MYANMAR WORDS SORTING

Su Mon Khine, Yadana Thein

University of Computer Studies, Yangon

ABSTRACT

Myanmar word sorting is very important in indexing of search engine to optimize in the searching process of keywords. This paper proposed an efficient sorting algorithm for Myanmar words based on the weights of consonants, vowels, devowelizers, and consonant combination of each syllable of the words since Myanmar words are composed of one syllable or more than one syllable and finally the words are sorted based on Quick sort. The proposed algorithm is intended to design for Zawgyi_One font, which is mainly dominant in Myanmar Web pages.

KEYWORDS

Myanmar Script, Myanmar Word sorting.

1. INTRODUCTION

Word sorting plays a vital role in natural language processing applications for lexical analysis, information retrieval or language specific search engine in order to find user specified keywords in indexing database more quickly. In search engine indexing databases, keywords or terms should be stored in sorted order in order to find the relevance documents in the least possible time.

The sorting of Myanmar words is a difficult task in natural language processing since the nature of Myanmar script is complex rather than the English language. Myanmar sentences do not have white space to specify words boundaries and hence word segmentation is essential as a first step for word sorting.

Myanmar words cannot be sorted directly through using existing sorting algorithms since these sorting algorithm cannot be directly determined which words is greater or lesser than the other words. In this paper, Quick sort algorithm, the fastest sorting algorithm, is used to sort the input Myanmar words and the determination of which word is greater or lesser is computed by the proposed algorithm that sort as a dictionary order [5]. Although there are various fonts in Myanmar text, the proposed algorithm is implemented for Zawgyi_One fonts of Myanmar words, which is mainly dominant in web pages.

The objective of this paper is to know how to sort Myanmar words in Myanmar language and to make more efficient in searching time of indexing database of search engine that are indexed based on sorted Myanmar words in future.

This paper is organized into six sections. Literature reviews are discussed in the next section. Section 3 describes about Myanmar scripts and Section 4 describes about Myanmar words and segmentation .The proposed algorithm will be explained in Section 5. Experimental results will be discussed in Section 6 and the proposed system will be concluded in Section 7.

DOI: 10.5121/ijnlc.2015.4502

2. LITERATURE REVIEWS

The Myanmar script, also known as Burmese script and it is used to write Burmese language. The Myanmar writing style derives from a Brahmi-related script borrowed from South India in about the eighth century [4].

To sort Myanmar words, syllable segmentation and word segmentation are first important steps. Zin Maung Maung , Yoshiki Mikami [1], proposed a rule based syllable segmentation algorithm for Myanmar text and segmentation rules were created based on the syllable structure of Myanmar script. They tested a training corpus of 32,283 Myanmar syllables and they achieved 99.96% accuracy for syllable segmentation.

Hla Hla Htay, Kavi Narayana Murthy [2] described words segmentation process for Myanmar words. The training corpus contained more than 75000 sentences from various web sites and normalized various fonts to Win Innwa font before segmentation. They removed stopwords which are frequently occurred in sentences and segmented Myanmar words by using N-gram statistics of Text::Ngrams which is developed by Vlado Keselj [3]. Manual checking was performed to build lists of valid words and they have collected about 100,000 Myanmar words.

Hla Hla Htay, Kavi Narayana [6] also proposed Myanmar word segmentation using the syllable level longest Matching. Firstly, they have collected 4550 syllables from available sources of 2,728 sentences and build the words lists from available sources including dictionaries and by generating syllable n-grams as possible words [2], a total of 800000 words. Secondly, word segmentation is carried out with the longest syllable word matching using their 800,000 strong stored word list. Finally, they tested the algorithm over 5000 sentences and the algorithm correctly recognized 34633 words of 34943 words with achieving 99.11% precision and 98.81% recall respectively.

