MYANMAR WORD SEGMENTATION USING HYBRID APPROACH

KHINE MYINT MYAT

M.C.Sc. **MAY, 2018**

MYANMAR WORD SEGMENTATION USING HYBRID APPROACH

By

KHINE MYINT MYAT

B.C.Sc. (Hons:)

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

MAY, 2018

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis.

First and foremost, I would like to express my gratitude and my sincere thanks to **Dr. Mie Mie Thet Thwin**, the Rector of the University of Computer Studies, Yangon, for allowing me to develop this thesis.

I would like to express my special appreciation and my sincere special thanks to **Dr. Thi Thi Soe Nyunt**, Professor, and Head of Faculty of Computer Science Department, for her administrative supports and encouragements in development of the thesis.

I am deeply thankful to my supervisor, **Dr. Khin Mar Soe**, Professor, Natural Language Processing Lab, University of Computer Studies, Yangon, for her invaluable guidance, encouragement, superior suggestion and supervision on the accomplishment of this thesis.

I would like to thank **Daw Aye Aye Khine,** Associate Professor and Head of English Department, University of Computer Studies, Yangon, for editing my thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers for their support not only for the fulfillment of the degree of M.C.Sc. but also for my life.

I also thank my friends and colleagues for supporting in various ways to complete this thesis.

ABSTRACT

Word segmentation is a basic task and an important problem in natural language processing. In Myanmar language, words composed of single or multiple syllables are usually not separated by white space. Myanmar word segmentation is to determine the boundaries of words for languages without word separators in orthography. This system uses a 2-step longest matching approach. The first step was syllable segmentation and second uses Hybrid Approach of left-to-right syllable maximum matching and hierarchical expectation maximization approach. This system is intended 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 experimental result shows 93% of accuracy based on a collection of 300 articles from the business, entertainment and sports sections of the Myanmar newspaper nearly 35,000 words. The proposed word segmentation is implemented as a web-based tool using C# .Net language.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 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 and morphological analysis of Chinese, Japanese, and Korean, for which the standard, state-of-the-art technique using conditional random fields has achieved satisfactory performance. This proposed system, focuses on applying word segmentation techniques to an understudied language, Myanmar. It adopted a simple dictionary based approach and used hierarchical expectation maximization approach. An important reason for this is that there are few linguistically annotated resources for Myanmar and then only dictionary based approach and hierarchical expectation maximization approach are feasible.[1]

The work of the proposed system is based on three domains that are business, entertainment, and sports. This proposed system, tests and compares word segmentation techniques, including dictionary based approach and hierarchical expectation maximization approach. From the experimental results, it is shown that the accuracy is 93% based on a collection of 300 articles from the "Business, Entertainment and Sports" sections of the Myanmar newspaper like that Kyaymon newspaper and Myanmarahlin 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 the data size is still not comparable to the large-scale corporation of Burmese word segmentation.

Although there can be multiple plausible segmentations of a given Myanmar sentence, only a single correct segmentation of each sentence is kept. A single correct segmentation of a sentence can be assumed for two reasons. The first one is of its simplicity. The second one is due to the fact that are not currently aware of any effective way of using multiple segmentations in typical applications concerning Myanmar processing. In each experiment, 90% of the gold test set is taken as training set, and 10% as test set.

1.1 Overview of System

The proposed system has two steps: syllable segmentation phase and hybrid segmentation phase.

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

Syllable segmentation 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.

Maximum matching by looking in a prepared dictionary, extracting segments based on syllabled words and matching the longest substring in an input sentence is a classic word segmentation approach.

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

The Hierarchical Expectation Maximization Approach is used for unsupervised text segmentation. That consists of two processes: The E-step, and the M-step. The E-step generates all morphemes from the training corpus *C*, learn a probability distribution over morphemes, and segment the original training corpus *C* into a morpheme sequence *G*. The M-step maximizes a probability distribution over segment G and segment G into a word sequence W.

1.2 Objective of Thesis

The objectives of the thesis are as follows:

- To propose a hybrid approach in word segmentation
- To be able to use this proposed system as a pre-processing tool in Myanmar text processing such as Machine Translation, Information Retrieval, Search Engine using Myanmar language
- To develop this system as a web-based online system that can be used separately for every people

1.3 Organization of the Thesis

This thesis is organized into five chapters.

Chapter 1 includes introduction of the proposed system, overview and objectives of this system.

Chapter 2 describes the background theory of this system, Myanmar Language, syllable segmentation rules and word segmentation approaches.

Chapter 3 explains the design of the proposed system, class diagram and flow diagram of the system.

Chapter 4 presents implementation of proposed system which includes system algorithms and screen designs of the proposed system.

The final chapter, Chapter 5 presents the conclusion of this thesis, benefits and, limitations and further extension of the system.

CHAPTER 2

THEORITICAL BACKGROUND

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. While natural language processing is not a new science, the technology is rapidly advancing thanks to an increased interest in human-to-machine communications, plus an availability of big data, powerful computing and enhanced algorithms. Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important. Human language is astoundingly complex and diverse. We express ourselves in infinite ways, both verbally and in writing. Not only are there hundreds of languages and dialects, but within each language is a unique set of grammar and syntax rules, terms and slang. When we write, we often misspell or abbreviate words, or omit punctuation. There are several different tasks that NLP can be used to accomplish, and each of those tasks can be done in many different ways. To understand how NLP works, we have to take a look at the two main components of it, NLU and NLG. These two parts of NLP are very different from each other and are achieved by using different methods. The most difficult part of NLP is understanding, or providing meaning to the natural language that the computer received. the computer should understand the meaning of what you said. There are several challenges in accomplishing this when considering problems such as words having several meanings (polysemy) or different words having similar meanings (synonymy), but developers encode rules into their NLU systems and train them to learn to apply the rules correctly. NLG is much simpler to accomplish. NLG translates a computer's artificial language into text, and can also go a step further by translating that text into audible speech with text-to-speech.[13]

2.1 Myanmar Language

The Myanmar language, also known as Burmese, is the official language of the Union of Myanmar and is more than one thousand years old. It is spoken by 32 million. Texts in the Myanmar language use the Myanmar script, is a member of the Tibeto-Burman languages, which is a subfamily of the Sino-Tibetan family of languages, is a phonologically based script, adapted from Mon and is descended from the Brahmi script of ancient South India. Other Southeast Asian descendants, known as Brahmic or Indic scripts, include Thai, Khmer and Lao.[15] Myanmar characters are rounded in shape. Myanmar writing is different from other language because its writing is not used white spaces between words or between syllables. Thus, the computer has to determine syllable and word boundaries by means of an algorithm. Moreover, a Myanmar syllable can be composed of multiple characters. Syllable segmentation is the process of determining word boundaries in a piece of text. Myanmar language consists of one or more morphemes that are linked more or less tightly together. Typically, a word consists of a root or stem and zero or more affixes. Words can be combined to form phrases, clauses and sentences. A word consisting of two or more stems joined together is known as a compound word. Word segmentation is the process of determining morpheme boundaries in a piece of text.

2.1.1 Basic Consonants

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.[16] Myanmar characters can be classified into three groups: consonants, medials and vowels. The basic consonants in Myanmar can be multiplied by medials. Syllables or words are formed by consonants combining with vowels. However, some syllables can be formed by just consonants, without any vowel. Other characters in the Myanmar script include special characters, numerals, punctuation marks and signs.

There are 34 basic consonants in the Myanmar script, as displayed in Figure 2.1. They are known as “Byee” in the Myanmar language. Consonants serve as the base characters of Myanmar words, and are similar in pronunciation to other Southeast Asian scripts such as Thai, Lao and Khmer.

Basic Consonants (ဗျည်းအက္ခရာများ)				
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	

Figure 2.1. Basic Consonants

2.1.2 Vowels

Vowels are known as “Thara”. Vowels are the basic building blocks of syllable formation in the Myanmar language, although a syllable or a word can be formed from just consonants, without a vowel as shown in Figure 2.1.2.1. Like other languages, multiple vowel characters can exist in a single syllable.

Vowels (သရများ)				
◌ေ	◌ိ	◌ီ	◌ာ	◌ဲ
	◌ို	◌ူ		

Figure 2.2 Vowels

2.1.3 Medials

Medials are known as “Byee Twe” in Myanmar. There are 4 basic medials and 6 combined medials in the Myan-mar script as shown in Table 3. The 10 medials can modify the 34 basic consonants to form 340 additional multi-clustered consonants. Therefore, a total of 374 consonants exist in the Myanmar script, although some consonants have the same pronunciation.

Medials (ဗျည်းတွဲများ)			
ျ	ြ	ွ	ှ

Figure 2.3 Medials

2.1.4 Special Characters

Special characters for Myanmar language are used as prescription noun and conjunctions words between two or more sentences.

Special Characters				
ဌ်	၍	၏	ဩော်	ဤ

Figure 2.4 Special Characters

2.2 Syllable Segmentation

Myanmar script uses no space between words and syllable segmentation represents a significant process in many NLP tasks such as word segmentation, sorting, line breaking and so on. Segmentation rules were created based on the syllable structure of Myanmar script and a syllable segmentation algorithm was designed based on the created rules. The lack of official standard encoding hinders localization of Myanmar language and no previous work on the syllable segmentation of Myanmar script was found. Most approaches use a dictionary for syllable segmentation. However, the segmentation accuracy depends on the quality of the dictionary used for analysis and unknown words can reduce the performance. In this study, uses rule-based syllable segmentation.[6]

Table 2.1 Classification of Myanmar Script

Category Name	Name	Glyph	Unicode Code Point
C	Consonants	ကခဂဃငစဆဇဈညဋဌဍဎဏတ ထဒဓနပဖဗဘမယရလဝသဟဠအ	U+1000…U+1021
M	Medials	ျ ြ ွ ှ	U+103B…U+103E
V	Dependent Vowel Sign	ာ ိ ီ ု ူ ေ ဲ	U+102B…U+1032
A	Myanmar Sign Asat	်	U+103A
F	Dependent Various Signs	ံ ့ း	U+1036…U+1038
I	Independent Vowels, Independent Various Signs	ဤ ဧ ဩ ဪ ၌ ၍ ၏	U+1024; U+1027; U+102A; U+104C; U+104D; U+104F;
E	Independent Vowels, Myanmar Symbol Aforementioned	ဣ ဥ ဦ ဩ ၎	U+1023; U+1025; U+1026; U+1029; U+104E;
G	Myanmar Letter Great Sa	ဿ	U+103F
D	Myanmar Digits	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉	U+1040…U+1049
P	Punctuation Marks	၊ ။	U+104A…U+104B
W	White space		U+0020

Table 2.2 Syllable Structure with examples

Syllable	Example	Unicode Point
C	က	U+1000
CF	ကံ	U+1000 U+1036
CCA	ကင်	U+1000 U+1004 U+103A
CCAF	ကင်း	U+1000 U+1004 U+103A U+1038
CV	ကာ	U+1000 U+102C
CVF	ကား	U+1000 U+102C U+1038
CVVA	ကော်	U+1000 U+1031 U+102C U+103A
CVVCA	ကောင်	U+1000 U+1031 U+102C U+1004 U+103A
CVVCAF	ကောင်း	U+1000 U+1031 U+102C U+1004 U+103A U+1038
CM	ကျ	U+1000 U+1038

Where,

C – Consonant V- Vowel

F – Dependent Various Sign M- Medial

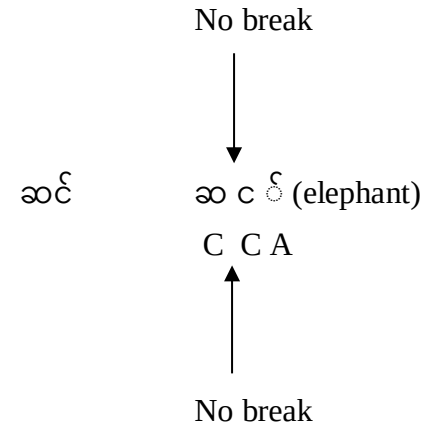
A – Asat

2.2.1 Syllable Segmentation Rules

Typically, a syllable boundary can be determined by comparing pairs of characters to find whether a break is possible or not between them. However, in some cases it is not sufficient to determine a syllable boundary by just comparing two characters. The following sections explain these cases and give examples.[5]

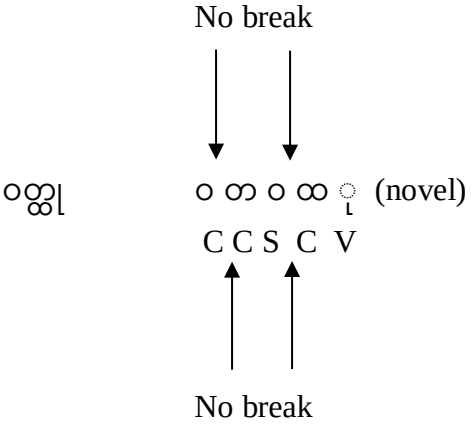
2.2.1.1 Devowelising

In one syllable, a consonant may appear twice but the second consonant is used for the devowelising process in conjunction with an Asat (U+103A MYANMAR SIGN ASAT). Therefore the character after the second consonant should be further checked for an Asat. If the character after the second consonant is an Asat, there should be no syllable break before the second consonant.



