Syllable Segmentation Algorithm for Myanmar Language

Cho Cho Hnin[1], Naw Naw[2]

[1]Department of Information Science, University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
hninhnin288@gmail.com

[2]Department of Information Secience, University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
dr.nawnaw@utycc.edu.mm

Abstract: ***Myanmar language does not have word boundary or white space between words. Thus, it is problematic to tokenize these words into the meaningful words before the process of text mining. There are many word segmentation methods using Unicode standard encoding for Myanmar language. Many people still use Myanmar Zawgyi-One font especially in social media contents. Thus, it is relevant to focus on the word segmentation for different social media text mining. Some Myanmar informal texts, especially in social media contents, can contain English words among Myanmar words. In such case, it is necessary to segment both of Myanmar words and English words from these mixed informal texts. This paper proposes a syllable segmentation algorithm for Myanmar text(Zawgyi-One Standard) and for the text with the combination of Myanmar words and English words.***

Keywords: Natural language processing, Myanmar language, word segmentation, syllable segmentation**.**

1. Introduction

Word segmentation is a major pre-processing task for text mining in natural language. It is the process of determining the boundaries of words in the text. It is also necessary before the tasks of text mining such as machine translation, information retrieval, opinion mining, sentiment analysis and other social media text mining. As the Myanmar language does not have a delimiter between words, it is difficult for the computer to determine word boundaries for Myanmar syllable which is composed of several characters.

There are many word segmentation methods for different encoding schemes. Among these, Zawgyi-One is widely used in Myanmar Web documents because many web developers in Myanmar still develop their web contents using this one. Moreover, most of the internet users are familiar with Zawgyi-One keyboard layout, and they do not want to learn unfamiliar ones. These facts are the reason to propose our segmentation work for Myanmar language.

This paper is organized as follows. Related works are outlined in Section 2 and the structure of Myanmar language is presented in Section 3. Section 4 describes our proposed system for Myanmar word segmentation including an algorithm for syllable segmentation. The evaluation for this system is presented in Section 5. Finally, we conclude this paper in Section 6.

2. Related Works

The [9], a rule-based approach for syllable segmentation for Myanmar text is proposed. They created segmentation rules based on the syllable structure of the Myanmar script. This segmentation task focus on the UTN11-2 encoding model for Myanmar text. But, they did not considered for the characters of non-Myanmar scripts. In [2], the rule-based heuristic approach and statistical approach are used for analysis of Myanmar word boundary and segmentation. They applied the deterministic finite state automata(DFA) to tokenize a word boundary for Myanmar language. They also used the corpus-based dictionary and lexicon database.

T. T. Thet et al, in [8], developed a word segmentation approach for Myanmar language using Unicode standard encoding. They also used a rule-based heuristic approach for syllable segmentation and a dictionary-based statistical approach for syllable merging. After testing their system on 16 document, they achieved the precision, recall and F-measure above 98%.

In [6], the authors proposed a POS-based word-splitting algorithm for Thai word segmentation which splits words in order to increase POS tags. They used the conditional random fields (CRF) model in order to achieved remarkably accurate segmentation. They also introduced a dictionary-based word-merging algorithm, which merges all kinds of compound words. They demonstrated their methods with three applications.

A. Klahan et al, in [1], proposed Thai word safe segmentation algorithm using dictionary to solve the problem of obtaining unreasonable search results of creating the inverted index. They evaluated their segmentation method with several implementations called Safe Analyzer. According to the experimental results, the linked-list Trie and Protostuff library gave the outstanding results. Their segmentation method could not solve the misspell within text accurately although it can definitely solve the ambiguity problem.

The authors in [4] proposed a method for Chinese word segmentation based on conditional random fields (CRF) with character clustering. They used two different clustering algorithm, K-means and Brown clustering algorithm, to get the clusters of character embedding. Their system achieves and F-score of 95.67%.

In [3], the authors describes a method to solve the segmentation problem. They used a dictionary-based approach to segment the text by applying the Maximum Matching algorithm to segment the text forwards (FMM) and backwards (BMM). Then, they applied SVM classifier to define the word boundaries based on the difference between them. Their system achieve an F-measure of 99.0 for overall segmentation.