Due to the practice of Buddhism and study of Buddhist literature in Myanmar, Pali words are influenced and adopted on Myanmar language .Zin Maung Maung [7] presented an algorithm to identify Myanmar –adopted Pali words in Myanmar language. The algorithm combined rule-based syllable segmentation and a dictionary based longest matching method to identify Myanmar Pali words. Firstly, the algorithm converted to Unicode font from Zawgyi_One font and then, syllable segmentation is carried out by [1]. Pali words were identified by using a dictionary-based syllable-level longest matching method, which scans the input text by sequentially reading each syllable from the input text and matching the syllable against stored Pali words in dictionary [8]containing 3477 Myanmar adopted Pali words. The newly found Pali words are added to the dictionary in the Pali word inventory process. Incorrectly identified Pali words are flittered in Pali word disambiguation process by matching them with Myanmar words. The algorithm tested on a total of 12536 sentences and achieved precision of 97.59 % and recall to 99.04 %.

Rsesearcher, Di Jiang [10] discussed the problem of sorting orders of Tibetan dictionaries and Tibetan electronic dictionaries. Firstly, this paper stated that three reasons for the disagreement among different dictionaries: there is no criteria on sorting orders of Tibetan-transliterated letters from Sanskrit, the nature of some character forms is not clear and whether a dictionary ought to contain those forms of Tibetan sounds newly-emerged in its phonology or not. Therefore, they proposed three standardizing principles for compiling Tibetan dictionaries: agreement, compatibility and rationality to overcome these problems. For the aspects of standardization, this paper carried out a referential scheme to Tibetan dictionaries or electronic dictionaries on its sorting order, by setting up reasonable sorting values. Finally, they designed a set of digital codes for each Tibetan letter or character and assign distinctive sorting values to all existing words in the electronic dictionary with algorithm that revise errors of the [11] , according to standardized principle of compiling dictionaries.

Syllable segmentation in this paper is used by rule based syllable segmentation described in [9], in which syllables were segmented by using rule based syllable segmentation for Zawgyi_One fonts. In this paper, words are segmented by finding the longest word in the dictionary that matches the ending point of the phases, which is widely used for segmentation in Myanmar language and words are sorted according to the Myanmar to English dictionary[5] sorting order such as " " (dance / first letter of Myanmar Alphabet), " " (period), " " (car), " " (tool),....," " (remain)", " " (muscle) .

3. MYANMAR SCRIPT

Myanmar language is the official language of Myanmar, spoken as first language by two thirds of the population of 60 million and 10 million as a second language, particularly ethnic minorities in Myanmar. Myanmar script draws its source of Brahmi script which flourished in India from about 500 B.C.to over 300 AD and a system of writing constructed from consonants, vowels symbols related to the relevant consonants, consonant combination symbols, devowelizer and digits. Myanmar script is composed of 33 consonants, 12 basic vowels, 8 independent vowels, 11 consonant combination symbols and 38 devowelizer [5] and is written from left to right in horizontal line.

The combination of one or more characters but not more than eight characters will become one syllable; combination of one or more syllables becomes one word and combination of one or more than one words becomes one phase and phases are combined into sentences. Finally, a paragraph is formed by one sentence or more than one sentences.

Figure 1 shows the structure of Myanmar sentence and Figure 2 shows structure of Myanmar syllable ( ) that is equivalent to 'school' in English and that contains 7 characters

- .