2.2.1.2 Syllable Chaining

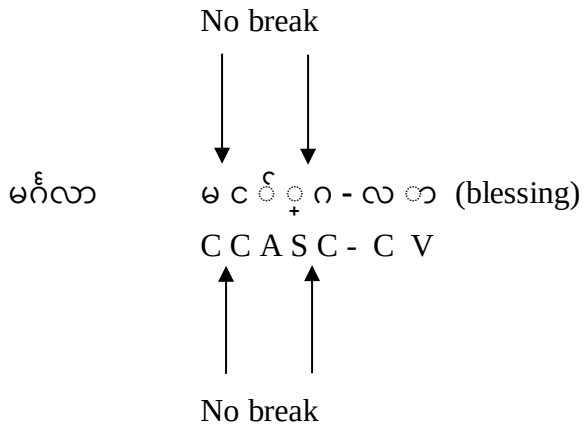
Subjoined characters are shown by using an invisible Virama sign (U+1039 MYANMAR SIGN VIRAMA) to indicate that the following character is subjoined and should take a subjoined form. In this case, if the character after the second consonant is an invisible Virama sign, there should be no syllable break before the second and third consonant. Although there are two syllables in a subjoined form, it is not possible to separate them in written form and they are therefore treated as one syllable.



2.2.1.3 Kinzi

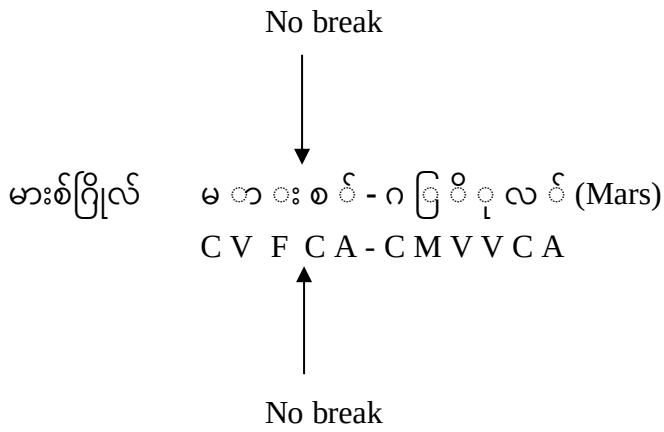
Kinzi is a special form of devowelised Nga (U+1004 MYANMAR LETTER NGA) with the following letter underneath, i.e., subjoined. In this case, if the character after the second consonant is an Asat and the next character after Asat is an invisible Virama sign (U+1039 MYANMAR SIGN VIRAMA) then there should be

no syllable break before the second and third consonant. Kinzi also consists of two syllables but it is treated as one syllable in written form.



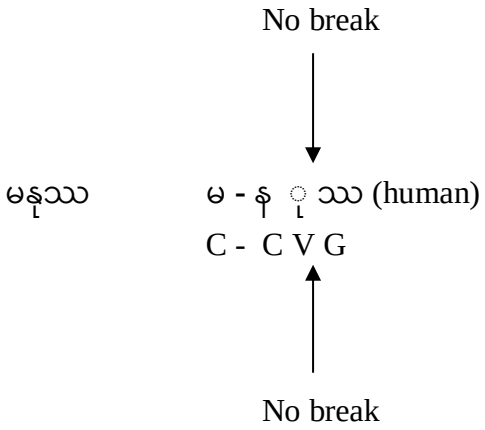
2.2.1.4 Loan Words

Usage of loan words can be found in Myanmar text. Although loan words do not follow the Myanmar syllable structure, their usage is common and the segmentation rules for these words are considered in this study.



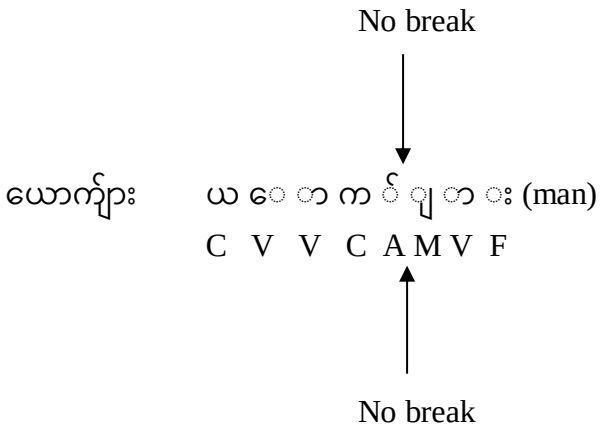
2.2.1.5 Great Sa

There should be no syllable break before great Sa (U+103F MYANMAR LETTER GREAT SA) as great Sa acts like a stacked သ. and devowelises the preceding consonant.



2.2.1.6 Contractions

There are usages of double-acting consonants in Myanmar text. The double acting consonant acts as both the final consonant of one syllable and the initial consonant of the following syllable. There are two syllables in a contracted form but they cannot be segmented in written form and there should be no syllable break between them.



2.3 Word Segmentation

Formally, the task of Myanmar word segmentation is the process of the most important subtask of natural language processing. The Myanmar script is an *abugida* (*alphasyllabary*) in the Brahmic family, containing 33 basic consonant letters. Each standalone consonant letter can form a complete syllable by itself with the help of an inherent vowel. The inherent vowel can be changed to other vowels through the use of various diacritic marks. Further diacritic marks include tone marks, dependent consonant marks for syllable onset clusters, and a *virama* ("*asat*", which means *kill* in Burmese) used to suppress the inherent vowel of a consonant letter for nasal or glottal syllable.The first apply a syllable segmentation process on input, that is, semgent into syllable boundaries, and then decide how the syllables form words.[12]

2.3.1 Dictionary-Based Matching

Maximum matching by looking in a prepared dictionary and matching the longest substring in an input sentence is a classic word segmentation approach for other languages. The matching can be conducted from the beginning of a sentence to its end or in reverse. The former method is referred to as forward maximum matching (fmm) and the latter as backward (or reverse) maximum matching (rmm). The two directional processes can be combined to form a bi-directional maximum matching (bmm), in which further heuristic rules are used to select the better result from those of fmm and rmm. Although the maximum matching approach is simple, its performance can be mediocre, and it is typically used as a baseline approach in word segmentation tasks. Because the simplicity and speed of this classic approach is a great advantage, the approach is still widely used in practical engineering and has been studied in recent research.[7]

2.3.2 Statistical Approaches

Word segmentation using a statistical language model is more reasonable than dictionary-based matching, because it uses the probabilities of words in real textual data. Consequently, segmentation results containing more common words are better than those containing obscure words. Statistical approaches correspondingly require more training data than a dictionary-based approach. Given a statistical language

model, for example, an N-gram language model, the word segmentation task becomes a search problem to determine the segmentation with the highest probability according to the model. A simple Viterbi-like (or Dijkstralike) dynamic programming algorithm can be applied by scanning an input sentence to generate the best segmentation up to each syllable until the best segmentation of the entire sentence is constructed. Other N-gram language models are the maximum-likelihood estimated uni-gram model (uni.mle), the absolute discounting uni-gram model (uni.abs), and the modified Kneser-Ney discounting uni-, bi-, and tri-gram models (uni.mkn, bi.mkn, and tri.mkn, respectively).[14]

2.3.3 Machine Learning Approaches

Word segmentation task can be treated as a classification task in a machine learning framework, or, more specifically, a sequence labeling task, due to the properties of textual data, and several standard learning frameworks have been established and developed. Almost them, CRF (Conditional Random Fields) in the CRF++ toolkit and SVM (Support Vector Machine) in the KyTea toolkit. Feature engineering for input and tag-set design for output are import issues for machine learning approaches.[12]

2.3.3.4 Expectation Maximization Approach

Many unsupervised methods have been proposed for segmenting raw character sequences with no boundary information into words. Most current approaches are based on using some form of EM to learn a probabilistic speech or text model and then employing Viterbi-decoding-like procedures to segment new speech or text into words. One reason that EM is widely adopted for unsupervised learning is that it is guaranteed to converge to a good probability model that locally maximizes the likelihood or posterior probability of the training data. For the problem of word segmentation, EM is typically applied by first extracting a set of candidate multi-grams from a given training corpus, initializing a probability distribution over this set, and then using the standard iteration to adjust the probabilities of the multi-grams to increase the posterior probability of the training data.[8]

2.4 Performance Measure

Performance measure is the process of collecting, analyzing and reporting information regarding the performance of and individual, group, organization, system or component. One of the most important aspects to be considered in relation to performance measure process is that the performance measures work qualitatively to provide the useful information about product, process, system or component. Implementation of performance measure is a great way to understand and manage and improve what a group, organization, system or component does.

2.4.1 Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if a system have high accuracy then that system is best. Yes, accuracy is a great measure but only when the system have symmetric datasets where values of false positive and false negatives are almost same.

Accuracy = (TP+TN)/(TP+FP+FN+TN),

where, TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

True positive and true negatives are the observations that are correctly predicted. False positive and false negatives, these values occur when actual class contradicts with the predicted class. These terms are a bit confusing.

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. E.g. if actual class value indicates that this passenger survived and predicted class tells the same thing.

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. E.g.

if actual class says this passenger did not survive and predicted class tells the same thing.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells that this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class valued indicates that this passenger survived and predicted class tells that passenger will die.

2.4.2 Precision, Recall and F1 Score

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High prevision relates to the false positive rate.

Precision = TP/TP + FP

Recall is the ratio of correctly predicted positive observations to the all observations in actual class – yes. The question recall answers is: OF all the passengers that truly survived, how many did label?

Recall = TP/TP + FN

F1 Score is the weighted average of Prevision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.[17]

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

CHAPTER 3
DESIGN OF THE SYSTEM

The system determine the boundaries of words for those languages that do not have word separators in orthography. In this system, client can segment Myanmar word from Myanmar sentences as line by line paragraph or as a paragraph by paragraph. The word “segmentation” is useful for quick access to the data, when one wants to hear the acoustic realization of a certain word.

3.1 Overview of System Design

This system has three steps: Text Analysis, Syllable Segmentation, and Hybrid Word Segmentation. Last step is hybrid so it has another two approaches: Dictionary Approach and Expectation Maximization Approach.

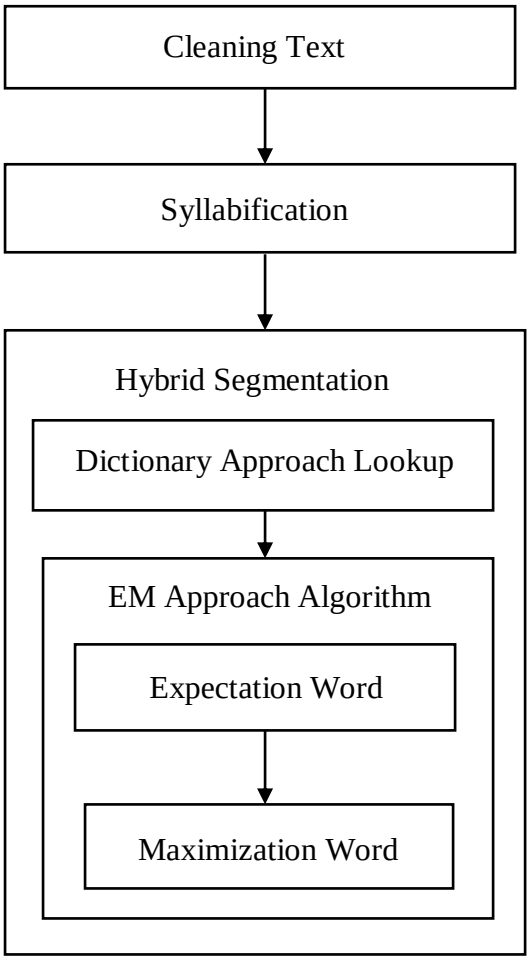


Figure 3.1 Design of the Myanmar Word Segmentation

3.2 Class Diagram of the System

There are four classes in this system. They are Main, Tokenization, DictionaryApproach and EMApproach. The Main class includes cleanText function, syllabification function and segmentation function. Tokenization has tokenization function, tokenize function, tokenizedWordList function and tokensyllabification function. The third class is DictionaryApproach that is composed of getDataList function, checkInList-ForDictionary function and changeArrayListtoString function. The last class is EMApproach. It contains getDataList function, checkInListForEM function, Expectation function and Maximization function.

