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 ប្រ 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 ។ ៕ () ◌ៈ ? ! ៖ ៘ ។ល។ . , ៙៚ " /

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.

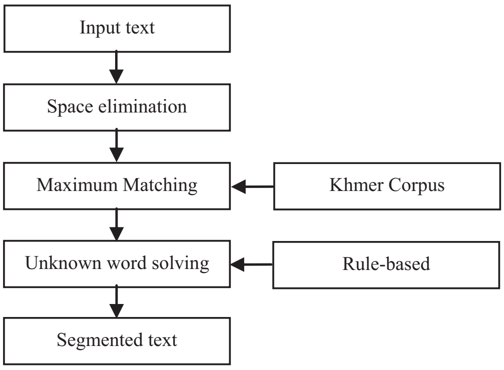

Fig.1 Khmer Word Segmentation flow chart

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: 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: 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 ហង្ស ស្កក ប្រឈម មណ្ឌល ពលរដ្ឋ សម្ផស្ស 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 ជ័រ រយៈ ខណៈ មរណៈ វចនៈ របស់ ខ្សត់ ឆន្ទៈ ជ្រលក់ វប្បធម៌...

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 ឯកវចនៈ ឋានៈ ឧស្ម័ន ឧបករណ៍ ឧទ្វេណ៍ ...

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 រោគ ប្រមុខ ប្រយោគ សម្ពោជ ប្រាកដ ចៅក្រម ប្រភេទ អាល័យ វិស័យ អនាម័យ និស្ស័យ ប៊ិច 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 រសជាតិ ផ្សំ រន្ទះ ...

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: ឥស្វរ 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 ឥដ្ឋា 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 ឧទ្ទរណ៍ 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 មូមៅ នានា ឆាឆៅ ញ៉េះញ៉ោះ រ៉ាំរ៉ៃ 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 អ៊ូអរ ទូទាត់ អែអង់ រារាំង 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 ហួងហែង 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 ខ្ជះខ្ជាយ ស្វះស្វែង ឆ្អើងឆ្អែ ឆ្គាំឆ្គង ញញីញញ័រ ផ្លុះផ្លាយ នៃបនិត្យ សាបសូន្យ ខ្ទេចខ្ទី ថ្កុំថ្កើង ស្គឹមស្គែ...

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 ខ្សឹបខ្សៀវ ខ្វាត់ខ្វែង ជ្រោមជ្រែង ញញឹមញញែម ចំតិតចំតូង ថ្កើងថ្កាន ទទឹងទទែង ផ្តេសផ្តាស រប៉ាតរប៉ាយ ផ្ងាញ់ផ្ងាល រខេករខាក...

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: បង្ខិតបង្ខំ អន្ទះអន្ទែង ស្រមោលស្រមៃ បញ្ជើចបញ្ជើ ក្សេមក្សាន្ត 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 ច្រងេងច្រងាង...

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: មេបា សេនា ចេនឡា ទេសនា វេលា មេឃា ទេវតា ចេតនា វោហា 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.