For Vietnamese word segmentation, T. P. Nguyen and A. C. Le [7] proposed a hybrid approach to detect word boundary. They used logistic regression as a binary classifier combining with longest matching algorithm. Their system achieved as F-measure of 98.82%. In [5], the authors developed a model for Japanese word segmentation and POS tagging in the microblog. Their model with lexical normalization could handle the orthographic diversity of microblog text. But, it cannot handle certain types of ill-spelled words and spelling errors.

3. Structure of Myanmar Language

Myanmar language, also called Burmese language, the official language of Myanmar (Burma), spoken as a native language by the majority of Burmans and as a second language by most native speakers of other languages in the country. A Myanmar alphabet is written from left to right and requires no space between words, although modern writing usually contains spaces after each clause to enhance readability.

Myanmar characters can be classified into three groups: consonants, medials and vowels. The Burmese script has 33 letters to indicate the initial consonant of a syllable as shown in Table 1. There are 4 medials in the medial group and eight vowels in the dependent vowel group. Myanmar Sign Asat is used over any of the syllable-final consonants when no stacking takes place(eg. ဘယ်). There are two punctuation marks in Myanmar script(" ၊ " and " ။ "). The various signs and symbols of Myanmar scripts are shown in Table 2.

Table 1: Myanmar Consonants

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

Table 2: Myanmar Scripts

Name	Examples
Medials	ျ , ြ , ွ , ှ
Dependent Vowel Sign	-ါ , -ာ , -ိ, -ီ, -ု , -ူ , ေ , -ဲ
Various Signs	-ံ , -့ , -း
Independent Vowels and Independent Various Signs	ဤ ဧ ဪ ၌ ၍ ၏ ဣ ဥ ဿ ၎
Myanmar Digits	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉
Punctuation Marks	၊ ။

3.1 Syllable Structure

A word can contain one or more syllables. A syllable can comprise multiple consonants, multiple medials, multiple vowels and various signs. These can appear in different orders to form a syllable. For example, a word "စားသောက်ဆိုင်"(restaurant) is composed of a sequence of syllables, " စား+သောက်+ဆိုင် ". The syllable "စား" ("စ" + " ာ" + " း ") includes a consonant "စ" followed by a vowel " ာ" followed by a sign called Visarga " း ". Firstly, it is necessary to segment the syllable boundary for each sequence of characters. And then, the correct boundary of word could be defined by combining the relevant sequence of syllables. Thus, the task of syllable segmentation greatly impacts that of word segmentation. Our proposed work mainly focuses on the syllable-level segmentation for both formal text and informal text.

4. Myanmar Word Segmentation

The main pre-processing task of natural language processing is the word segmentation . The overall process of our proposed system is presented in Figure 1. The text file with Myanmar Zawgyi standard is accepted as an input. Firstly, the input texts are divided into the individual sentences by using Myanmar sentence delimiter " ။ " as " . " in English sentence. Then, two main processes: syllable segmentation and syllable merging are performed to segment Myanmar text. The system outputs the segmented texts that are ready to use in different text mining.

