Myanmar Word Segmentation Using Hybrid Approach

Khine Myint Myat, Khin Mar Soe
University of Computer Studies, Yangon
khinemm21@gmail.com, khinmarsoe@ucsy.edu.mm

Abstract

Word segmentation is a basic task and an important problem in natural language processing. In Myanmar text, words composed of single or multiple syllables are usually not separated by white space. Word segmentation to determine the boundaries of words for languages without word separators in orthography is a basic task in natural language processing. This system uses a 2-step longest matching approach. The first step was syllable segmentation, in the second was Hybrid Approach of left-to-right syllable maximum matching and hierarchical expectation maximization approach. This system is to be able to use as a pre-processing tool in Myanmar text processing such as Machine Translation, Information Retrieval, Search Engine using Myanmar language. The experiment result shows that 96% of accuracy in word segmentation.

Key words: *Machine Translation, Information Retrieval, Search Engine using Myanmar language*

1. **Introduction**

Word segmentation is a basic task and an important problem in natural language processing. It is to determine the boundaries of words for some languages without word separator in orthography are not delimited by white-space but instead must be inferred from the basic character sequence.

For Asian languages, most research on this task has focused on the segmentation for which the standard, state-of-the-art technique using conditional random fields has achieved satisfactory performance.

In this paper, we focus on applying word segmentation techniques to an understudied language, Myanmar. It adopted a simple dictionary based approach and used hierarchical expectation maximization approach.

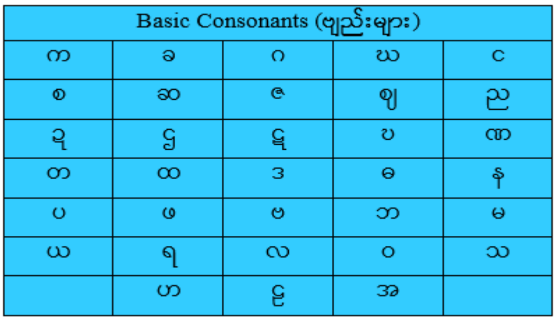
The remaining part of the paper are organized as follow. In section 2, Myanmar Language nature will be discussed and related work can be seen in section 3. The background theory and the overview of the proposed system will be explained in section 4 and 5 respectively. Section 6 contain the Experimental Work. For last two sections, section 7 and section 8, conclusion is described and reference paper concluded.

2. Myanmar Language

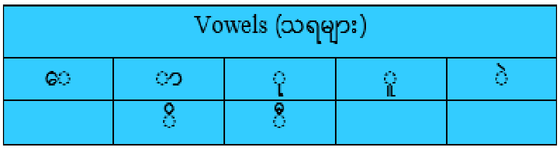
Myanmar language, also known as Burmese, is the official language of the Union of Myanmar and is more than one thousand years old. Burmese is a tonal and analytic language using the Burmese script. This is a phonologically based script, adapted from Mon, and ultimately based on an Indian (Brahmi) prototype.

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right without regular inter-word spacing, although inter-phrase spacing may sometimes be used.

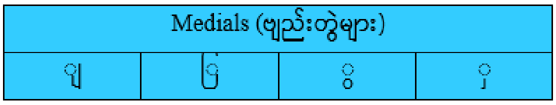
Myanmar characters can be classified into three groups: consonants, medials and vowels.



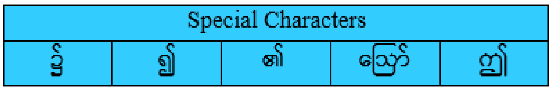
Basic Consonants (ဗျည်းများ)				
က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
	ဟ	ဠ	အ	



Vowels (သရများ)				
ေ◌	◌ာ	◌ု	◌ူ	◌ဲ
	◌ိ	◌ီ		



Medials (ဗျည်းတွဲများ)			
◌ျ	◌ြ	◌ွ	◌ှ



Special Characters				
ဦ	၍	၏	ဪ	ဤ

3. Related Works

- **A Hybrid Approach to Word Segmentation of Vietnamese Texts**: The work of (Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussamaly, Tuong Vinh Ho) are based on Finite-state automata technique, regular expression parsing and the maximal-matching strategy. This system gives a relatively high accuracy about 96%.
- **A Hybrid Approach to Word Segmentation of Chinese Texts:** The work of (Li Hazhou and Yuan Baosheng, Kent Ridge Digital Labs, Singapore.) are based on Lexical knowledge based method and linguistic knowledge based method. This system accuracy is high.
- **A Hierarchical EM Approach to Word Segmentation:** The work of (Fuchun Peng and Dale Schuurmans) are based on Two-level hierarchical probability model. This system performs 49.1% word precision, 60.1% word recall and 53.8% word F-measure.