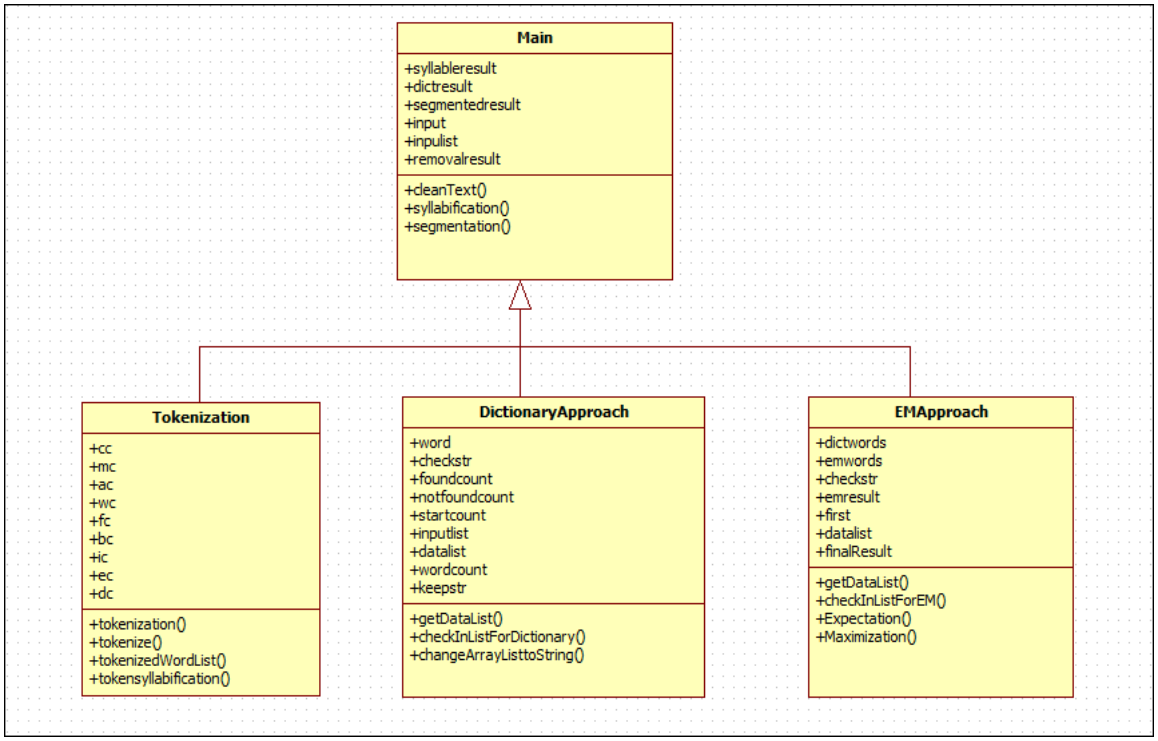


Figure 3.2 Class Diagram of the Myanmar Word Segmentation

3.3 Procedure of system

In this system, client input Myanmar text as line by line sentences or paragraph. We clean input text by using Text Analysis. After that, segment syllables from output of Text Analysis by using SyllableSegmentation. And then, segment words from output of Syllable Segmentation by using Hybrid Word Segmentation.

3.3.1 Cleaning Text

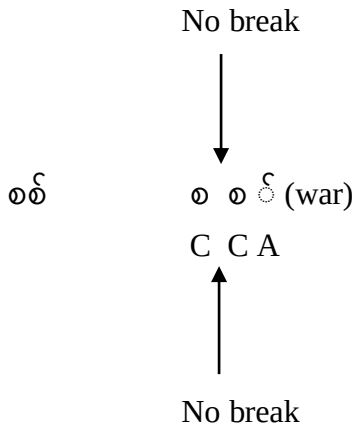
When input Myanmar text, will perform text analysis. Text analysis takes input in the form of text and remove unwanted words (“-”, “॥”, “।”, “(”, “)”, “?”)etc.

3.3.2 Syllabification

Syllable Segmentation rules were created based on the syllable structure of Myanmar script and a syllable segmentation algorithm was designed based on the created rules. A syllable boundary can be determined by comparing pairs of characters to find whether a break is possible or not between them. However, in some cases it is not sufficient to determine a syllable boundary by just comparing two characters. So, another syllable segmentation rules are used. In this system, Devowelising and Great Sa has been used.

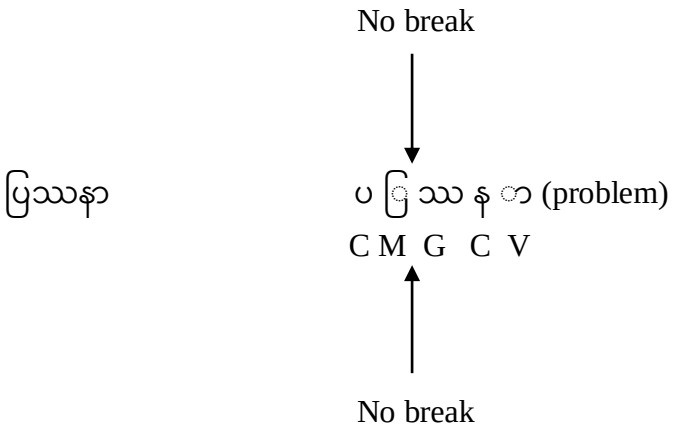
3.3.2(A) Devowelising Process

In one syllable, a consonant may appear twice but the second consonant is used for the devowelising process in conjunction with an Asat (U+103A MYANMAR SIGN ASAT). Therefore the character after the second consonant should be further checked for an Asat. If the character after the second consonant is an Asat, there should be no syllable break before the second consonant.



3.3.2(B) Great Sa Process

There should be no syllable break before great Sa (U+103F MYANMAR LETTER GREAT SA) as great Sa acts like a stacked သ and devowelises the preceding consonant.



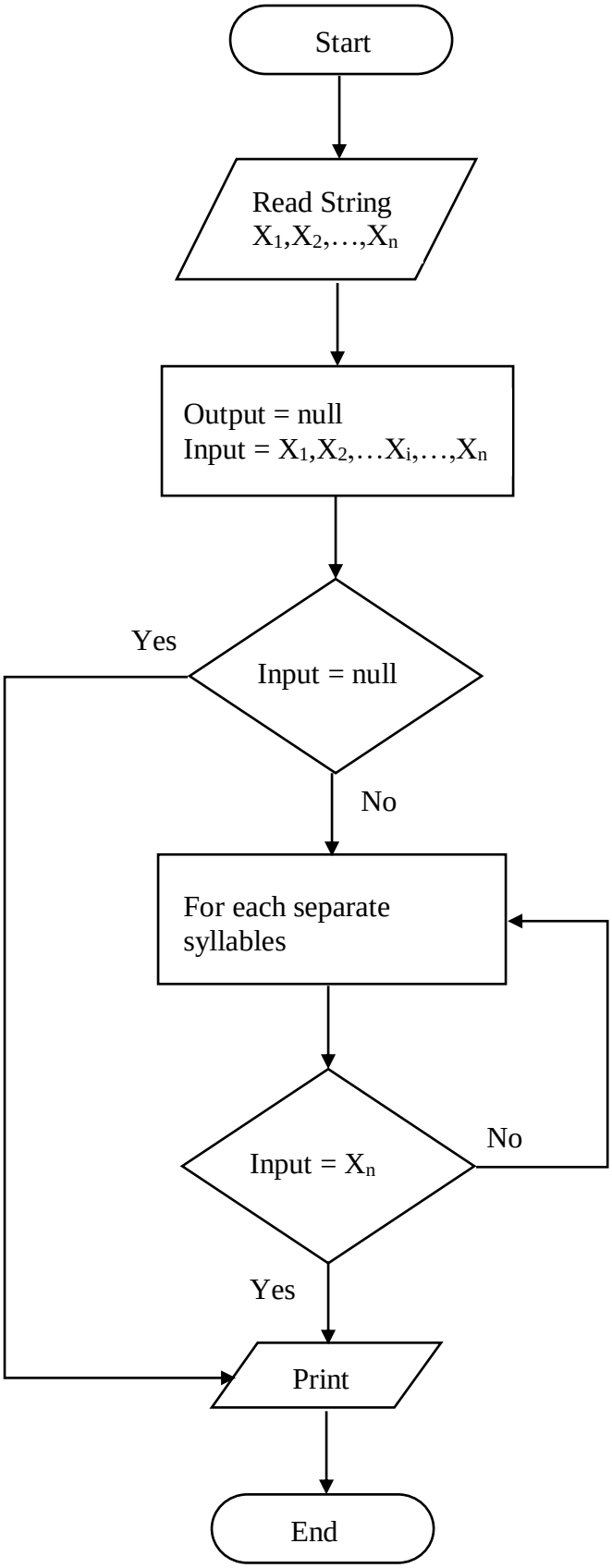


Figure 3.3 Syllable Segmentation Flow Chart

3.3.3 Hybrid Word Segmentation

Words for Myanmar language without word separator in orthography are not delimited by white-space but instead must be inferred from the basic character sequence. Moreover, Myanmar syllables can also part of multi-syllable words whose syllables are separated by word segmenter between them. We present a hybrid approach to automatically segment Myanmar text. The approach combines both dictionary approach and expectation maximization approach.

3.3.3.1 Dictionary Approach

Dictionary approach use maximum matching method. Maximum matching is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation. Maximum matching by looking in a prepared dictionary, extracting segments based on syllabled words and matching the longest substring in an input sentence is a classic word segmentation approach.

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

3.3.3.2 Hierarchical Expectation Maximization Approach

The Expectation Maximization Approach is used for unsupervised text segmentation. That consists of two processes: The E-step, and the M-step. In the E-step, generate all morphemes from the training corpus *C*, learn a probability distribution over morphemes, and segment the original training corpus *C* into a morpheme sequence *G*. In the M-step, maximize a probability distribution over segment G and segment G into a word sequence W.[2]

Overall, the E-step determines,

$$G^* = prob(G,C) \quad \text{---------- Equation 1}$$

and the M-step determines,

$$W^* = \max\{G^*\} \quad \text{---------- Equation 2}$$

The Expectation Maximization algorithms in both levels are identical except that in the E-step the basic observation unit is character and in the M-step the basic unit is morpheme.

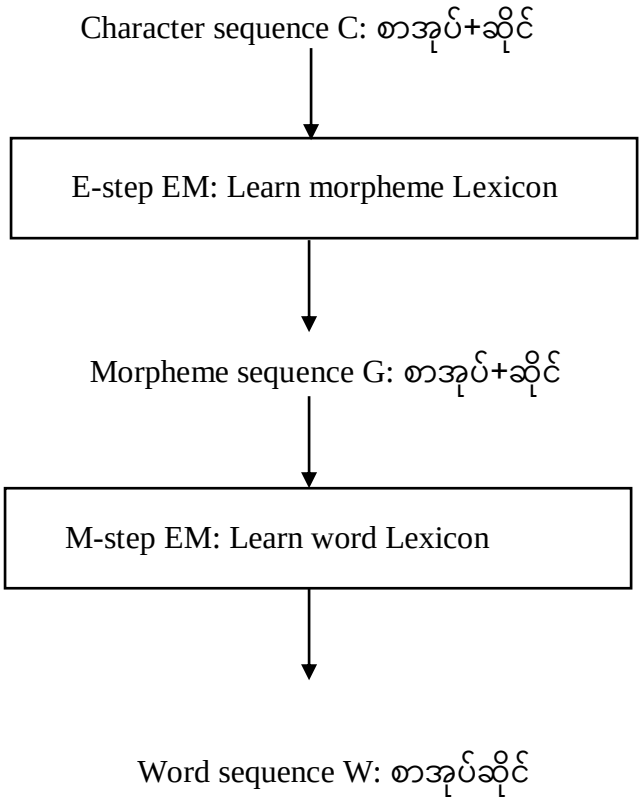


Figure 3.4 Expectation Maximization Model

CHAPTER 4
IMPLEMENTATION OF THE SYSTEM

This chapter presents how to implement the system. There are seven algorithms in the system. They are cleanText Algorithm for Text Analysis, syllabification Algorithm for Syllibification, hybridSegmentText Algorithm for Hybrid EM Segmentation Approach, dictionaryApproach Algorithm for Dictionary Approach, EMApproach Algorithm for Expectation Maximization Approach, expectation Algorithm for Expectation Approach, maximization Algorithm for Maximization Approach.

4.1 cleanText Algorithm

In cleanText algorithm, the input is Myanmar text, this step removes unwanted words (“-”, “॥”, “।”, “(”, “)”, “?”) etc. and the output is the Myanmar text that clean unwanted words.