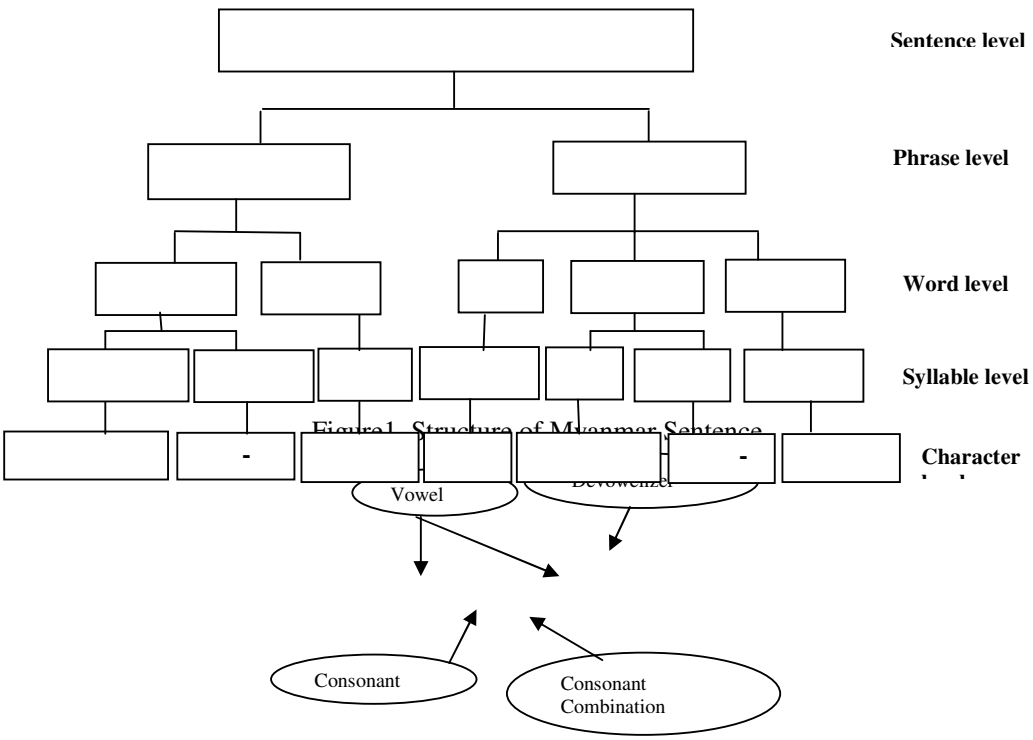

Figure1. Structure of Myanmar Sentence

Figure2. Structure of Myanmar Syllable

3.1 Consonant (C)

Myanmar word consists of one consonant or more than one consonant. There are basically 33 consonants, also known as "byee" [5] and these consonants are described in Table 1. An example of Myanmar word "□□□□□" (bear) consists of two syllables "□□□" and "□□". The first syllable consists of two consonants " " and " ", and the second syllable consists of one consonant " ".

Table1. Myanmar Consonant

3.2 Vowel (V)

Basically, there are 12 vowels in Myanmar writing [5]. They are " " ," ", " ", " ", " " , " ", " ″ ," "," ", " ", " " ," " . The variation " " ," ", " ", " ", " " , " ", " , " " ," ", " ", " " ," ", can be written . These variations on 12 basic vowels can be extended with the employment of tone markers (-) and (-) and therefore there are totally 23 vowels in Myanmar scripts shown in Table 2. The vowels " ", " ", " " , " "," ", " ", " " and the other vowels " "," "," ", "· ", can also called independent vowels .The variation of 12 basic vowels which are mostly occurred in Myanmar words rather than independent vowels is used to sort Myanmar words in this proposed algorithm.

Table2. Myanmar Vowels

-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	

3.3 Devowelizer (D)

There are 38 devowelizers in Myanmar scripts [5] which are formed by combining the character " " (asat) to some consonants and that are shown in Table 3.

Table3. Myanmar Devowelizers

3.4 Consonant Combination (CC)

There are four basic consonant combinations in Myanmar scripts [5]such as "□" , " ", " " , " " , and one or more consonant combination can be combined with consonant to become the sound of consonant. Totally, there are 11 consonant combination symbols in Myanmar scripts and that are described in Table 4.