4. Background Theory

Natural Language Processing is an interdisciplinary field of artificial intelligence, computer science and computational linguistics. It deals with the interactions between computers and human languages. Every aspect of NLP is used in script recognition, optical character recognition, sentiment analysis, part of speech tagging, information extraction, social media analysis etc. Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks.

In Myanmar text, words composed of single or multiple syllables are usually not separated by white space. Word segmentation to determine the boundaries of words for languages without word separators in orthography is a basic task in natural language processing.

Word Segmentation is pre-processing step of many NLP applications such as:

- Machine Translation
- Information Retrieval
- Search Engine

5. Overview of the Proposed System

The proposed system design can be seen in figure 1. This system has two steps: syllable segmentation phase and hybrid segmentation phase.

The first step was syllable segmentation, in the second step left-to-right syllable maximum matching word segmentation with a dictionary was performed and hierarchical expectation maximization approach.

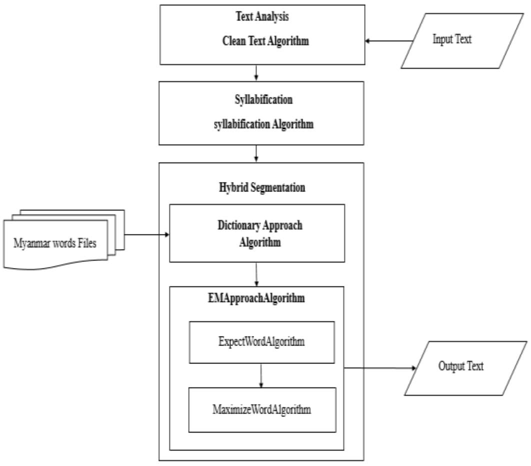


Figure 1. Proposed System Design

Text Analysis is the process of removing unwanted words (“-”, “။”, “၊”, “(”, “)”, “?”) etc from the input text. It outputs as readable sentences.

Syllabification is the ability to identify how many syllables there are in a word. A syllable boundary can be determined by comparing pairs of characters to find whether a break is possible or not between them. The detail flow can be seen in figure 2.

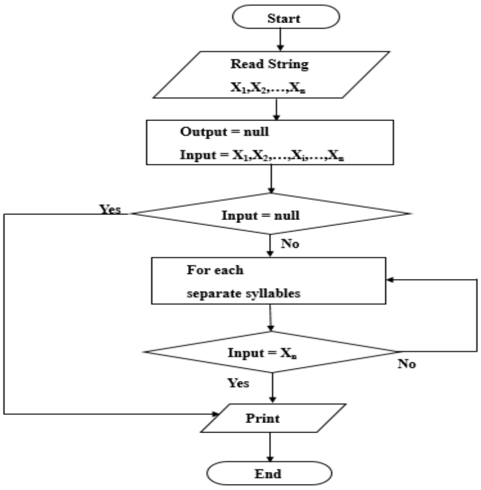


Figure 2. Syllable Segmentation Flow Chart

Dictionary Approach is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation. It is by looking in a prepared dictionary and matching the longest substring in an input sentence is a classic word segmentation approach.

This method segments using segments chosen from a dictionary. It strives to segment using the longest possible segments. The segmentation process may start from either end of the sequences.

The Hierarchical EM Approach is used for unsupervised text segmentation. It has two levels. In the first level, generate all morphemes with one to five characters from the training corpus *C*, use EM to learn a probability distribution over morphemes, prune low probability morphemes, and segment the original training corpus *C* into a morpheme sequence *G*. In the second level, generate a large word lexicon from *G* (multi-grams over morphemes), use EM to learn a probability distribution over words, and segment *G* into a word sequence *W*.

