Morpheme Based Myanmar Word Segmenter
Sin Thi Yar Myint, Hanni Htun, Myat Myo Nwe Wai
Faculty of Computer Science, Myanmar Institute of Information Technology, Mandalay, Myanmar

How to cite this paper: Sin Thi Yar Myint | Hanni Htun | Myat Myo Nwe Wai "Morpheme Based Myanmar Word Segmenter" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.911-914, https://doi.org/10.31142/ijtsrd26520

ABSTRACT
Myanmar script has no fixed delimiters between words or syllables. Therefore, to achieve meaningful and correct segmented words from the text is a challenging task. This paper has proposed a morpheme-based Myanmar word tokenizer which combines rule-based syllable breaking and dictionary lookup syllable merging methods with longest string matching approach. The proposed approach is tested on a Monolingual dictionary that contains useful information for the word segmentation. It also contains above 32,581 words including headwords, stop words and essential words with Myanmar3 font. These words are collected from Myanmar and Essential Words dictionaries. According to the experimental results, it can provide the promising segmentation accuracy of Myanmar text.

KEYWORDS: Syllable breaking; Morpheme; style; styling

INTRODUCTION
Word segmentation is prerequisite for any Myanmar language processing such as part of speech (POS) tagging, search engine, translation, information retrieval, and word sense disambiguation and many more of Natural Language Processing (NLP) activities. Currently, there has no Myanmar word segmentation approach based on the morpheme of the word in Myanmar text using a dictionary approach. Morpheme represents the root of a specific word. According to the Myanmar language nature, a morpheme is a vital role for the machine translation of Myanmar text. By exploiting the power of morpheme word, it can achieve the easy way of translation of Myanmar text.

In this paper, we proposed the word segmentation approach which is not applied to statistical methods with the corpus. This approach is very useful when there is no linguistic resource such as corpus and copra for Myanmar language. We simply build the monolingual lexicon which is inspired by morpheme Myanmar words collected from Myanmar and Essential Words dictionaries. Syllables tokenization is defined as preprocessing. Syllable segmentation is done by using the rules on the syllable structure of Myanmar script for the input sentence. To determine word boundaries of the segmented syllables, the proposed approach is applied forward longest matching dictionary. This system can segment into morpheme-based Myanmar words from the input sentence of text by comparing one by one character from the input string with the monolingual dictionary. This approach is very simple but it proved that this is a practical approach which is not available the applicable linguistic resources.

RELATED WORK
In this section, previous works on Myanmar word segmentation are reviewed. Win Pa Pa and NiLar Thein experimented Disambiguation in Myanmar Word Segmentation. The authors solved the word ambiguity problems by combining Forward Maximum Matching, Backward Maximum Matching and Joint Entropy. And then, they tried to solve the ambiguity problem using a statistical approach with the corpus. The authors described Precision of word segmentation for this approach was 92% and recall is 94%. Tun Thura Thet, Jin-Cheon Na, Wunna Ko, was a proposed word segmentation for the Myanmar language. They applied rule-based syllable segmentation and also used dictionary-based statistical syllable merging, for the word ambiguity. The authors combined with Mutual Information by calculating collocation strength with the corpus. They showed that Precision 98.94%, Recall 99.05%, F-Measure98.99.

“Myanmar Word Segmentation using Syllable level Longest Matching” was proved by Hla Hla Hay, Kavi Narayana Murthy. They used word List above 800,000 words including inflected forms. The authors also applied to stop word removal first and also used the Ngram approach for syllable matching. They achieved Recall 98.81%, Precision 99.11%, F-Measure 98.95%, also tested on the sentence level which is collected from web documents, grammar books and stories.

MYANMAR LANGUAGE
Myanmar language is the official language of the Union of the Republic of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which is descended from the Brahmi script of ancient South India.

A. Myanmar Script
A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right.
Myanmar script contains 33 consonants, 8 vowels (free-standing and attached), 2 diacritics, 11 medials, a vowel killer or ASAT, 10 digits and 2 punctuation marks [4].