Table 4. Myanmar Consonant combination

3.5 Myanmar Fonts, encoding system and normalization of Zawgyi_One font

The first generation of Myanmar encoding systems were ASCII code in which Latin English glyphs were replaced by the Myanmar script glyphs to render the Myanmar script which was no standardization of encoding characters. Firstly, Myanmar script was added to Unicode Consortium in 1999 as version 3.0 and improved Unicode5.1 in 2008 and Myanmar3, Padauk and Parabaik fonts are in the range of U+1000 to U+109F. And then, various fonts such as Myazedi, Zawgyi_One have been created. Although Zawgyi_One is not Unicode standard, over 90% of Web sites use Zawgyi_One font. Unicode stores text in only one order and render correctly. Zawgyi–One can store text in several ways but superficially appears correct. Therefore, the proposed sorting algorithm is developed for Zawgyi_One fonts and normalizes various writing style to one standard style. For example, the user can write the vowels sing "- " to ' ' ,' ' or ' ', ' ' after writing consonant ' ' for syllable ' ' that is equivalent to 'Ko' in English. Table 5 shows different encoding sequences of Unicode and Zawgyi_One and Table 6 shows some examples of normalization of Zawgy_One font.

Table 5. Sequence style of using Unicode and Zawgyi_One for Myanmar Syllable “ ”

Fonts	Sequence Style
Unicode	**+ + =**
	+ + =
Zawgyi-One	**+ + =**
	+ + =

Table 6. Normalization of Zawgyi_One font

Various forms of writing sequence	Normalize sequence
- , -	-
- , -	-
- , - , - , -	-
- , -	-
- , -	-
- , - , - , - '	-
.....	....
- , - , - , -	-

4. MYANMAR WORDS AND SEGMENTATION

Although Myanmar sentences are defined by a sentence bounder marker such as " ", words are not separated by special character such as white space or punctuation marks. Spaces may sometimes use in sentences, but it may not be meaningfully separated as a word. In Myanmar text, meaningful words are composed of syllables and syllables are formed by combining the consonants, vowels, devowelizer and consonant combination characters. In this paper, syllable segmentation is carried out by rule based syllable segmentation [9] for Zawgyi_One font. After the syllable segmentation is carried out, words are segmented by finding the longest word in the dictionary that matches the ending point of the phases, which is widely used for segmentation in Myanmar language. Finally, these words are sorted by the proposed algorithm.

Myanmar words can be divided into simple words, compound words, pali words (Patsint), and loan words and these words are shown in Table 7. Patsints are written by combining two syllables into one syllable by omitting the last character of devowelizer of first syllable " "and these can be repeated to form a pali or patsints word. The following words are some examples of Myanmar Pali words.

" " (paper) , " " (eye) , " " (species of land lily) , " " (copy), " " (box), " "(clever or skillful) .
The proposed algorithm can be sorted any types Myanmar words written by users.

Table7. Types of Myanmar Word

Types of Myanmar word	Examples
Simple words	(moon) (bear)
Compound words	(kettle) => (hot water) + (pot)
Pali words (Patsint)	" " (paper) " " (box)
Loan words	(computer) (radio)

5. THE PROPOSED SORTING ALGORITHM

After word segmentation is carried out, Myanmar words are sorted by the proposed algorithm according to the following step.

Firstly, weights are assigned to consonants, vowels, devowelizer and consonant combination symbols in ascending order. These weights are assigned according to Myanmar dictionary sorting order.s

a. Weights 1 to 33 are assigned to 33 consonants (C)
b. Weights 34 to 56 are assigned to 23 vowels (V)
c. Weights 57 to 94 are assigned to 38 devowelizers (D) and
d. Weights 95 to 105 are assigned to consonant combination symbols (CC).

Secondly, the input words are separated into syllable (□□□□□□□□□□□=> _) and then compute and sort according to their weight of each syllable.

The following steps are used to sort Myanmar words.

1. Firstly, separate each syllable of the words into consonant, vowel, devowelizer and consonant combination.
2. Then, the algorithm compares the weight of the consonants of first syllable of each compared words.
3. If they are the same consonants according to their weight values, then compare the weight of consonant combinations of syllable.
4. If the weights of consonant combinations are equal, then compare to devowelizer and if the devowelizers of syllables are equal, then compare to vowels of syllables.
5. Finally, if the weights of vowel of syllables are equal, then next syllables of words are compared as shown in above until end of the length of the shorted syllable is reached.