Overall, the first level determines,

$$G^* = \underset{G}{\operatorname{argmax}}\{prob(G,C|\Theta_g)\} \text{ ------- Eq:1}$$

and the second level determines

$$W^* = \operatorname{argmax}\{prob(W,G|\Theta_w)\} \text{ ------- Eq:2}$$

where Θ_g and Θ_w are the distributions over morpheme lexicon and word lexicon respectively. The EM algorithms in both levels are identical except that in the first level the basic observation unit is character and in the second level the basic unit is morpheme. The hierarchical EM flow is depicted in figure 3.

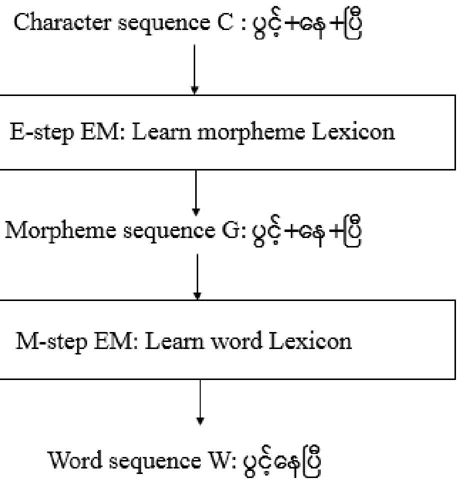
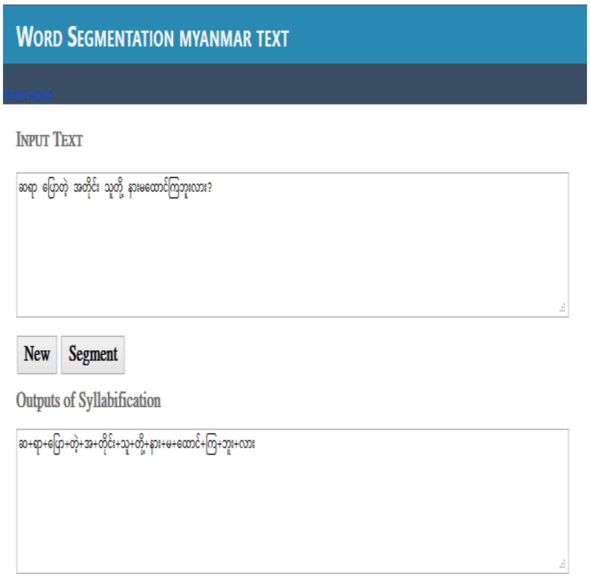

Character sequence C : ပွင့်+နေ+ပြီ
E-step EM: Learn morpheme Lexicon
Morpheme sequence G: ပွင့်+နေ+ပြီ
M-step EM: Learn word Lexicon
Word sequence W: ပွင့်နေပြီ

Figure 3. Hierarchical EM segmentation model

Example 1: (Test with one sentence)


WORD SEGMENTATION MYANMAR TEXT

Home About

INPUT TEXT

ဆရာ ပြောတဲ့ အတိုင်း သူတို့ နားမထောင်ကြဘူးလား?

New Segment

Outputs of Syllabification

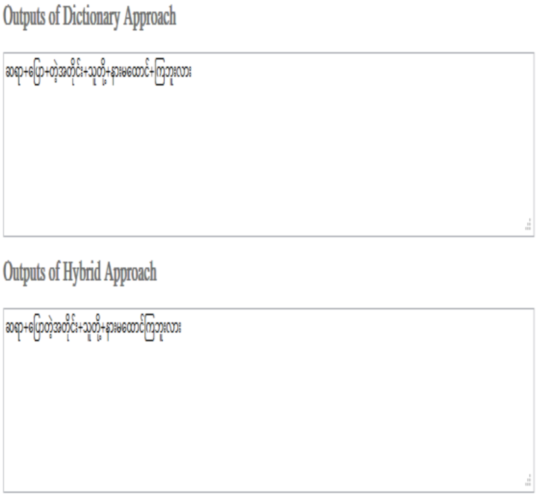
ဆ+ရာ+ပြော+တဲ့+အ+တိုင်း+သူ+တို့+နား+မ+ထောင်+ကြ+ဘူး+လား

Input text is

"ဆရာ ပြောတဲ့ အတိုင်း သူတို့ နားမထောင်ကြဘူးလား?"