B. Syllable Breaking in Myanmar Text

Syllable breaking is the process of identifying syllable boundaries in a text. The syllable is the smallest unit of language. In Myanmar text, a syllable can start with a consonant may be followed by a medial consonant. After the vowel, a syllable may end with nasalization of the vowel or an unreleased glottal stop. At the end of syllable, a final consonant usually has an ‘asat’ sign above it, to show that there is no inherent vowel. In multisyllabic words derived from an Indian language such as Pali, where two consonants occur internally without intervening vowel, the consonants tend to be stacked vertically, and the asat sign is not used. There are a set of Myanmar numerals, which are used just like Latin digits [2]. Firstly, syllable segmentation is done by using the rules on the syllable structure of the Myanmar script. Syllable breaking rules are based on combining consonant and vowel, devocalizing and kinzi, contractions, syllable chaining, distinct letter, single character and loan words. In syllable breaking stage, the proposed system determines a syllable boundary by comparing pairs of characters to find whether a break is possible or not between them. Moreover, the accuracy results of syllable segmentation are described in Table I and Table II.

1. Combining consonant and vowel

2. Devowelizing and Kinzi Devowelising and Kinzi

3. Syllable Changing

4. Single Character

5. Contraction

**MYANMAR WORD SEGMENTATION**

Word segmentation is the process of parsing concatenated text (i.e., text that contains no spaces or other word separators) to infer where word breaks exist. Myanmar script doesn’t need to put white spaces between words or syllables. Modern writing style contains spaces after each clause in order to enhance readability [5]. Generally, a word is a basic unit of language that carries meaning and can be spoken or written. A Myanmar word can consist of one or more morphemes that are linked more or less tightly together.

**Example:** အ ထွက် သျင် စိုက် ဆောင် ရွက် သော လိုအပ်

And then a Myanmar word will consist of a root or stem and zero or more affixes.

**Example:** ပေး မယ် ပေး မယ် ပေး မယ် ပေး မယ်

Moreover, Myanmar words can be combined to form phrases, clauses and sentences.

In addition, a word consisting of two or more stems joined together is known as a compound word[3].

**Example:** ပြ မြဲ လေး မြဲ လေး မြဲ လေး မြဲ လေး

And then, the next step was to merge the segmented syllables into the meaningful word from the input sentence. Syllable merging is done by using the longest matching approach and mapped with the lexicon. The algorithm starts from the beginning of a sentence, finding the longest matching word compared with the lexicon and then repeating the process until it reaches the end of the sentence. This system can segment into a morpheme-based word from the input sentence by comparing one by one character from the input string with the monolingual dictionary. The process of word segmentation is shown in Figure 2. This system is tested on all types of simple and complex sentence types of Myanmar text including one or more clauses and phrases. The accuracy results are mentioned in Table 2. There may be some problems in syllable merging of the proposed system. Because of the longest matching approach, it cannot give the correct segmentation of all words in the input sentence. It can find segment conflicts in some word in the sentence.

**Example:** အ လာ ကြည့် သျင် ဆောင် ရွက် သော လိုအပ်

With the longest matching approach, this sentence is segmented to the wrong word into

**Fig. 1. Process of Word Segmentation**
The structure of the sentence in Myanmar language may be simple and compound or complex. Generally, the sentence is subdivided into phrases. The phrase is subdivided into words. Word is subdivided into syllables. A syllable is the smallest unit of the language [3]. In this case, either simple or compound sentence can be contained with one more phrases and one or more clauses. A group of words, which makes sense, but not complete sense, is called a Phrase. It is a group of related words without a Subject and a Verb. Examples: in the east, on a wall, with blue trimming, on the bridge, with red hair [2]. A clause is a group of words that contains both a subject and a predicate but cannot always be considered as a full grammatical sentence. Clauses can be either independent clauses (also called main clauses) or dependent clauses (also called subordinate clauses) [2]. Like an English sentence, Myanmar sentence is also composed of one or more clauses and phrases. Myanmar script contains 33 consonants, 8 vowels (free-standing and attached, 2 diacritics, 11 medials, a vowel killer or ASAT, 10 digits and 2 punctuation marks [4].

1. **Examples for adding adjective & adverb phrase in a simple sentence**
   - အိမ်အတွင်း ရိုက်ထည်မှု ရှိသည်။ (There is a painting in the house.)
   - ရိုက်ထည်မှု ရှိသည်။ အိမ်အတွင်း (There is a painting in the house.)

2. **Examples for adding phrases in a simple sentence**
   - မိဘကို မြင့်မြတ်စေချင်သည်။ (I want to go to my mother.)
   - မိဘကို မြင့်မြတ်စေချင်သည်။ မျက်စွာ (I want to go to my mother.)

