)			Maximum Matching with Rule-based		
	P	*R*	*F*	*P*	*R*	*F*
Agriculture	91.55	95.91	93.67	95.27	95.91	95.58
Magazine	93.39	95.19	94.28	95.23	96.15	95.68
Newspaper	91.02	96.50	93.67	95.15	97.18	96.15
Technology	91.92	98.01	94.86	94.90	98.67	96.74
Health	94.30	97.47	95.85	98.28	96.63	97.45
History	90.75	95.81	93.21	94.21	96.51	95.34
Mean	91.43	96.46	93.87	95.19	97.00	96.08

TABLE IV. COMPARISION OF ACCURACY

Article	Maximum Matching	Maximum Matching with Rule-based	% increase
Agriculture	88.34	92.02	3.68
Magazine	89.28	91.96	2.68
Newspaper	88.17	92.93	4.76
Technology	90.36	93.97	3.61
Health	92.12	95.27	3.15
History	87.38	91.48	4.10
Mean	88.55	92.81	4.26

V. CONCLUSION

In this paper, a new Khmer Word Segmentation technique has been developed. This method was divided into two steps. First, we did the segmentation based on the dictionary by using Maximum Matching algorithm and then, the Rule-based approach was used to solve the problem of words that does not exist in the dictionary. There are 21 rules and they were created based on the principle of Khmer grammar books [8] and [9]. With the proposed solution, we got the accuracy of 88.55% in word segmentation based on the corpus and it increased to 92.81% when the rules were applied.

REFERENCES

[1] Channa Van and Wataru Kameyama, "Khmer Word Segmentation and Out-of-Vocabulary Words Detection Using Collocation Measurement of Repeated Characters Subsequences," Graduate School of Global Information and Telecommunication Studies, Waseda University, 2013.

[2] Narin Bi and Nguonly Taing, "Khmer Word Segmentation based on Bi-Directional Maximal Matching for Plaintext and Microsoft Word Document" Royal University of Phnom Penh, Cambodia, 2014.

[3] Shuhui Huang, Xin Yan and Qingling Lei, "Construction of Khmer Entity Annotation Corpus Based on Constrained Conditional Random Fields," Kunming, China, 2016.

[4] Chea Sok Huor, Top Rithy, Ros Pich Hemy and Vann Navy, "Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation," PAN Localization Cambodia, 2006.

[5] Tan Yuan Liu and Kan Xu Shen, "The Word Segmentation Methods for Chinese Information Processing Hua University Press and Guang Xi Science and Technology Press, 1994.

[6] Channa Van and Wataru Kameyama, "Query Expansion for Khmer Information Retrieval," Proceedings of the 8th Workshop on Asian Language Resources, Beijing, China, 2010.

[7] Tran Van Nam, Nguyen Thi Hue and Phan Huy Khanh, "Building a Syllable Database to Solve the Problem of Khmer Word Segmentation," Department of Computer Engineering, Polytechnic University of Da Nang, Vietnam, 2017

[8] Thon Hin, "Khmer Grammar for Phumseuksa," Phnom Penh, Cambodia, 2011.

[9] Chhorn Chheang, "Khmer Grammar for General Studies," Phnom Penh, Cambodia, 2002.

AUTHOR BIOGRAPHY

Author:	Mr. Pakrigna Long
Degree:	Master of Engineering
Date of Graduation:	17th July 2018
Date of Birth:	13th April 1993
Place of Birth:	Kampong Cham, Cambodia

Undergraduate and Graduate Education:

Master of Engineering in Computing in Engineering Systems, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 2018

Bachelor's degree in informatics engineering, STMIK IKMI Cirebon, West Java, Indonesia, 2015

Major: Computing in Engineering Systems

Presentations and Publications:

[1] Long, P. and Boonjing, V."Khmer Word Segmentation and Rule-based Techniques for Khmer Word Segmentation", The 2018 - 10th International Conference on Knowledge and Smart Technology, Jan 31- Feb 3, 2018,Chiang Mai, Thailand, 978-1-5386-4015-9/18/$31.00 @2018 IEEE.