and after click Segment button, the outputs of syllabification is

"ဆ+ရာ+ပြော+တဲ့+အ+တိုင်း+သူ+တို့+နား+မ+ထောင်+ကြ+ဘူး+လား "


Outputs of Dictionary Approach

ဆရာ+ပြော+တဲ့အတိုင်း+သူတို့+နားမထောင်+ကြဘူးလား

Outputs of Hybrid Approach

ဆရာ+ပြောတဲ့အတိုင်း+သူတို့+နားမထောင်ကြဘူးလား

And then the outputs of dictionary approach is

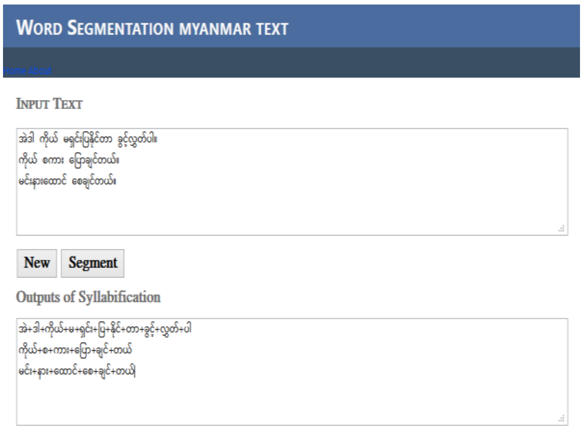
"ဆရာ+ပြော+တဲ့အတိုင်း+သူတို့+နားမထောင် +ကြဘူးလား"

that input text is the outputs of syllabification and the outputs of hybrid approach is

“ဆရာ+ပြောတဲ့အတိုင်း+သူတို့+နားမထောင်ကြဘူးလား”

that input text is the outputs of dictionary approach and using expectation maximization approach.

Example 2: (Test with line by line paragraph)



အဲဒါ ကိုယ် မရှင်းပြနိုင်တာ ခွင့်လွှတ်ပါ။
ကိုယ် စကား ပြောချင်တယ်။
မင်းနားထောင် စေချင်တယ်။

အဲ+ဒါ+ကိုယ်+မ+ရှင်း+ပြ+နိုင်+တာ+ခွင့်+လွှတ်+ပါ
ကိုယ်+စ+ကား+ပြော+ချင်+တယ်
မင်း+နား+ထောင်+စေ+ချင်+တယ်၊

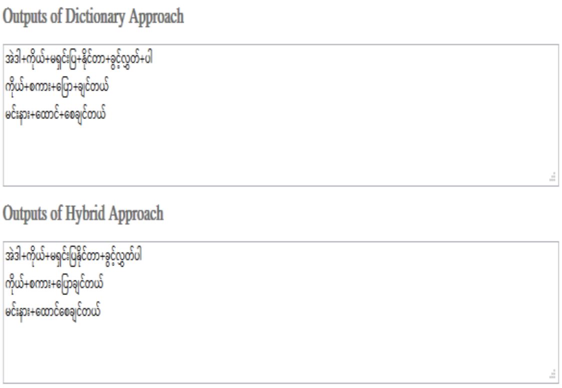
Input text is

“ အဲဒါ ကိုယ် မရှင်းပြနိုင်တာ ခွင့်လွှတ်ပါ။
ကိုယ် စကား ပြောချင်တယ်။
မင်းနားထောင် စေချင်တယ်။”

with line by line paragraph format and after click Segment button, the outputs of syllabification is

“အဲ+ဒါ+ကိုယ်+မ+ရှင်း+ပြ+နိုင်+တာ+ခွင့်+လွှတ်+ပါ
ကိုယ်+စ+ကား+ပြော+ချင်+တယ်
မင်း+နား+ထောင်+စေ+ချင်+တယ် ”

with line by line paragraph format defined by (**“။”**).