3. **Examples for adding time phrases in a simple sentence**
   - စားသော်ကြည်သည်။ (I eat at 7 PM.)
   - စားသော်ကြည်သည်။ ကျန်ရှိ (I eat at 7 PM.)

4. **Examples for adding accusative phrases in a simple sentence**
   - မိဘကို မြင့်မြတ်စေချင်သည်။ (I want to go to my mother.)
   - မိဘကို မြင့်မြတ်စေချင်သည်။ မျက်စွာ (I want to go to my mother.)

5. **Examples of a compound sentence with a dependent clause and independent clause**
   - ယူနီတွင် ရိုက်ထည်မှု ရှိသည်။ (There is a painting in the room.)
   - ရိုက်ထည်မှု ရှိသည်။ ယူနီတွင် (There is a painting in the room.)

6. **Examples of a compound sentence with three clauses**
   - မိဘကို မြင့်မြတ်စေချင်သည်။ စားသော်ကြည်သည်။ (I want to go to my mother and eat at 7 PM.)
   - စားသော်ကြည်သည်။ ကျန်ရှိ (I want to go to my mother and eat at 7 PM.)

7. **Examples of sentence Hidden object**
   - ကျွန်တော်တို့ အလုပ်ကို လုပ်ရန် မဖော်ပြထားသည်။ (We are not to do the work.)
   - လုပ်ရန် မဖော်ပြထားသည်။ ကျွန်တော်တို့ (We are not to do the work.)

8. **Examples of sentence changing the position of subject and reason**
   - ကျွန်တော်တို့ အလုပ်ကို လုပ်ရန် မဖော်ပြထားသည်။ (We are not to do the work.)
   - လုပ်ရန် မဖော်ပြထားသည်။ ကျွန်တော်တို့ (We are not to do the work.)

**EXPERIMENT RESULTS**

Table I and Table II show the experimental results of word segmentation system for syllable breaking and syllable merging word segmentation. Accuracy result for syllable breaking is 100% correct.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>Accuracy Results on syllable Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syllable Type</td>
<td>NCseg</td>
</tr>
<tr>
<td>Unique Syllable</td>
<td>1903</td>
</tr>
<tr>
<td>Tokens</td>
<td>7069</td>
</tr>
<tr>
<td>Sentence</td>
<td>1226</td>
</tr>
</tbody>
</table>

Accuracy=NCseg/Nseg*100

NCseg=the number of correctly segmented syllables by the program on the input

Nseg=the number of total segmented syllables verified manually
TABLE II Accuracy Results on Word Segmentation

<table>
<thead>
<tr>
<th>Syllable Type</th>
<th>Ncseg</th>
<th>Ntseg</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Syllable</td>
<td>7069</td>
<td>6679</td>
<td>95.77%</td>
</tr>
<tr>
<td>Tokens</td>
<td>1226</td>
<td>926</td>
<td>95.53%</td>
</tr>
</tbody>
</table>

\[ \text{Accuracy} = \frac{\text{NCmg}}{\text{NTmg}} \times 100 \]

NCmg = the number of correctly merge syllables by the program on the input

Ntmg = the number of total merge syllables verified manually

Tested Dictionary contains 32,581 tokens. Sentences are tested upon all kind of sentence types, namely (simple, compound or complex). Covers on all complex sentence type including a sentence with one clause, two clauses, and three clauses.

CONCLUSION
This paper has proposed an approach for Myanmar word segmentation by using rule-based syllable breaking and dictionary lookup syllable merging methods. In the syllable breaking stage, the proposed system determines a syllable boundary by comparing pairs of characters to find whether a break is possible or not between them. And then, it merges the segmented syllables into a meaningful word by using the dictionary lookup approach with the longest string matching algorithm. Moreover, this proposed system can produce correct morpheme-based Myanmar words from the input sentence. It can also solve to segment the words with one or more phrases and clauses of in the written Myanmar sentences. It can give the correct segmented words which contain one or more dependent clauses and independent clauses on all types of simple and compound sentences of Myanmar text. So, it can support many benefits to Myanmar to English translation system and further (NLP) tasks such as information retrieval, noun phrase identification, verb phrase identification, named entity recognition, word sense disambiguation and many more of NLP activities.

References