1. function cleanText(text)
Input : Myanmar Text
Output : clean text
Process :
2. **begin**
3. Define unwanted symbols as {”/”,”?”,…}
4. Remove these symbols from input text
5. Return clean text
6. **end**

4.2 syllabification Algorithm

In syllabification algorithm, the input is the output of cleanText algorithm. This step combines both devowelising and great sa processes. The devowelising process is the process of segmentation between consonants. If the character after the second

consonant is an Asat, there should be no syllable break before the second consonant. Great Sa process is the segmentation process based on collocation word like a stacked သ. The output is the syllabled word of input text.

```
1. function syllabification(text)
Input        : output of cleanText Algorithm
Output       : syllable word of input text
Process      :
2.    begin
3.    cc = {"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"};
4.    ac = {"–ိ"};
5.    fc = {"–ု"};
6.    bc = {"–ံ", "– း"};
7.    ic = {"ဣ","ဥ","ဿ", "ဧ","ြော","ၚ","သ"};
8.    ec = {"ဣ" ,"ဥ", "ဦ","ဉ","ြ", "ဎ"};
9.    dc = {"ဝ","ဿ","ျ","ဲ","ွ","ှ","ဖ","ဂ","ဈ","ဧ"}
10.   let character  = output of cleanText Algorithm
11.   for(; i<character.length; i++)
12.   begin
13.   if ic contain character[i] then
14.   begin
15.       save character[i] to wordForm
16.       flag = true;
17.   end
18.  else
19.     begin
```

```
20.   if cc contain character[i] or ec contain character[i] then
21.                 begin
22.                     int j = i;
23.                     boolean skip = false;
24.   if (j+1) is less than the size of character and ac contain
      character [++j] or wc contain character[i] or fc contain character[j] then
25.             begin
26.                 save character[i] to word
27.           skip = true;
28.       end
29.   if (i-1) is greater than or equal 0 and wc contain character[j-2] then
30.       begin
31.           save character[i] to word
32.           skip = true;
33.       end
34.       if skip is not true then
35.       begin
36.          save character[i] to wordForm
37.         if i is equal 0 then
38.            save word to wordForm
39.         else
40.            flag = true;
41.               end
42.   end
43.   else
44.        save character[i] to word
45. end
46. if flag is true then
47.        begin
```

```
48.          if word is not null then
49.                save word to wordlist
50.             save wordForm to word
51.             flag = false;
52.    end
53.    end
54.     if word is not null then
55.          save word to wordList
56.      return wordList
59.  end
```

4.3 hybridSegmentText Algorithm

The hybridSegmentText algorithm accepts the output of syllabification. This step combines both dictionaryApproach function and EMApproach function. The dictionaryApproach strives to segment using the longest possible segments. The EMApproach is used for unsupervised text segmentation. The output is segmented word of input text.

```
1. function hybridSegmentText(text)
Input          : output of syllabification algorithm
Output         : segment text
Process        :
2.       begin
3.       define terminater symbol as "॥"
4.       save split with terminater input text to inputstring
5.       foreach input in inputstring
6.       begin
7.          if input is not null then
8.          begin
```

```
11.         call function syllabification(input)
12.         save returned syllabledword
13.         save wordlist splitting by "+"  of syllabledword
14.         for(int i=0; i< wordlist ; i++)
19.         begin
22.             call function dictionaryApproach(syllabledword)
23.             save returned dictresult
24.         end
31.         for(int i=0; i< wordlist ; i++)
32.         begin
35.             call function EMApproach(dictresult)
36.             save returned emresult
37.         end
44.       end
45.     end
46.  end
```

4.3.1 dictionaryApproach Alogorithm

The dictionaryApproach algorithm accepts the output of syllabification. This step strives to segment using the longest possible segments. The output is the longest substring of words.

```
1. function dictionaryApproach(text)

Input       : output of syllabification Algorithm
Output      : longest substring of words
Process     :
2.      begin
3.      Let syllabledword = output of syllabification Algorithm
4.      save split with "+" from syllabledword to  inputlist
5.     for(int i=0; i<inputlist.Length; i++)
```

```
6.      begin
7.       call function getDataList(inputlist[i], "");
8.       save returned list to dataList
9.       while(i<inputlist.length)
10.          begin
11.           save inputlist[i] to checkstr
12.           check checkstr with dataList
13.           if found then
14.           begin
15.                save checkstr to word
16.           end
17.           else
18.           begin
19.               if word is not null
20.               begin
21.                       keepstr.Add(word);
22.                end
23.           end
24.          end
25.          if word is not null then
26.              save word to keepstr
27.     end
28.   end
```

4.3.2 EMApproach Algorithm

The EMApproach algorithm accepts the output of dictionaryApproach algorithm. This step segments unsupervised text and combines the E-step and the M-step. The output is the segmented text.

```
1. function EMApproach(text)
Input          : output of dictionaryApproach Algorithm
Output         : segment text
Process        :
2.      begin
3.       Let dictionary_result = output of
4.       save split with "+" from dictionary_result to dictwords
5.       foreach letter in dictwords
6.       begin
7.        save letter to checkstr
8.        call function getDataList(checkstr, "corpus.txt");
9.        save returned list to dataList
10.       check checkstr with dataList
11.       if found then
12.       begin
13.            save match with previous letter into emwords
14.            remove previous letter from finalResult
15.            save letter to emwords
16.           end
17.       else
18.       begin
19.               if emwords not null then
20.               begin
21.                       call function Expectation(emwords, dataList)
22.                       save returned expectation result to dict
23.                       call function Maximization(dict)
24.                       save returned maxmimum result to  finalResult
25.               end
```

```
26.     save letter to finalResult
27.             save space to checkstr
28.        end
29.     end

30.     if emwords not null then
31.      begin
32.             call function expectation(emwords, dataList)
33.             save returned expectation result to dict
34.             call function Maximization(dict)
35.             save returned maxmimum result to finalResult
36.             save space to emwords
37.      end
38.     call changeArrayListtoString(finalResult)
39.     save returned changestring to emresult;
40. end
```

4.3.2(A) expectation Algorithm

The expectation algorithm accepts the output of unsupervised data list and data list for checking. This step generate all morphemes from the training corpus *C*, learn a probability distribution over morphemes, and segment the original training corpus *C* into a morpheme sequence *G*. The output is probability of unsupervised word.

```
1. function expectation(datalist, checkdatalist)
Input       : word of dictionaryApproach Algorithm's datalist, checkdatalist
Ouput       : key and value based on word
Process     :
2.    begin
3.    Let check_text = output of dictionaryApproach Algorithm
```

```
4.      save split with "+" from check_text to syllabledword
5.      save syllabledword[0] to checkstr
6.      foreach letter in syllabledword
7.      begin
8.       check check_text with checkdatalist
9.       if found then
10.      begin
11.             if (i+1) less than syllabledword.length then
12.                    save later letter to checkstr
13.             else
14.             begin
15.                    call function getCount(checkdatalist, "total", "")
16.                    save returned data to total
17.                    call function getCount(checkdatalist, "per", checkstr)
18.                     save returned data to per
19.                    save checkstr as key and (per/total) as value to dict
20.             end
21.          end
22.          else
23.          begin
24.             if (checkstr is not contain in dict) then
25.                   save checkstr as key and 0.0 as value to dict
26.           end
27.       end
28.  end
```

4.3.2(B) maximization Algorithm

The maximization algorithm accepts the output of expectWord algorithm. This step is maximize a probability distribution over segment G and segment G into a word sequence W. The output is maximization result.

```
1. function maximization(text)
Input          : output of expectWordAlgorithm
Output         : maximization result
Process        :
2.      begin
3.      save key of expectation result to keylist
4.      save value of expectation result to valueList
5.      for(int i=0; i<valueList.Length; i++)
6.      begin
7.        save keylist[i] to sumstr
8.        save valueList[i] to sumValues
9.       end
10.       save sumstr to keylist
11.       save sumValues to valueList
12.       save valueList[0] to max
13.      for(int i=0; i<valueList.Length; i++)
14.      begin
15.              if max less than or equal valueList[i] then
16.              begin
17.                      save valueList[i] to max
18.                      save i to index
19.              end
20.       end
21.       save keyList[index] to result
22.  end
```

4.4 Implementation of the system

The implementation of this system is tested by using Myanmar texts new contents relating “Business”, “Entertainment” and “Sports”. The input text can be used as line by line format or one paragraph format. The main page can be shown in figure 4.1.

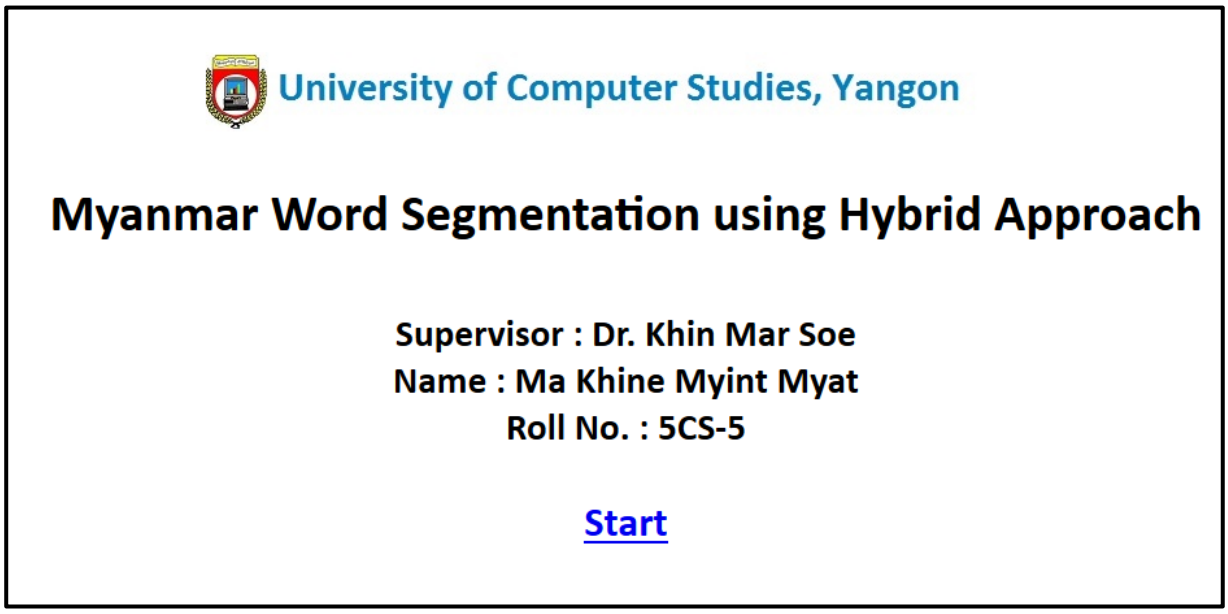


Figure 4.1 First Page of the System

After clicking the start button, the following page can be seen as in figure 4.2.

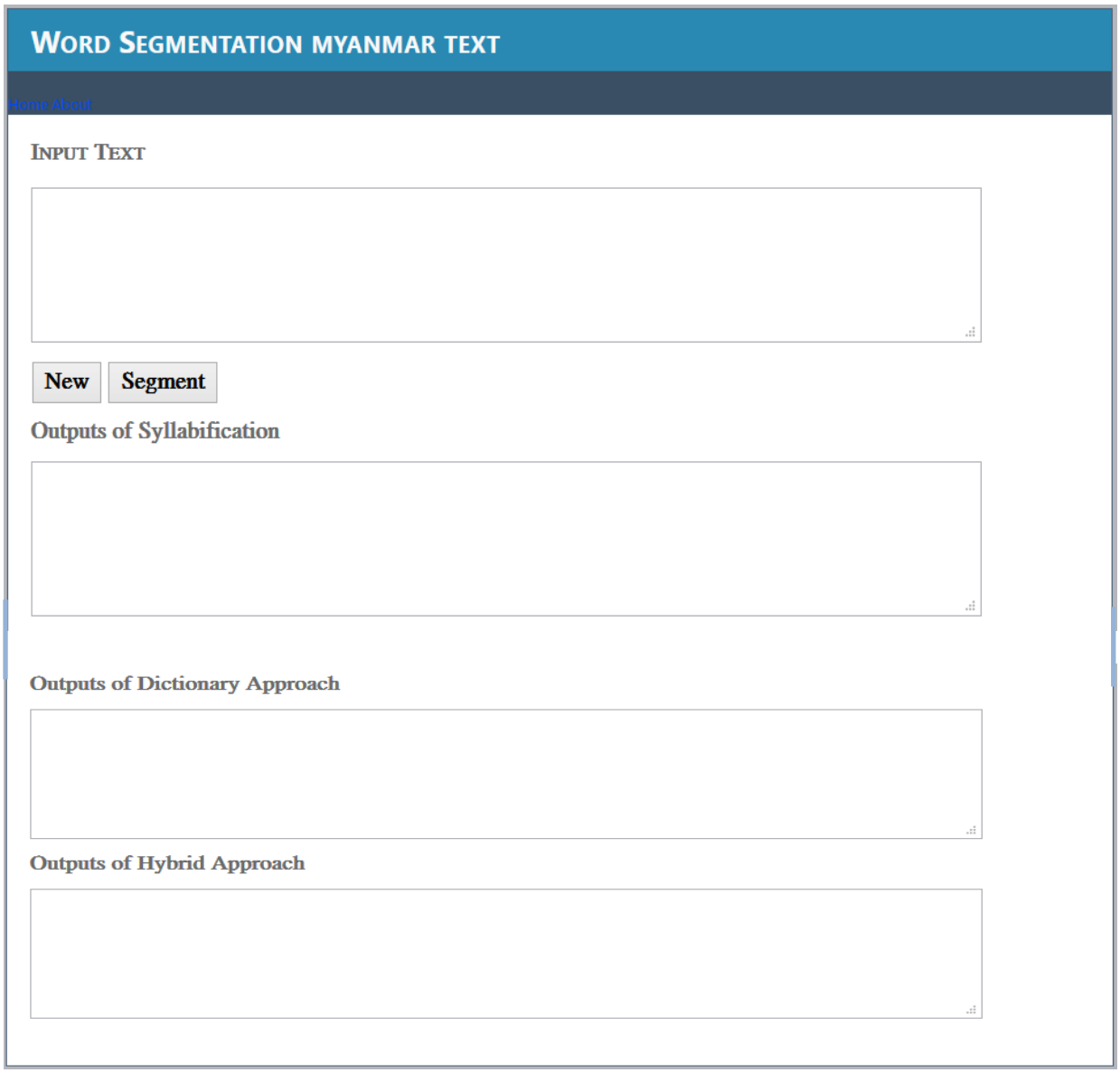


Figure 4.2 Second Page of the System

4.4.1 Evaluation of Business article

This system can be accessed as line by line paragraph format or one paragraph frormat. The input text as can be placed in the input text box that has line by line paragraph format as shown in figure 4.3 and one paragraph format as shown in figure 4.4.