အဲဒါ+ကိုယ်+မရှင်းပြ+နိုင်တာ+ခွင့်လွှတ်+ပါ
ကိုယ်+စကား+ပြော+ချင်တယ်
မင်းနား+ထောင်+စေချင်တယ်

အဲဒါ+ကိုယ်+မရှင်းပြနိုင်တာ+ခွင့်လွှတ်ပါ
ကိုယ်+စကား+ပြောချင်တယ်
မင်းနား+ထောင်စေချင်တယ်

And then the outputs of dictionary approach is

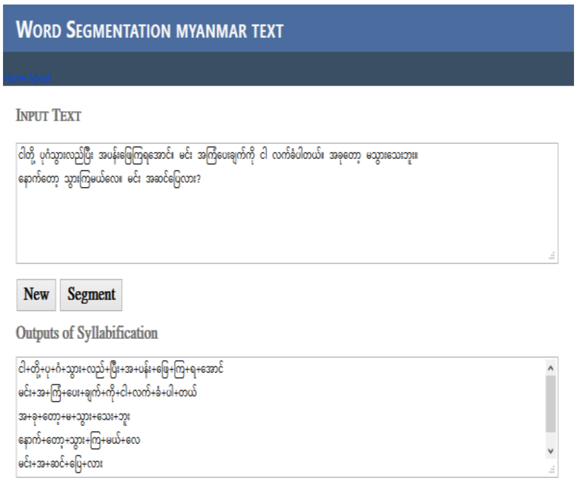
“အဲဒါ+ကိုယ်+မရှင်းပြ+နိုင်တာ+ခွင့်လွှတ်+ပါ
ကိုယ်+စကား+ပြော+ချင်တယ်
မင်း+နားထောင်+စေချင်တယ် ”

with line by line paragraph format defined by (**“။”**) that input text is the outputs of syllabification and the outputs of hybrid approach is

“အဲဒါ+ကိုယ်+မရှင်းပြနိုင်တာ+ ခွင့်လွှတ်ပါ
ကိုယ်+စကား+ပြောချင်တယ်
မင်း+နားထောင်စေချင်တယ် ”

with line by line paragraph format defined by (**“။”**) that input text is the outputs of dictionary approach and using expectation maximization approach.

Example 3: (Test with one paragraph)



ငါတို့ ပုဂံသွားလည်ပြီး အပန်းဖြေကြရအောင်။ မင်း အကြံပေးချက်ကို ငါ လက်ခံပါတယ်။ အခုတော့ မသွားသေးဘူး။
နောက်တော့ သွားကြမယ်လေ။ မင်း အဆင်ပြေလား?

ငါ+တို့+ပု+ဂံ+သွား+လည်+ပြီး+အ+ပန်း+ဖြေ+ကြ+ရ+အောင်
မင်း+အ+ကြံ+ပေး+ချက်+ကို+ငါ+လက်+ခံ+ပါ+တယ်
အ+ခု+တော့+မ+သွား+သေး+ဘူး
နောက်+တော့+သွား+ကြ+မယ်+လေ
မင်း+အ+ဆင်+ပြေ+လား

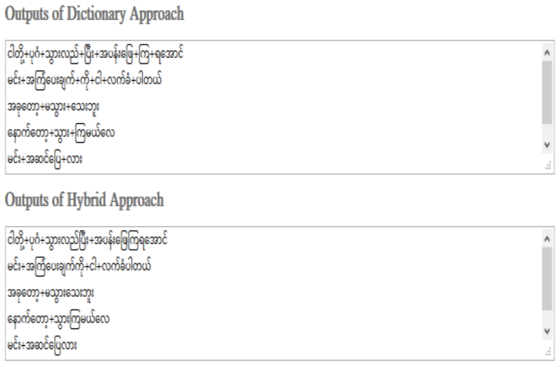
Input text is

“ငါတို့ ပုဂံသွားလည်ပြီး အပန်းဖြေကြရအောင်။
မင်း အကြံပေးချက်ကို ငါ လက်ခံပါတယ်။
အခုတော့ မသွားသေးဘူး။
နောက်တော့ သွားကြမယ်လေ။
မင်း အဆင်ပြေ လား?”

with one paragraph format and after click Segment button, the outputs of syllabification is

“ငါ+တို့+ပု+ဂံ+သွား+လည်+ပြီး+အ+ပန်း+ဖြေ+ကြ
+ရ+အောင်
မင်း+အ+ကြံ+ပေး+ချက်+ကို+ငါ+လက်+ခံ+ပါ+တ
ယ်
အ+ခု+တော့+မ+သွား+သေး+ဘူး
နောက်+တော့+လောက်+သွား+ကြ+မယ်+လေ
မင်း+အ+ဆင်+ပြေ+လား”

with line by line paragraph format defined by (**“။”**).