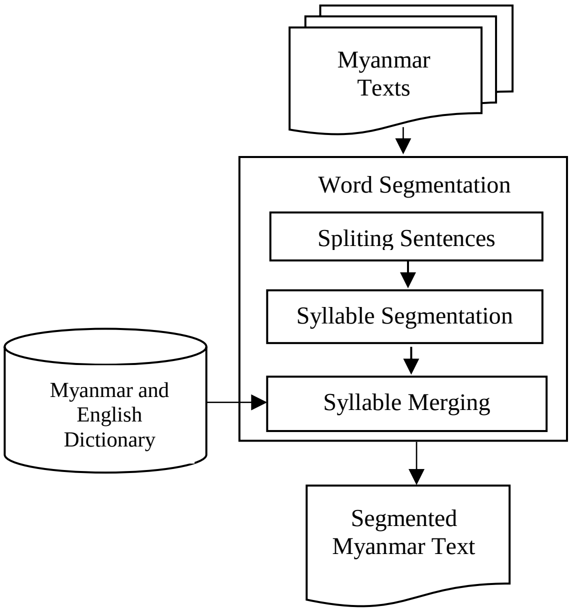

Figure 1: Proposed system design for Myanmar word segmentation

4.1 Syllable Segmentation

```
Input: Myanmar Text Document "D"
Output: Set of Segmented Syllables "Seg"
1. Seg = {   }
2. for each sentence S in document D do
3.    for each character $c_i$ in S do
4.       if ((($c_i \in$ mainAlphabet) && ($c_{i+1} \neq$ " ် " )&&
               ($c_{i-1} \notin$ prefixAlphabet )) ||
5.          (($c_{i-1} \in$ prefixAlphabet) && ($c_{i+1} \neq$ " ေ ")) ||
6.          ($c_i \in$ singleAlphabet) ||
7.          (($c_i \in$ EngAlphabet) && ($c_{i-1} \notin$ EngAlphabet))) then
8.             Seg = Seg + "-" +$c_i$
9.     else if ( ($c_i$ = = " ။ ") ||  ($c_i$ = = " ၊ ")) then
10.            Seg = Seg +"-" +$c_i$
11.    end if
12.   end for
13. end for
14. Output : Set of Segmented Syllables
```

The algorithm elaborates the process the syllable segmentation for Myanmar language. With the provided set of Myanmar text document set and the several segmentation rules, the output of this algorithm is a set of segmented syllables Seg.

In the above algorithm, Seg represents the list of segmented syllables and is also the output syllables produced by the algorithm. The mainAlphabet is for Myanmar consonants, the prefixAlphabet is the list containing the medials (" ြ " , " ြ " , " ြ " , " ြ ") and a dependent vowel sign (" ေ "), the singleAlphabet is the list of the independent vowels and independent various signs. The EngAlphabet is for English letters (A – Z and a – z).

The algorithm starts with the reading of Myanmar sentences and then searches for the delimiter where the set of syllables will have to segmented. The sample input Myanmar text before performing any segmentation process is shown in Table 4. Firstly, the program determines whether or not the character is a member of the main alphabets/consonants. If it is true and the next alphabet is Myanmar sign Asat " ် " and the previous alphabet is not a member of the prefixAlphabets, then the program defines a boundary before the current pointing character as a segmented syllable. If the first condition is not true, the program checks another conditions until a syllable boundary is identified. Finally, the algorithm produces the segmented syllables of input documents. After the syllable segmentation, the segmented text for the input sentences for Table 5 is shown in Table 6.

4.2 Syllable Merging

The segmented syllables obtained from the above syllable segmentation phase are merged into the meaningful words by using a maximum matching approach. This task is performed by matching the longest substring in the input sentence with Myanmar dictionary for Myanmar words and English dictionary for English words. The segmented words after the merging phase is described in Table 6.

Table 4: Sample Myanmar Text before Syllable Segmentation

ဧည့်ခန်း လှပပြီးခံ့ညားတယ်။ (ဧ+ည+်+ ့+ခ+န+်+း+ +လ+ှ+ပ+ြ+ီ+း+ခ+ံ+ ့+ည+ာ+း+တ+ယ+်+။) (The lobby is grand and elegant.)
ဝန်ဆောင်မှုနဲ့ အပြင်အဆင် ကောင်းပါတယ်။ (ဝ+န+်+ေ+ဆ+ာ+င+်+မ+ှ+ု+န+ဲ+ ့+ +အ+ြ+ပ+င+်+အ+ဆ+င+်+ +ေ+က+ာ+င+်+း+ပ+ါ+တ+ယ+်+။) (The service and decoration is good.)
အစားအစာ အရမ်းကောင်းပါတယ်။ (အ+စ+ာ+း+အ+စ+ာ+ +အ+ရ+မ+်+း+ေ+က+ာ+င+်+း+ပ+ါ+တ+ယ+်+။) (The foods are very delicious.)
အင်တာနက်လိုင်း သိပ်မကောင်းဘူး။ (အ+င+်+တ+ာ+န+က+်+လ+ိ+ု+င+်+း+ +သ+ိ+ပ+်+မ+ေ+က+ာ+င+ ်+း+ဘ+ူ+း+။) (The internet connection is not so good.)