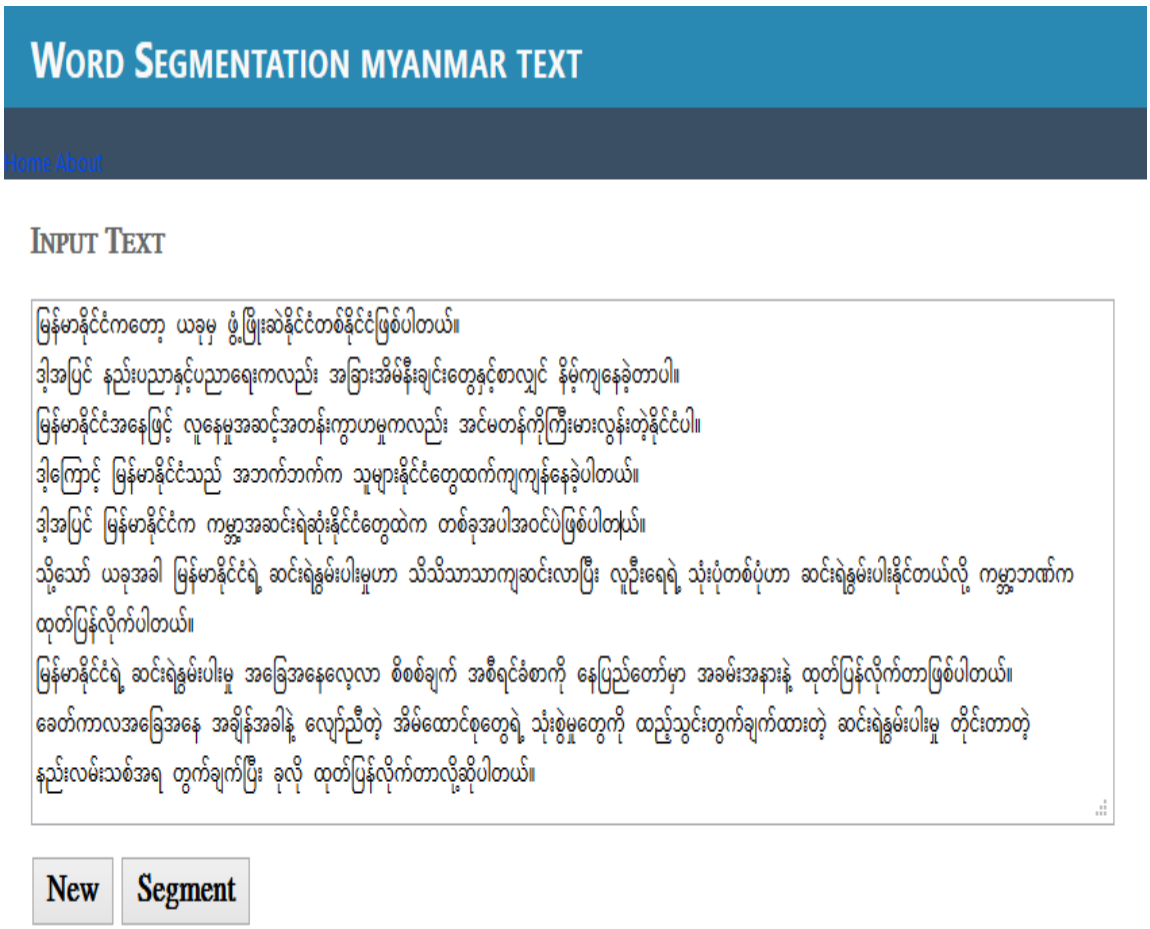


Figure 4.3 Input Text for business article with line by line paragraph format

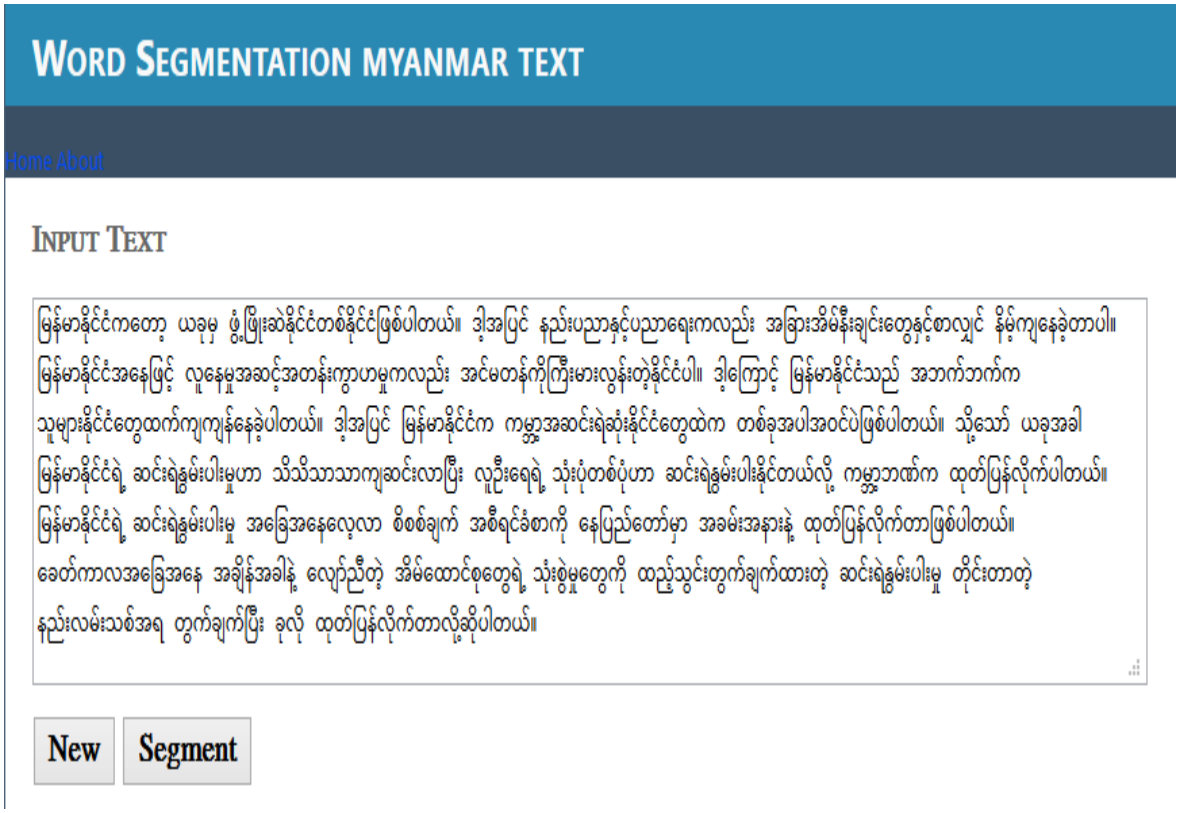


Figure 4.4 Input Text for business article with one paragraph format

If user clicks segment button, the following output of syllabification can be seen as shown in figure 4.4. The output of syllabification can be shown as line by line paragraph format.

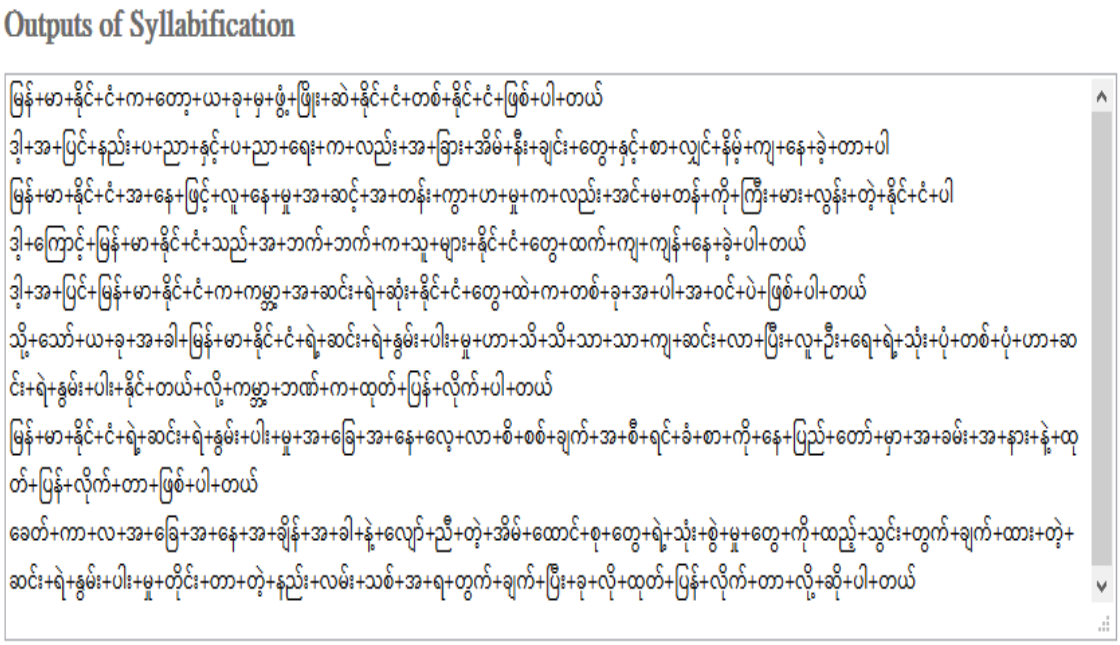


Figure 4.5 Outputs of Syllabification for business article

The input text for dictionary approach is the output of syllabification and the output of dictionary approach can be placed as shown in figure 4.6.

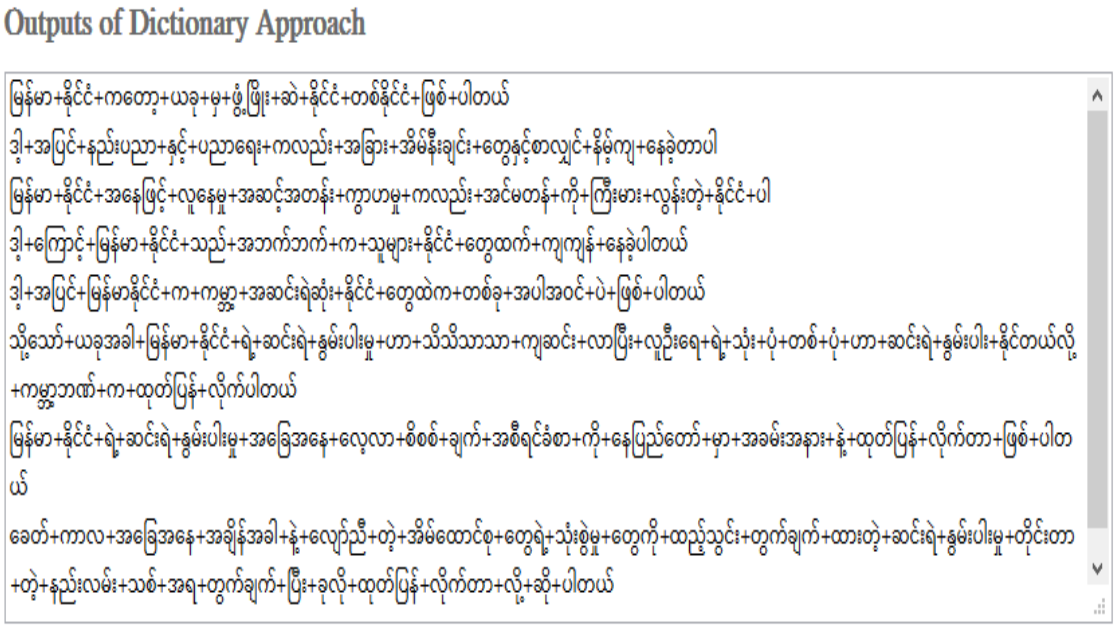


Figure 4.6 Outputs of Dictionary Approach for business article

The input text is the output of dictionary approach, the output of hybrid approach can be seen as shown in figure 4.7.