Outputs of Dictionary Approach

ငါတို့+ပုဂံ+သွားလည်+ပြီး+အပန်းဖြေ+ကြ+ရအောင်
မင်း+အကြံပေးချက်+ကို+ငါ+လက်ခံ+ပါတယ်
အခုတော့+မသွား+သေးဘူး
နောက်တော့+သွား+ကြမယ်လေ
မင်း+အဆင်ပြေ+လား

Outputs of Hybrid Approach

ငါတို့+ပုဂံ+သွားလည်ပြီး+အပန်းဖြေကြရအောင်
မင်း+အကြံပေးချက်ကို+ငါ+လက်ခံပါတယ်
အခုတော့+မသွားသေးဘူး
နောက်တော့+သွားကြမယ်လေ
မင်း+အဆင်ပြေလား

And then the outputs of dictionary approach is

“ငါတို့+ပုဂံ+သွားလည်+ပြီး+အပန်းဖြေ+ကြရအောင်
မင်း+အကြံပေးချက်+ကို+ငါ+လက်ခံ+ပါတယ်
အခု+တော့+မသွား+သေးဘူး
နောက်+တော့+သွား+ကြ+မယ်+လေ
မင်း+အဆင်ပြေ+လား ”

with line by line paragraph format defined by (**“။”**) that input text is the outputs of syllabification and the outputs of hybrid approach is

“ငါတို့+ပုဂံ+သွားလည်ပြီး+အပန်းဖြေကြရအောင်
မင်း+အကြံပေးချက်ကို+ငါ+လက်ခံပါတယ်
အခုတော့+မသွားသေးဘူး
နောက်တော့+သွားကြမယ်လေ
မင်း+အဆင်ပြေလား ”

with line by line paragraph format defined by (**“။”**) that input text is the outputs of dictionary approach and using expectation maximization approach.

6. Experimental Work

We present in this section the experimental setup and give a report on results of experiments with the hybrid approach presented in the section 5. We also describe briefly hybrid approach of word segmentation for Myanmar texts.

6.1 Myanmar Words Constitution

The training words upon which we evaluate the performance of the maximum matching is a collection of 750 articles from the “Business, Entertainment and Sports” sections of the Myanmar newspaper like that kyaymon newspaper and myanmar ahlin newspaper, www.phothutaw.com , www.7daydaily.com , for a total of nearly 35,000 words that have been manually spell-checked and segmented by associated editors. Although there can be multiple plausible segmentations of a given Myanmar sentence, only a single correct segmentation of each sentence is kept. In each experiment, we take 90% of the gold test set as training set, and 10% as test set. We present in the next paragraph the performance measure of training documents and results.

6.2 Performance Measure

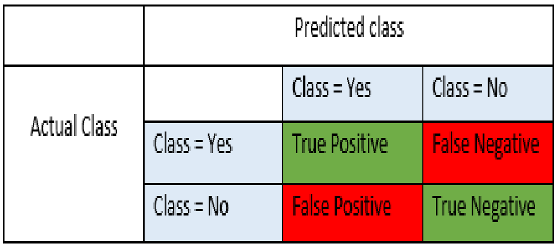
The performance of this proposed system can be measured by its efficiency and its effectiveness. This system measures the accuracy by using the following method.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.



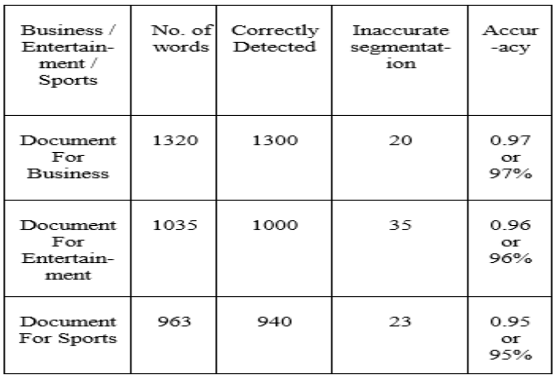
	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy - Accuracy is the most intuitive performance measure and can be defined as the ratio of correctly predicted class (TP+TN) to the total testing class (TP+FP+FN+TN). One may think that, if we have high accuracy then our model is best.