Table 5: Sample Myanmar Text after Syllable Segmentation

ဧည့်-ခန်း-လှ-ပ-ပြီး-ခံ့-ညား-တယ်။
ဝန်-ဆောင်-မှု-နဲ့- အ-ပြင်-အ-ဆင်-ကောင်း-ပါ-တယ်။
အ-စား-အ-စာ-အ-ရမ်း-ကောင်း-ပါ-တယ်။
အင်-တာ-နက်-လိုင်း- သိပ်-မ-ကောင်း-ဘူး။

Table 6: Sample Myanmar Text after Syllable Merging

ဧည့်ခန်း-လှပ-ပြီး-ခံ့ညား-တယ်။
ဝန်ဆောင်မှု-နဲ့- အပြင်အဆင်-ကောင်း-ပါ-တယ်။
အစားအစာ-အရမ်း-ကောင်း-ပါ-တယ်။
အင်တာနက်လိုင်း- သိပ်-မ-ကောင်း-ဘူး။

4.3 Experiments

We implements our work by using the collected Myanmar texts which are no segmented text. These documents includes many informal sentences because they are user reviews obtained from the hotel sites. Consider the following sentence, " ဒီဟိုတယ်မှာ ဧည့်ခန်း လှပပြီးခံ့ညားတယ်။" (The lobby in this hotel is grand and elegant). The syllable segmentation and word segmentation processes of this sample sentence are shown as an example in Table 7.

Table 7: Example of Segmentation for Only Myanmar Text

Input Text:	ဒီဟိုတယ်မှာ ဧည့်ခန်း လှပပြီးခံ့ညားတယ်။
Sequence of Syllables:	ဒ+ီ+ဟ+ိ+ု+တ+ယ+်+မ+ှ+ာ+ +ဧ+ည+်+ ့+ခ+န+်+း+ +လ+ှ+ပ+ြ+ီ+း+ခ+ံ+ ့+ည+ာ+း+တ+ယ+်+။
Segmented Syllables:	ဒီ-ဟို-တယ်-မှာ-ဧည့်-ခန်း-လှ-ပ-ပြီး-ခံ့-ညား-တယ်-။
Segmented Words:	ဒီ-ဟိုတယ်-မှာ-ဧည့်ခန်း-လှပ-ပြီး-ခံ့ညား-တယ်-။ (This+hotel+in+the lobby+grand+and+ elegant+is+.)

In the above example, the first syllable "ဒီ" is composed of two characters(ဒ+ီ), one consonant and one dependent vowel. The next followed syllable is "ဟို" which is composed of three characters(ဟ+ိ+ု). Since the character "ဒ" is a member of the main alphabet and the next one " ီ " is not Asat (်), it will be segmented in front of "ဒ" according to the line 4 in the algorithm. Then, it analyzes whether the position between "ဒ" the next character " ီ" is segmented or not. But, it is not necessary to segment this position because there is no match any condition in the algorithm. For the character "ဟ", the condition at line 4 is true. Thus, the position before the character "ဟ" is segmented as a correct syllable. The two characters("ဒ" and " ီ ") is defined as a correct syllable("ဒီ"). The process of syllable segmentation is continued until the remaining characters in the sentence are segmented as the same. For the above example sentence, there are 13 segmented syllables produced from the syllable segmentation algorithm.

After the syllable segmentation phase, the segmented syllables are merged to form a meaningful words. These pairs of syllables are merged by matching with the Myanmar dictionary. Finally, there are 7 words obtained by performing the maximum matching approach for the sentence in Table 7.

Some texts often contain English words among Myanmar words. In this case, the segmentation process have to performed for both Myanmar texts and English texts. The segmentation process for the combination of Myanmar and English texts is shown in Table 8 as an example. The first two characters("ဒ +" ီ") are checked as the above example sentence shown in Table 7. The next character "h" is an English character and the previous character " ီ" is not an English alphabet. The position before the character "h" is identified as a syllable boundary according to line 7. Thus, the syllable "ဒီ" is defined as a segmented syllable. Then, the program checks whether the position after the character "h" is segmented or not. It is not needed to segment between "h" and "o" as there is no match any condition in the algorithm. After checking the next characters, there are 11 segmented syllables. There are 6 Myanmar words and 2 English words after merging the segmented syllables as shown in Table 8.

