International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

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**

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 the Myanmar language, there is no statistical corpus resources and training data to test the word segmentation algorithm for Myanmar language.

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 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%, FMeasure98.99.

"Myanmar Word Segmentation using Syllable level Longest Matching" was proved by Hla Hla Htay, 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 syallable, 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 with no 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

သူ = သ + ူ = cosonant + vowel

သူပြုံးလိုက်သည့်အခါတိုင်းသူ၏ပါးချိုင့်လေးများပေါ်လာသည်။

သူ/ ပြုံး /လိုက် /သည့်/ အ/ ခါ /တိုင်း/ သူ /၏/ ပါး /ချိုင့် /လေး /များ /ပေါ် /လာ/ သည်/။

2. Devowelizing and Kinzi Devowelising and Kinzi

ကင် = က+င+် = consonant+ consonat + asat

ခွေးသည်အမဲကင်ကိုပစ်ချခ်ဲ့လွတ်ရာသို့ပြေးလေ၏။

ခွေး /သည် /အ/ မဲ /ကင်/ကို/ ပစ် /ချ /ခ်ဲ့/လွတ် /ရာ /သို့ /ပြေး/ လေ /၏ /။

သင်္ကြန် = သ+ိ်+က+ ြ+န+်=consonat+kinzi+ consonat+ medial + consonat +asat

ကျွန်ုပ်တို့သည် သင်္ကြန်ပွဲတော်တွင် ရေ ကစားသည်။

ကျွန်ုပ်/တို့/သည် /သင််/ကြန်/ပွဲ/တော်/တွင် /ရေ /က/စား/သည်/။

3. Syllable Changing

လိမ္မာ=လ+ိ+မ္+မ+ာ=consonat + vowel +consonat=+syllable chain+ consonat +vowel

သားသမီးလိမ္မာလျှင်မိဘစိတ်ချမ်းသာမည်။

သား /သ/ မီး /လိ/ မ္ /မာ/ လျှင်/ မိ/ ဘ /စိတ် /ချမ်း /သာ /မည်/ ။

4. Single Character

ဝိဿာ=ဝ+ိ+သ+ာ=consonat+vowel+singlecharacter+vowel

အမေသည်ငါးခြောက်တစ်ဝိဿာတိတိဝယ်လာသည်။

အ /မေ /သည် /ငါး /ခြောက်/ တစ် /ဝိ/ သာ /တိ/ တိ/ ဝယ် /လာ /သည်/ ။

5. Contraction

ယောကျ်ား=ယ+ ော + ာ+ က+ ်+ ျ+ာ+း=consonat+medial+ vowel+consonat+asat+medial+vowel+tonemark

ယောကျ်ားလေးများသည်အလွန်အပြေးမြန်သည်။

ယောကျ်ား/လေး/များ/သည်/အ/လွန်/အ/ပြေး/မြန်/သည်/။

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: စပါး, နှင်းဆီ, ဆံပင်, ပန်း, မီးပူ

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

Example: လှ, လှခြင်း, လှပမှု, အလှ, လှပသည်, လှလှပပ, လပသော

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

တာဝန်သိသောရွာသားများသည်
ပျက်စီးနေသောလမ်းကို
အင်တိုက်အားတိုက်ပြုပြင်ကြသည်။

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

ရေအိုး, ကူးသန်းရောင်းဝယ်, ပြေးဖက်, သွားစား ရေးသားစပ်ဆို,

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 Figure2. 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 Table2. 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: သူမလိုင်အလွန်ကြိုက်သည်။

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

သူမ/ လိုင်/ အလွန်/ ကြိုက်/ သည် /.