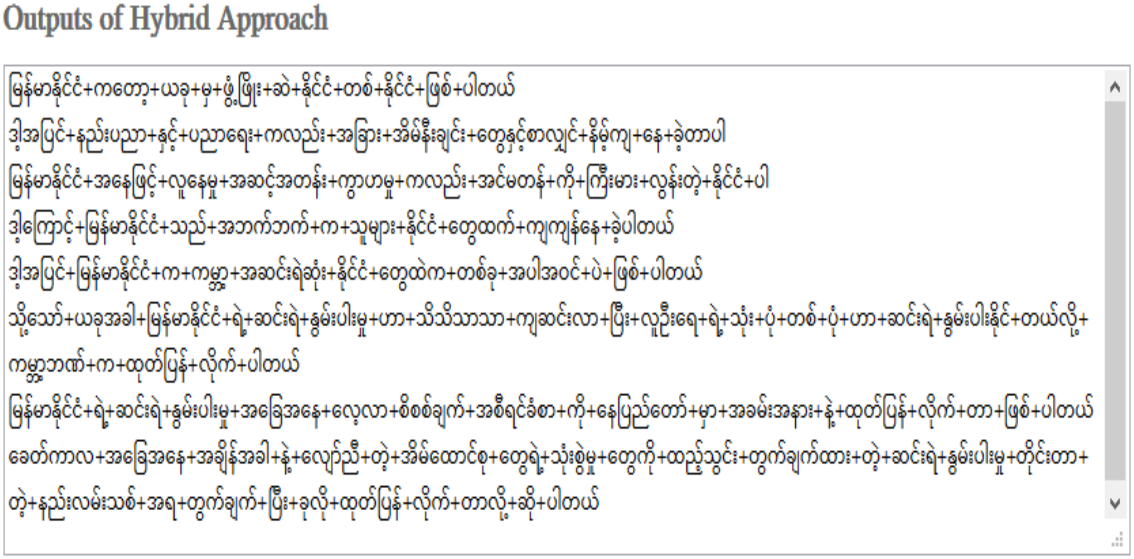


Figure 4.7 Outputs of Hybrid Approach for business article

4.4.2 Evaluation of Entertainment

This system can be accessed as line by line paragraph format or one paragraph frormat. The input text as can be placed in the input text box that has line by line paragraph format as shown in figure 4.8 and one paragraph format as shown in figure 4.9.

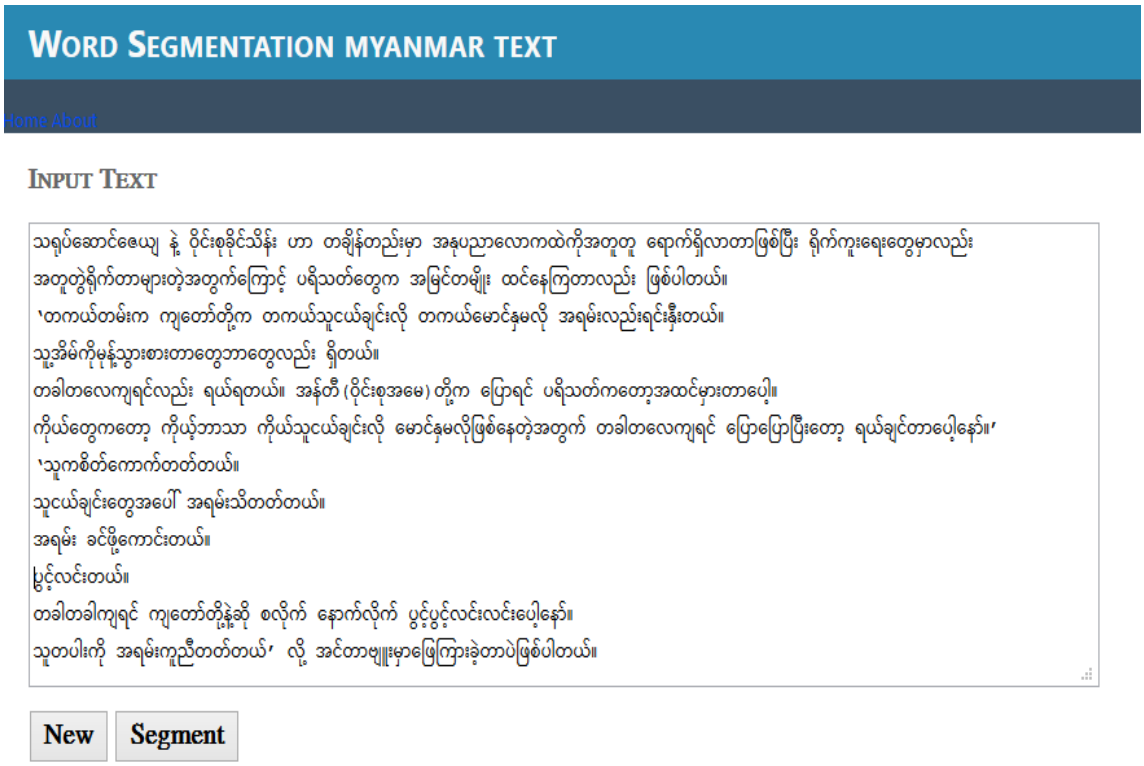


Figure 4.8 Input Text for entertainment article with line by line paragraph format

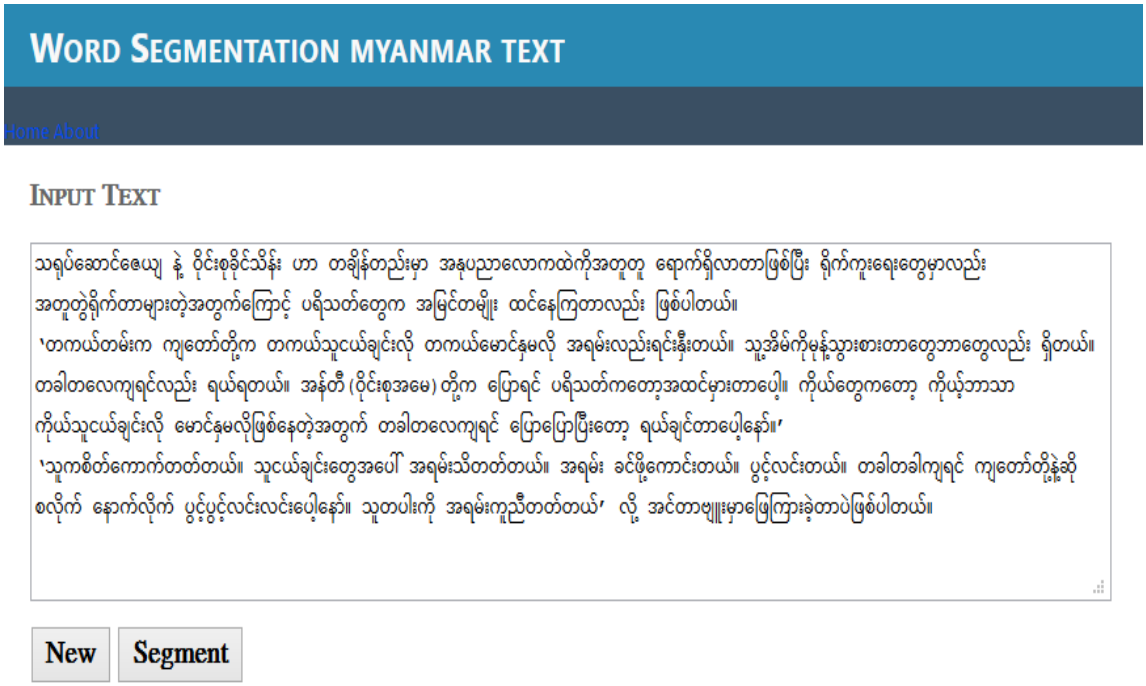


Figure 4.9 Input Text for entertainment article with one paragraph format

If user clicks segment button, the following output of syllabification can be seen as shown in figure 4.10.

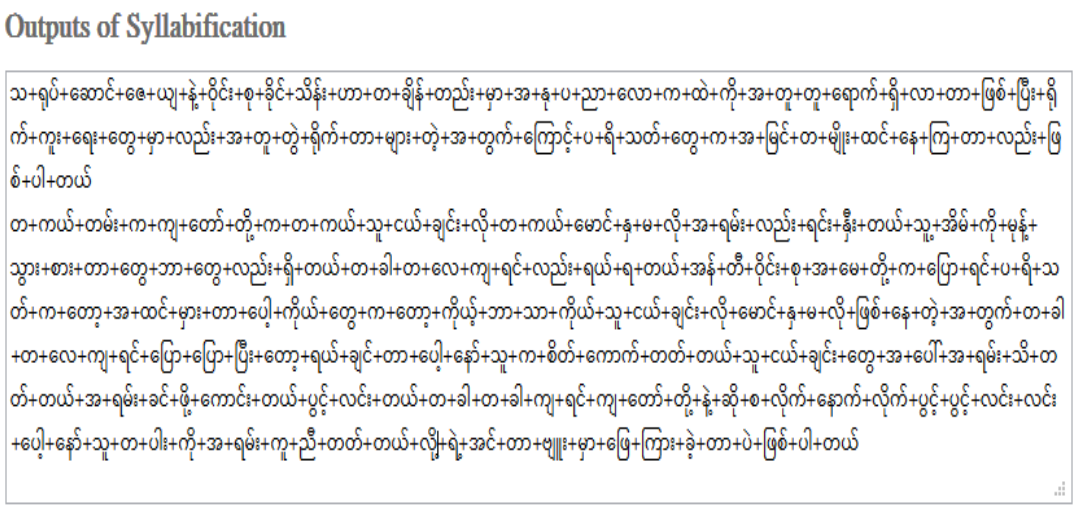


Figure 4.10 Outputs of Syllabification for entertainment article

The input text is the output of syllabification and the output of dictionary approach can be placed as shown in figure 4.11.

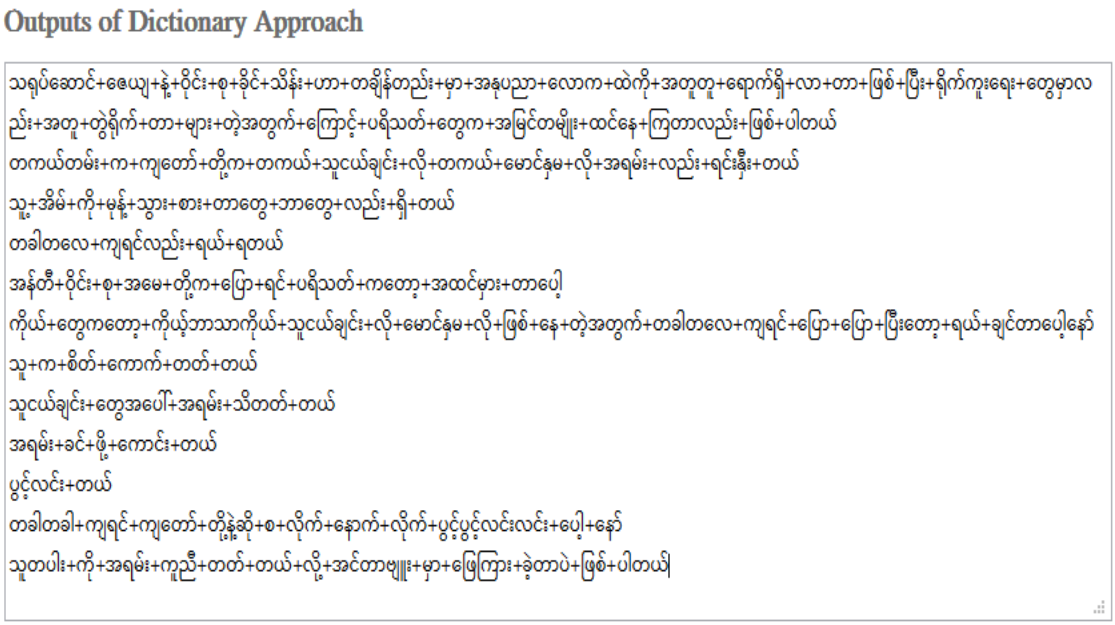


Figure 4.11 Outputs of Dictionary Approach for entertainment article

The input text is the output of dictionary approach, the output of hybrid approach can be seen as shown in figure 4.12.

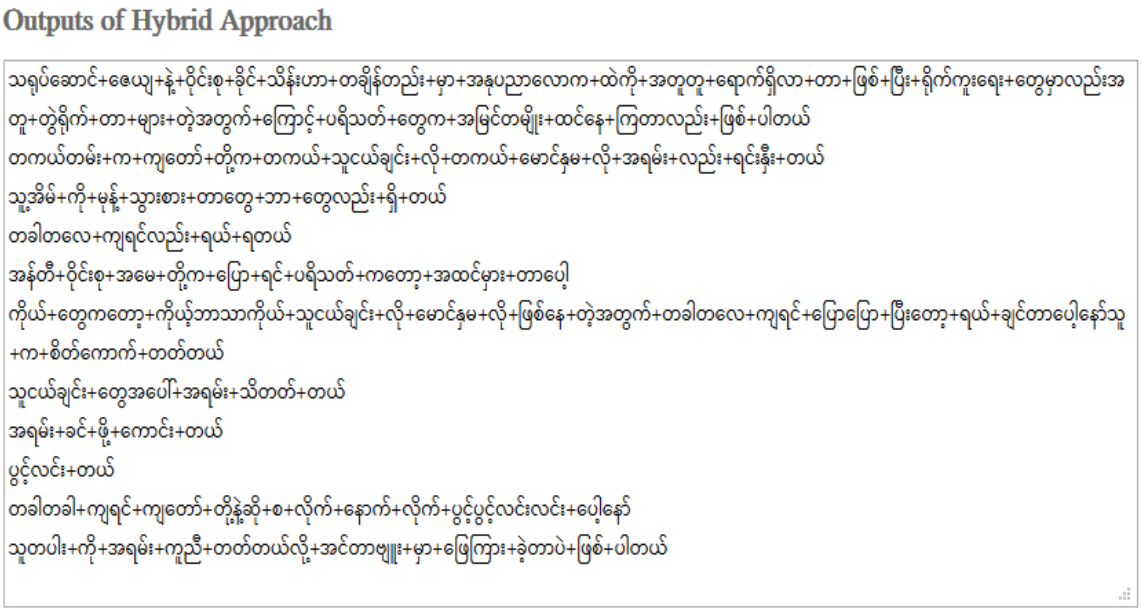


Figure 4.12 Outputs of Hybrid Approach for entertainment article

4.4.3 Evaluation of Sports article

This system can be accessed as line by line paragraph format or one paragraph frormat. The input text as can be placed in the input text box that has line by line paragraph format as shown in figure 4.13 and one paragraph format as shown in figure 4.14.