Table 8: Example of Segmentation for Myanmar and English Text

Input Text:	ဒီhotelမှာ decoration နဲ့ ဝန်ဆောင်မှု ကောင်းတယ်။ (Decoration and service in this hotel is good.)
Sequence of Syllables:	ဒ+ီ+h+o+t+e+l+မ+ှ+ာ+ +d+e+c+o+r+a+t+i+o+n+ +န+ဲ + ့ + ဝ+န+်ေ+ဆ+ာ+င+် +မ+ှ+ု+ေ+က+ာ+င+ ်+း+တ+ယ+်+။
Segmented Syllables:	ဒီ+hotel+မှာ+decoration+နဲ့+ဝန်+ဆောင်+မှု+ကောင်း+တယ်+။
Segmented Words:	ဒီ+hotel+မှာ+decoration+နဲ့+ဝန်ဆောင်မှု+ကောင်း+တယ်+။ (This+hotel+in+decorations+and+service+good+is+.)

5. Performance Evaluation

To measure the performance of the segmentation process in the proposed system, the segmented words are compared with ground truth segmented words that are manually read and segmented. The accuracy of Myanmar syllable segmentation process is 99% and that of word segmentation process is 96% on 25000 manually segmented words in the Myanmar text document. The accuracy of syllable-level segmentation is higher than that of word-level segmentation because some words such as names and pronouns, and some idioms that are not in the dictionary and they are rarely used.

6. Conclusions

The proposed system presents an approach with a new algorithm for syllable segmentation using Zawgyi-One encoding for Myanmar language. The syllable segmentation task is developed based on the structure of the Myanmar scripts. There are still some limitations in the phase of word segmentation as it depends on the dictionary in merging the syllable segments. But, this proposed work is effective for syllable segmentation process which is our main emphasized work, because of the 99% accuracy. Especially, our work will be an effective preprocessing task for social media text mining in Myanmar language without converting the texts written with Zawgyi-One to Myanmar Unicode texts. As a future work, testing and evaluation on a larger data set using a large corpus and more complete dictionary for better accuracy.

References

[1] A. Klahan, S. Pannoi, P. Uewichitrapochana, R. Wiangsripanaw, "Thai Word Safe Segmentation with Bounding Extension for Data Indexing in Search Engine," in Information and Communication Technology, pp. 83-92, 2018.

[2] A. M. Mon, M. M. Thein, S. S. Htay, S. L. Phyue and T. T. Win, "Analysis of Myanmar Word Boundary and Segmentation by using Statistical Approach", in 3rd International Conference on Advanced Computer Theory and Engineering(ICAETE), 2010.

[3] C. L. Goh, M. Asahara and Y. Matsumoto, " Chinese Word Segmentation by Classification of Characters", in the association for Computation Linguistics and Chinese Language Processing, pp. 381-396, 2005.

[4] L. Du, X. Li, C. Liu, X. Fan, J. Yang, D. Lin and M. Wei, "Chinese word segmentation based on conditional random fields with character clustering", International Conference on Asian Language Processing(IALP), 2016.

[5] N. Kaji and M. Kitsuregawa, "Accurate Word Segmentation and POS Tagging for Japanese Microblogs: Corpus Annotation and Joint Modeling with Lexical Normalization", in proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 99-109, 2014.

[6] R. Nararatwong, N. Cooharajananone, " Improving Thai Word and Sentence Segmentation Using Linguistic Knowledge", IEICE Transactions on Information and System, pp. 3218-3225, 2018.

[7] Tuan-Phong Nguyen and Anh-Cuong Le, "A hybrid approach to Vietnamese word segmentation", in IEEE RIVF International Conference on Computing & Communication Technologies, 2016.

[8] T. T. Thet, J. C. Na and W. K. Ko, " Word segmentation for the Myanmar language", in the Journal of Information Science, pp. 688-704, 2008.

[9] Z. M. Maung, Y. Mikami, "A Rule-based Syllable Segmentation of Myanmar Text", in the proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 51-58, 2008.