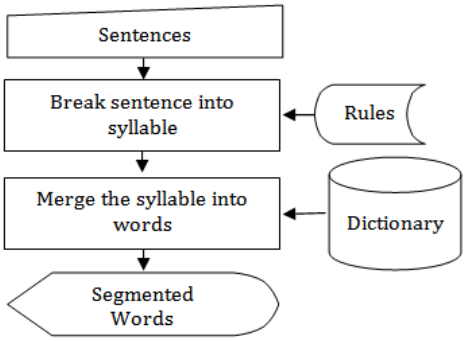

Fig1. Process of Word Segmentation

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

လသည်ကြည်လင်သောကောင်းကင်၌ရွှန်းရွှန်းပပသာနေသည်။

ကတ္တား +နာမဝိသေသန +နေရာပြ +ကြိယာဝိသေသန+ ကြိယာ

လသည်+ ကြည်လင်သော+ကောင်းကင်၌+ ရွှန်းရွှန်းပပ +သာနေသည်။

လ/သည်/ကြည်လင်/သော/ကောင်းကင်/၌/ရွှန်းရွှန်းပပ/သာ/နေ/သည်/

2. Examples for adding phrases in a simple sentence

တာဝန်သိသောရွာသားများသည်ပျက်စီးနေသောလမ်းကိုအင်တိုက်အားတိုက်ပြုပြင်ကြသည်။

နာမဝိသေသန +ကတ္တား +နာမဝိသေသန+ ကံ + ကြိယာဝိသေသန + ကြိယာ

တာဝန်သိသော ရွာသားများသည် ပျက်စီးနေသော လမ်းကို အင်တိုက်အားတိုက် ပြုပြင်ကြသည်။

တာဝန်သိ/သော/ရွာသား/များ/သည်/ပျက်စီး/နေ/သော/လမ်း/ကို/အင်တိုက်အား/တိုက်/ပြုပြင်/ကြ/သည်/

3. Examples for adding time phrases in a simple sentence

မိန်းကလေးများသည်လသာသောညတွင်ထုပ်ဆီးတိုးကြသည်။

ကတ္တားပုဒ် +အချိန်ပြပုဒ်စု +ကြိယာ

မိန်းကလေးများသည် လသာသောညတွင် ထုပ်ဆီးတိုးကြသည်။

မိန်းကလေး/များ/သည်/လ/သာ/သော/ည/တွင်/ထုပ်ဆီးတိုး/ကြ/သည်/

4. Examples for adding accusative phrases in a simple sentence

ရွာသားများသည်ကျွန်တော်တို့ကိုကျောက်ပန်းတောင်းကထန်းလျက်ဖြင့်ဧည့်ခံသည်။

ကတ္တား +ကံပုဒ် +အသုံးခံပြပုဒ်စု +ကြိယာ

ရွာသားများသည် ကျွန်တော်တို့ကိုကျောက်ပန်းတောင်းကထန်းလျက်ဖြင့် ဧည့်ခံသည်။

ရွာသား/များ/သည်/ကျွန်တော်/တို့/ကို/ကျောက်ပန်းတောင်း/က/ထန်းလျက်/ဖြင့်/ဧည့်ခံ/သည်/

5. Examples of a compound sentence with a dependent clause and independent clause

အမှီဝါကျ(dependent clause)	အမှီခံဝါကျ(independent)
သူမသည်ဧည့်သည်များကိုကျွေးရန်	အုန်းထမင်းချက်နေသည်။
သူမ/သည်/ဧည့်သည်/များ/ကို/ကျွေး/ရန်/အုန်းထမင်း/ချက်/နေ/သည်/	
အမှီဝါကျ	အမှီခံဝါကျ
ကျောင်းသားတို့သည်နွေရာသီ၌ချောင်းထဲကငါးများကိုဖမ်း၍	မြစ်ထဲသို့လွှတ်ကြသည်။
ကျောင်းသား/တို့/သည်/နွေရာသီ/၌/ချောင်း/ထဲက/ငါး/များ/ကို/ဖမ်း/၍/မြစ်/ထဲ/သို့/လွှတ်/ကြ/သည်/	