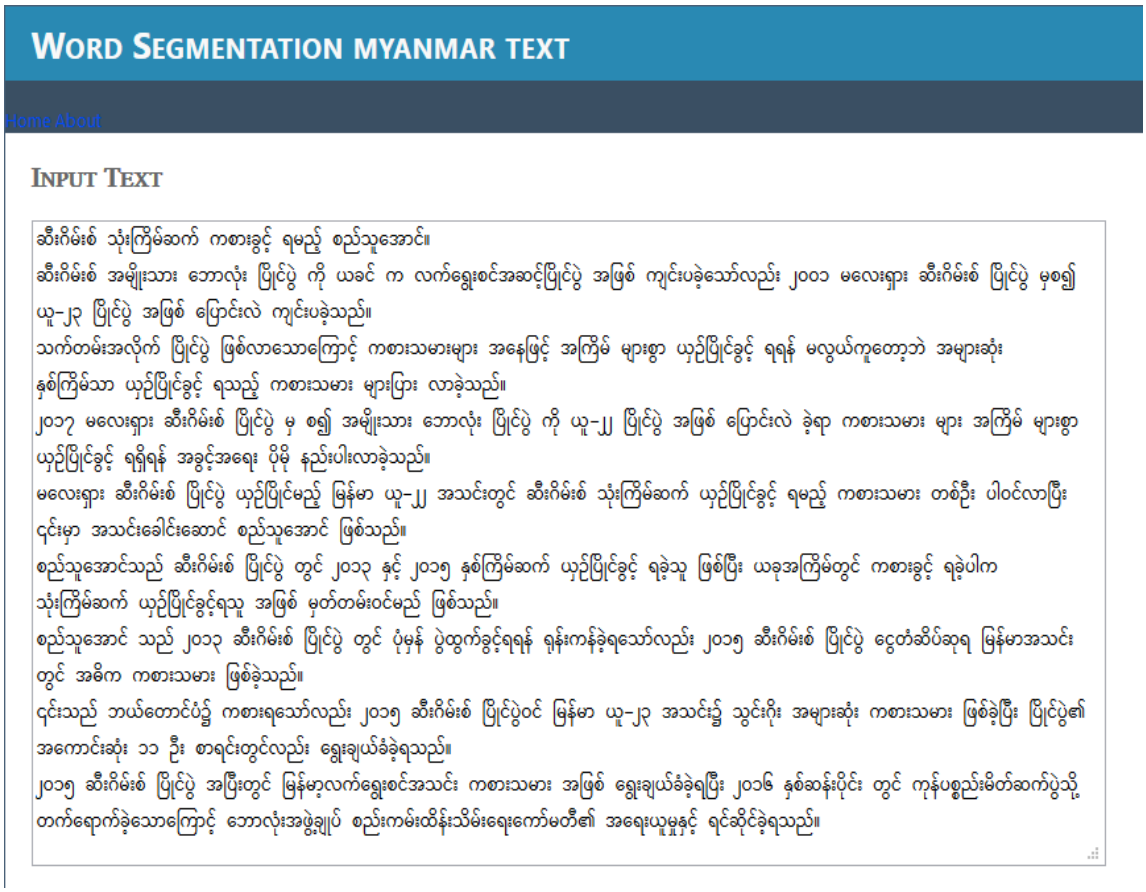


Figure 4.13 Input Text for sports article with line by line paragraph format

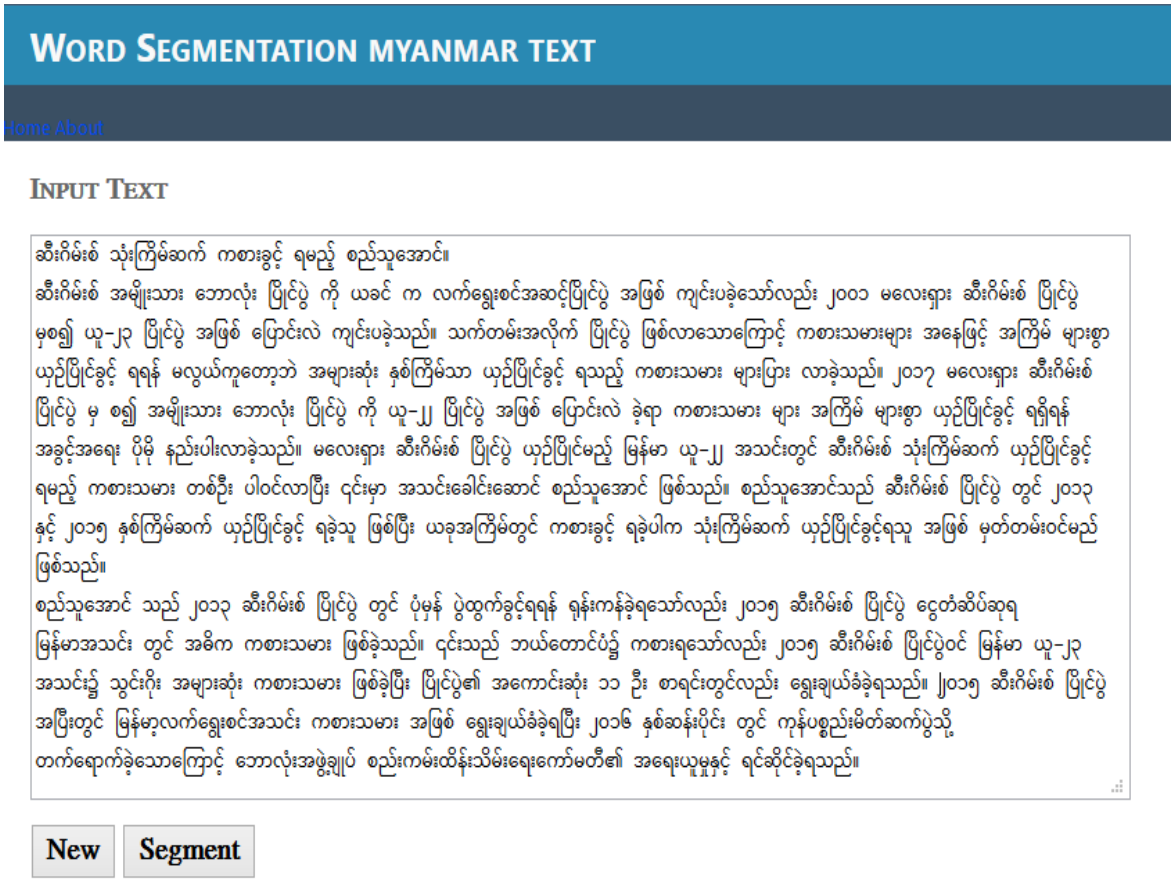


Figure 4.14 Input Text for entertainment article with one paragraph format

If user clicks segment button, the following output of syllabification can be seen as shown in figure 4.15.

Figure 4.15 Outputs of Syllabification for sports article

The input text is the output of syllabification and the output of dictionary approach can be placed as shown in figure 4.16.

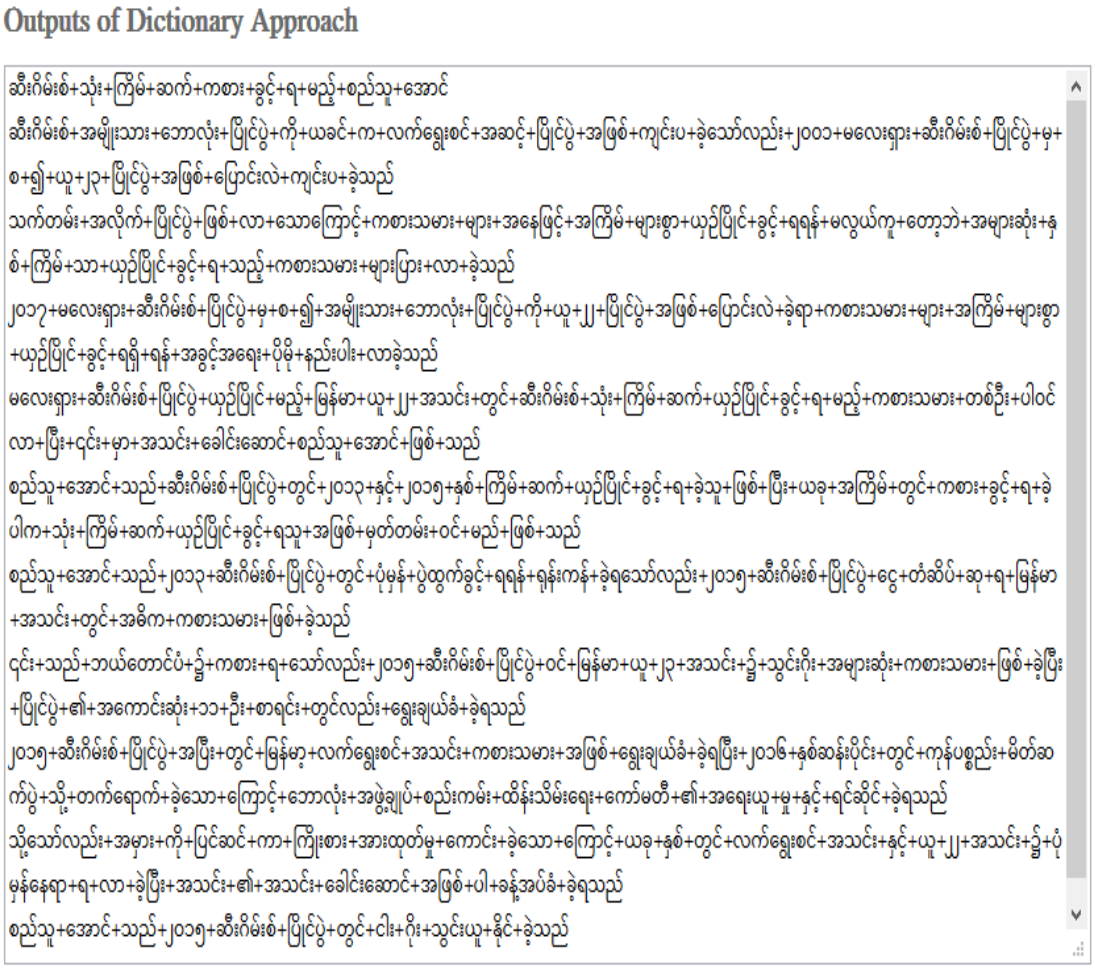


Figure 4.16 Outputs of Dictionary Approach for sports article

The input text is the output of dictionary approach, the output of hybrid approach can be seen as shown in figure 4.17.

Figure 4.17 Outputs of Hybrid Approach for sports article

4.5. Experimental Work

Some complete experimental results are show in the following sessions. The performance measures are discussed in session 4.5.1. Accuracy of word segmentation for particular articles are shown in session 4.5.2.

4.5.1 Performance measures

The performance of this system is measured by the accuracy.

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. E.g. if actual class value indicates that this sentences segmented and predicted class tells the same thing.

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. E.g. if actual class says this sentences didn't segmented and predicted class tells the same thing.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this sentences didn't segmented but predicted class tells that this sentences will segment.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this sentences segmented and predicted class tells that sentences will segment.

	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 it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best.

Accuracy = (TP+TN)/(TP+FP+FN+TN), where

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

4.5.2 Accuracy of word segmentation

The accuracy results are calculated by collection of 300 articles from the business, entertainment and sports sections of the Myanmar newspaper. The following sessions shows detail calculations for each article.