Accuracy = (TP+TN)/(TP+FP+FN+TN) Eq:3

where,	TP	=	True Positive
	TN	=	True Negative
	FP	=	False Positive
	FN	=	False Negative

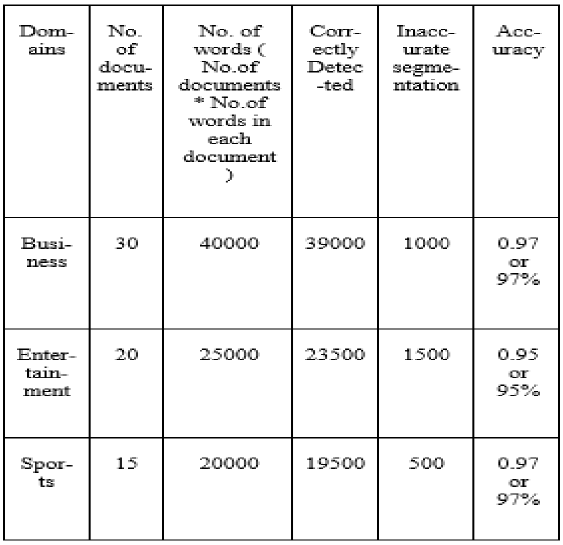
Table [6.1] Result of each document of each domain



Business / Entertainment / Sports	No. of words	Correctly Detected	Inaccurate segmentation	Accuracy
Document For Business	1320	1300	20	0.97 or 97%
Document For Entertainment	1035	1000	35	0.96 or 96%
Document For Sports	963	940	23	0.95 or 95%

Table [6.1] show that accuracy of each document of each domain (Business, Entertainment, Sports). Each document is article of each domain.

Table [6.2] Result of documents of each domain



Domains	No. of documents	No. of words (No.of documents * No.of words in each document)	Correctly Detected	Inaccurate segmentation	Accuracy
Business	30	40000	39000	1000	0.97 or 97%
Entertainment	20	25000	23500	1500	0.95 or 95%
Sports	15	20000	19500	500	0.97 or 97%

Table [6.2] show that accuracy of training documents and testing documents of each domain (Business, Entertainment, Sports). Each document is article of each domain.

7. Conclusion

The proposed system is used a hybrid approach to do word segmentation of Myanmar texts. This system is implemented by using ASP.NET and tested on IIS server. This system can be used as a pre-processing tool in Myanmar text processing such as search engine and machine translation. It can also provide as a web-based online system that can be used separately for every people. It is beneficial for the growth of neutrality and clarity of the word segmentation.

References

[1] "A Hybrid Approach to Word Segmentation of Vietnamese Texts" , Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussamaly, Tuong Vinh Ho, HAL ID : inria-00334761, 27 Oct 2008.

[2] "A Hybrid Approach to Word Segmentation of Chinese Texts", Li Hazhou and Yuan Baosheng, Kent Ridge Digital Labs, Singapore.

[3] "A Hierarchical EM Approach to Word Segmentation" , Fuchun Peng and Dale Schuurmans, Department of Computer Science, University of Waterloo, 200 University Avenue West, Water loo, Ontario, Canada, N2L 3G1.

[4] "Argmax and Max Calculus", Mark Schmidt, January 6, 2016.

[5] “Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer”, Tin Htay Hlaing and Yoshiki MIKAMI, Nagaoka University of Technology, JAPAN.

[6] “A Rule-based Syllable Segmentation of Myanmar Text” , Zin Maung Maung, Yoshiki Mikami, Nagaoka University of Technology, 1603-1 Kamitomioka, Nagaoka, Japan, January, 2008.

[7] “Deterministic Word Segmentation Using Maximum Matching with Fully Lexicalized Rules”, Manabu Sassano, Yahoo Japan Corporation, Midtown Tower, 9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan.

[8] “Expectation Maximization” , TLT-5906 Advanced Course in Digital Transmission, Jukka Talvitie, M.Sc. (eng), Department of Communication Engineering, Tampere University of Technology, December, 2013.

[9] “Tutorial on Expectation Maximization”, Stefanos Zafeiriou, Imperial College, London.

[10] “The Expectation Maximization Algorithm” , Sean Borman, July 18 2004.