6. Examples of a compound sentence with three clauses

တောတောင်ပတ်ဝန်းကျင်သာယာလှသည်။
နွေဦးရာသီ၌လေရူးကလေးများတိုက်ခတ်လာသည်။
ရွက်ဟောင်းများကြွေသည်။
ရွက်သစ်ရွက်ညွှန့်ပုရစ်ဖူးကလေးများအစီရရီထွက်ပေါ်လာသည်။
တောတောင်ပတ်ဝန်းကျင် သာယာလှသော နွေဦးရာသီ၌ လေရူးကလေးများ တိုက်ခတ်လာသည်နှင့်တစ်ပြိုင်နက် ရွက်ဟောင်းများကြွေကာ ရွက်သစ်ရွက်ညွှန့်ပုရစ်ဖူးလေးများ အစီရရီထွက်ပေါ်လာသည်။
တော/တောင်/ပတ်ဝန်း/ကျင်/သာယာ/လှ/သော/နွေဦး/ရာသီ/၌/လေရူး/ကလေး/များ/တိုက်/ခတ်/လာ/သည်/နှင့်/တစ်ပြိုင်နက်/ရွက်ဟောင်း/များ/ကြွေ/ကာ/ရွက်သစ်/ရွက်ညွှန့်/ပုရစ်ဖူး/လေး/များ/အစီရရီ/ထွက်ပေါ်/လာ/သည်/

7. Examples of sentence Hidden object

အဝတ်ကိုကန်ဘောင်ပေါ်၌မလျှော်ရ။

ကတ္တား +ကံ + နေရာပြ + ကြိယာ + အဝတ်ကို ကန်ဘောင်ပေါ်၌ မလျှော်ရ။

အဝတ်/ကို/ကန်ဘောင်/ပေါ်/၌/မ/လျှော်/ရ/

8. Examples of sentence changing the position of subject and reason

မဟာဓနကသည်ထီးနန်းကိုလုံ့လကြောင့်ရသည်။

ကတ္တား +ကံ +အကြောင်းပြ +ကြိယာ

မဟာဓနကသည် ထီးနန်းကို လုံ့လကြောင့် ရသည်။

မဟာဓနက/သည်/ထီးနန်း/ကို/လုံ့လ/ကြောင့်/ရသည်/

လုံ့လကြောင့်မဟာဓနကသည်ထီးနန်းကိုရသည်။

အကြောင်းပြ +ကတ္တား + ကံ + ကြိယာ

လုံ့လကြောင့် မဟာဓနကသည် ထီးနန်းကို ရသည်။

လုံ့လ/ကြောင့်/မဟာဓနက/သည်/ထီးနန်း/ကို/ရသည်/

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.I Accuracy Results on syllable Segmentation

Syallable Type	NCseg	NTseg	Accuracy
Unique Syllable	1903	1903	100%
Tokens	7069	7069	100%
Sentence	1226	1226	100%

Accuracy=NCseg/NTseg*100

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

Ntseg=the number of total segmented syllables verified manually

TABLE.II Accuracy Results on Word Segmentation

Syallable Type	NCseg	NTseg	Accuracy
Unique Syllable	7069	6769	95.77%
Tokens	1226	926	75.53%

Accuracy=NCmg/NTmg*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

[1] C. D. Manning, H. Schiitze," Foundations Of Statistical Natural Language Processing", The MIT Press, Cambridge, Massachusetts London, England, 2000. .

[2] https:// Myanmar script notes.htm, https:// what-isclause.html, https:// what-is-phrase.html.

[3] Myanmar Grammar, First Edition, Myanmar Language Commission, memorable for 30th anniversary, June 2005.

[4] Myanmar Orthography, Second Edition, Myanmar Language Commission, June 2003.

[5] M. T. Win & et.al, "Burmese Phrase Segmentation", Proceedings of Conference on Human Language Technology for Development, Egypt, May 2011.

[6] Lexique Pro_ Myanmar lexicon (Version-2), July, 2011.

[7] T. T. Thet, J. C. Na, W. K. Ko, "Word segmentation for the Myanmar Language", Journal of Information Science, 2007, PP. 1-17.

[8] T. H. Hlaing, "Manually Constructed Context-Free Grammar For Myanmar Syllable Structure", Nagaoka University of Technology Nagaoka, JAPAN, 2011.

[9] W. P. Pa, N. L. Thein "Disambiguation in Myanmar Word Segmentation", "Proceedings Of the Seventh International Conference On Computer Applications", Yangon, Myanmar,2009, PP. 1-4.

[10] H. H. Htay, K. N. Murthy, Myanmar Word Segmentation using Syllable Level Longest Matching, "Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, January 2008.

[11] PyinNya Kyaw, Essential Words Dictionary and Myanmar Dictionary, First Edition, February 2010, Yangon.