4.5.2.1 The accuracy of business article

The following calculation shows the accuracy result for evaluation of business article described in the previous session 4.4.1. True Positive of this is 60 (မြန်မာနိုင်ငံ, ယခု, ဖွံ့ဖြိုး, နိုင်ငံ, တစ်, နိုင်ငံ, ဖြစ်, ဒါ့အပြင်, နည်းပညာ, ပညာရေး, အခြား, အိမ်နီးချင်း, နိမ့်ကျ, မြန်မာနိုင်ငံ, လူနေမှု, အဆင့်အတန်း, ကွာဟမှု, ကြီးမား, နိုင်ငံ, ဒါ့ကြောင့်, မြန်မာနိုင်ငံ, အဘက်ဘက်, သူများ, နိုင်ငံ, ကျကျန်နေ, ဒါ့အပြင်, မြန်မာနိုင်ငံ, တစ်ခု, အပါအဝင်, ဖြစ်, သို့သော်, ယခုအခါ, မြန်မာနိုင်ငံ, သိသိသာသာ, ကျဆင်းလာ, လူဦးရေ, သုံး, တစ်, ကမ္ဘာဘဏ်, ထုတ်ပြန်, မြန်မာနိုင်ငံ, အခြေအနေ, လေ့လာ, စိစစ်ချက်, အစီရင်ခံစာ, နေပြည်တော်, အခမ်းအနား, ထုတ်ပြန်, ဖြစ်, လျော်ညီ, အိမ်ထောင်စု, သုံးစွဲမှု, ထည့်သွင်း, တွက်ချက်ထား, တိုင်းတာ, နည်းလမ်းသစ်, တွက်ချက်, ခုလို, ထုတ်ပြန်, ဆို), true negative is 5 (ကမ္ဘာ့အဆင်းရဲဆုံးနိုင်ငံ, ဆင်းရဲနွမ်းပါးမှု, ဆင်းရဲနွမ်းပါး, ဆင်းရဲနွမ်းပါးမှု, ဆင်းရဲနွမ်းပါးမှု), false positive is 4 (ဖွံ့ဖြိုးဆဲနိုင်ငံ တစ်နိုင်ငံဖြစ်ပါတယ်, အင်မတန်ကိုကြီးမားလွန်းတဲ့နိုင်ငံပါ, သူများနိုင်ငံတွေထက်ကျကျန်နေခဲ့ပါတယ်, တစ်ခုအပါအဝင်ပဲဖြစ်ပါတယ်) and false negative is 1 (ခေတ်ကာလအခြေအနေ အချိန်အခါ).

True Positive (TP) = 60

True Negative (TN) = 5

False Positive (FP) = 4

False Negative (FN) = 1

Accuracy = (TP+TN)/(TP+FP+TN+FN)

= (60+5)/(60+4+5+1)

= 65/70

= 0.9285714285

= 93%

The result shows good accuracy for business article.

4.5.2.2 The accuracy of entertainment article

The following calculation shows the accuracy result for evaluation of entertainment article described in 4.4.2. True Positive of this is 59 (သရုပ်ဆောင်, ဇေယျ, တချိန်တည်း, အနုပညာလောက, အတူတူ, ရောက်ရှိ, ဖြစ်, ရိုက်ကူးရေး, အတူ, တွဲရိုက်, များ, ပရိသတ်, အမြင်တမျိုး, ထင်နေ, ဖြစ်, တကယ်တမ်း, ကျတော်, တကယ်, သူငယ်ချင်း, တကယ်, မောင်နှမ, ရင်းနှီး, သူ့အိမ်, မုန့်, သွားစား, ဘာ, ရှိ, တခါတလေ, ရယ်, အန်တီ, ဝိုင်းစု, အမေ, ပြော, ပရိသတ်, အထင်မှား, ကိုယ်, ကိုယ့်ဘာသာကိုယ်, သူငယ်ချင်း, မောင်နှမ, ဖြစ်နေ, တခါတလေ, ပြောပြော, ရယ်, စိတ်ကောက်, သူငယ်ချင်း, သိတတ်, ခင်, ကောင်း, ပွင့်လင်း, တခါတခါ, ကျတော်,

စ, နောက်, ပွင့်ပွင့်လင်းလင်း, သူတပါး, ကူညီ, အင်တာဗျူး, ဖြေကြား, ဖြစ်), true negative is 3 (ဝိုင်းစုခိုင်သိန်း, အရမ်းသိတတ်တယ်, အရမ်းကူညီတတ်တယ်), false positive is 9 (အနုပညာလောကထဲကိုအတူတူ, ရောက်ရှိလာတာဖြစ်ပြီး, အတူတွဲရိုက်တာများတဲ့အတွက်ကြောင့်, အရမ်းလည်းရင်းနှီးတယ်, သူ့အိမ်ကိုမုန့်သွားစားတာတွေဘာတွေလည်းရှိတယ်, ပရိသတ်ကတော့အထင်မှားတာပေါ့, သူကစိတ်ကောက်တတ်တယ်, ခင်ဖို့ကောင်းတယ်, အင်တာဗျူးမှာဖြေကြားခဲ့တာပဲဖြစ်ပါတယ်) and false negative is 0.

True Positive (TP) = 59

True Negative (TN) = 3

False Positive (FP) = 9

False Negative (FN) = 0

Accuracy = (TP+TN)/(TP+FP+TN+FN)

= (59+3)/(59+9+3+0)

= 62/71

= 0.8732394366

= 87%

The result shows good accuracy for entertainment article.

4.5.2.3 The accuracy of sports article

The following calculation shows the accuracy result for evaluation of sports article described in previous session 4.4.3. True Positive of this is 95 (ဆီးဂိမ်းစ်, သုံးကြိမ်ဆက်, ကစားခွင့်, ရ, ဆီးဂိမ်းစ်, ယခင်, ကျင်းပ, မလေးရှား, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, စ, ပြောင်းလဲ, ကျင်းပ, သက်တမ်း, ပြိုင်ပွဲ, ဖြစ်လာ, ကစားသမားများ, အကြိမ်, ယှဉ်ပြိုင်ခွင့်, ရ, မလွယ်ကူ, အများဆုံး, နှစ်, ယှဉ်ပြိုင်ခွင့်, ရ, ကစားသမား, များပြား, မလေးရှား, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, စ, ပြောင်းလဲ, ကစားသမားများ, အကြိမ်, ယှဉ်ပြိုင်ခွင့်, ရရှိ, အခွင့်အရေး, နည်းပါးလာ, မလေးရှား, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, ယှဉ်ပြိုင်, မြန်မာ, ဆီးဂိမ်းစ်, သုံးကြိမ်ဆက်, ယှဉ်ပြိုင်ခွင့်, ရ, ကစားသမား, ပါဝင်လာ, ၎င်း, ဖြစ်, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, နှစ်ကြိမ်ဆက်, ယှဉ်ပြိုင်ခွင့်ရခဲ့သူ, ဖြစ်, ကစားခွင့်, ရ, သုံးကြိမ်ဆက်, ယှဉ်ပြိုင်ခွင့်ရသူ, မှတ်တမ်းဝင်, ဖြစ်, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, ပုံမှန်, ပွဲထွက်ခွင့်, ရ, ရုန်းကန်, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, ငွေတံဆိပ်, ဆုရ, မြန်မာအသင်း, အဓိက, ကစားသမား, ဖြစ်, ၎င်း, ဘယ်တောင်ပံ, ကစား, ဆီးဂိမ်းစ်ပြိုင်ပွဲဝင်, မြန်မာ, သွင်းဂိုး, အများဆုံး, ကစားသမား, ဖြစ်, ပြိုင်ပွဲ, အကောင်းဆုံး, စာရင်း, ရွေးချယ်ခံ, ဆီးဂိမ်းစ်ပြိုင်ပွဲ, မြန်မာ့လက်ရွေးစင်အသင်း, ကစားသမား, ရွေးချယ်ခံ, နှစ်ဆန်းပိုင်း, ကုန်ပစ္စည်း, မိတ်ဆက်ပွဲ, တက်ရောက်, ဘောလုံးအဖွဲ့ချုပ်, စည်းကမ်းထိန်းသိမ်းရေး

ကော်မတီ, အရေးယူမှု, ရင်ဆိုင်), true negative is 3 (စည်သူအောင်, လက်ရွေးစင်အဆင့်ပြိုင်ပွဲ, စည်သူအောင်), false positive is 1 (များပြား လာခဲ့သည်) and false negative is 6 (အမျိုးသားဘောလုံးပြိုင်ပွဲ, ယူ–၂၃ ပြိုင်ပွဲ, အမျိုးသားဘောလုံးပြိုင်ပွဲ, ယူ–၂၂ ပြိုင်ပွဲ, ယူ–၂၂ အသင်း, ယူ–၂၃ အသင်း).

True Positive (TP) = 95

True Negative (TN) = 3

False Positive (FP) = 1

False Negative (FN) = 6

Accuracy = (TP+TN)/(TP+FP+TN+FN)

= (95+3)/(95+1+3+6)

= 98/105

= 0.9333333333

= 93%

The result shows good accuracy for sports article.

CHAPTER 5
CONCLUSION AND FUTURE WORK

This system implemented Myanmar word segmentation using hybrid approach. It includes three parts: syllabification, maximum-matching and expectation-maximization. The first part uses rules to do the syllabification of Myanmar language. The next approach, dictionary approach, uses a Myanmar word lexicon containing words. The system uses maximum-matching approach to set the exact words from the lexicon. To solve the ambiguity program in maximum-matching approach, the system uses Expectation Maximization approach for the ambiguous words. The effectiveness of method on corpora in Myanmar has been evaluated.

The work of the proposed system is based on three domains: business, entertainment, and sports. This proposed system, tests and compares word segmentation techniques, including dictionary based approach and hierarchical expectation maximization approach. The experimental results show that the system can get 96% accuracy and is also useful as a web-based online Myanmar word segmentor.

5.1 Benefits of the System

The proposed system has the following benefits:

- This system implemented a hybrid approach that include dictionary approach and expectation-maximization approach to word segmentation of Myanmar texts.
- This system is enable to use as a pre-processing tool in Myanmar text processing such as Machine Translation, Information Retrieval and Search Engine using Myanmar language.
- This system as a web-based online system that can be used separately for every person.

5.2 Limitation and Further Extension

The proposed system is organized by a collection of 300 articles Myanmar newspaper like that Kyaymon newspaper and Myanmarahlin newspaper, for a total of nearly 35,000 words that have been manually spell-checked and segmented by associated editors. Segmentation error can occur when the words are not listed in dictionary. In dictionary, no lexicon contains every possible word of Myanmar language. There always exist out-of-vocabulary words such as new derived words, new compounds words, morphological variations of existing words and technical words. And then segmentation errors can also occur due to the limitations of the left-to-right processing.

REFERENCES

[1] Chen Chen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita, "Word Segmentation for Burmese (Myanmar)" , in proceeding of Journal ACM Transactions on Asia and Low-Resource Language Information Processing (TALLIP), Volume 15, National Institute of Information and Communications Technology, New York, USA, May, 2016.

[2] Chuong B Do & Serafim Batzoglou , "What is the expectation maximization algorithm?", in proceeding of www.nature.com/articles, Nature Biotechnology volumne 26, pages 897-899, United State, August, 2008.

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

[4] Jukka Talvitie, "Expectation Maximization" , in proceeding of Advanced Course in Digital Transmission, TLT-5906, Department of Communication Engineering, Tampere University of Technology, December, 2013.

[5] Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly and Tuong Vinh Ho, "A Hybrid Approach to Word Segmentation of Vietnamese Texts", in proceeding of 2nd International Conference on Language and Automa Theory and Applications – Lata 2008, pp 240-249, Lecture Notes in Computer Science, Spain, Oct 2008.

[6] Li Haizhou and Yuan Baosheng, "Chinese Word Segmentation", in proceeding of Language, Information and Computation, (PACLIC12), Kent Ridge Digital Labs, Singapore, February 1998.

[7] Manabu Sassano, "Deterministic Word Segmentation Using Maximum Matching with Fully Lexicalized Rules", in proceeding of Yahoo Japan Corporation, Midtown Tower, 9-7-1 Akasaka, Minato-ku, Tokyo 107-6211, Japan, April 26, 2014.

[8] Mark Schmid, "Argmax and Max Calculus", in proceeding of UBC Department of Computer Science, UBC Computer Science, Canada, January 6, 2016.

[9] Sean Borman, “The Expectation Maximization Algorithm” , in proceeding of IEEE Journals and Magazine, June 28 2006.

[10] Stefanos Zafeiriou, “Tutorial on Expectation Maximization”, in proceeding of Adv. Statistical Machine Learning (course 495), Imperial College, London, 2016.

[11] Tin Htay Hlaing and Yoshiki MIKAMI, “Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer”, in proceeding of International Journal on Advanced in ICT for Emerging Regions (ICTer), Vol 6, No 2, University of Colombo School of Computing, Sri Lanka , 2013.

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

[13] web link : https://en.wikipedia.org/wiki/Natural-language_processing

[14] web link: https://machinelearningmastery.com/natural-language-processing/

[15] web link: https://en.wikipedia.org/wiki/Languages_of_Myanmar

[16] web link: https://en.wikipedia.org/wiki/Burmese_language

[17] web link: https://towardsdatascience.com/beyond-accuracy-precision-and-recall

PUBLICATION

[1] Khine Myint Myat , Khin Mar Soe, “Myanmar Word Segmentation Using Hybrid Approach”, to be published in the Proceedings of the 9th Conference on Parallel and Soft Computing (PSC 2017), Yangon, Myanmar, 2018.