These detail steps of the algorithm are shown in Figure 3.

Algorithm MW_Sorting(W_1, W_2)
1. Assign weight to C, V, D, CC in ascending order .
2. MaxW ← Φ
3. S1 (s_1w_1,s_2w_1,….,s_nw_1)←syll-segment(W_1) // input words are segmented into syllable
   S2 (s_1w_2,s_2w_2,….,s_nw_2)←syll-segment(W_2)
4. **for** (k=0;k<min(W_1,W_2).length-1;k++) **do**
5. Result← Compare_Syll (s_kw_1, s_kw_2)
6. **if** result equal to 3 then MaxW←W1 break;
7. **if** result equal to 2 then MaxW← W2 break;
   if result equal to 1 **, then**
8. **if** k is last index ,**then**
9. **if** W1 length is equal to W2, **then** W1 is equal to W2
10. **else if** W1.lenght>W2.length , **then** MaxW←W1
11. **else** MaxW←W2
12. **end for**
13. **return** MaxW

```
Function  Compare_Syll (s1w1, s1w2)
1.  [C, V, D, CC] ← s1w1      // separate each syllable into 4 parts
    [C, V, D, CC] ← s1w2
2.  result  ← compare_Val (  s1w1 [C], s1w2 [C])
3.        if    result  equal 1 , then
4.                for (k= s1w1.lenght-1; k>0;k--) do
5.                        result = compare_Val (s1w1.[k], s1w2.[k])
6.                        if   result ≠ 1  then
7.                                go to 10.
8.                end for
9. end if
10. return result.

Function compare_Val (v1,v2)
1.  if  (v1>v2) then result ←3; // compare their weight
2.  if  (v1<v2) then result←2; //
3.  else   result=1.
5. return result
```

Figure3. Proposed Myanmar word sorting algorithm

Examples of Myanmar word sorting

Example (1) W_1= □□□□□□□□□□ (word 1) , W_2= □□□□□□□□□ (word 2)
s_1w_1= □□□□□ (first syllable of word1), s_1w_2= □□□□□ (first syllable of word2)

Table8. Comparison of syllables "□□□□□" and "□□□□□"

C	Weight of C	V	Weight of V	D	Weight of D	CC	Weight of CC
	1	□-	56	-	-	□	95
	1	□-	56	-	-		96

Table8 shows example of comparison of syllable "□□□□□" and "□□□□□". In this example, the weights of two consonants " " and " " of syllables " " ,1 and " " ,1 are equal. So, the algorithm needs to compare the weights of the consonant combinations " " and " " of two syllables. The weights of two consonant combinations are " " to 95 and " " to 96 . Therefore, word 2 is greater than word1 and the algorithm does not need to compare the next syllables of words.

Example (2) W_1= , W_2=
s_1w_1= ,s_1w_2=

Table 9.Comparison of syllables " " and " "

C	Weight of C	V	Weight of V	D	Weight of D	CC	Weight of CC
	1	-	34	-	-		95
	1	-	34	-	-		95

Table 9 shows the comparison of syllable "□□□□□□□" and "□□□□□□□" .In the second example, the weights of consonant, vowel and consonant combination of first syllable of word1 and the weights of consonant ,vowel and consonant combination of first syllable of word2 are equal. So, the algorithm compares the next syllables of word1 and word2 ,"□□" and "□□□" are shown in Table 10.The weight of vowel " " ,40 is greater than the vowel "- ",35 and therefore, word1 is greater than word2 and shown in Table 10.

Table10. Comparison of syllables " " and " "

C	Weight of C	V	Weight of V	D	Weight of D	CC	Weight of CC
	28	-	40	-	-	-	-
	28	-	35	-	-	-	-

The proposed algorithm can also sort Myanmar pali words (Patsint) such as " ", " ". The pali word " " is formed by omitting the character " " of first syllable and then directly concat to second syllable of word " " and formed as a superscription. For these pali words, the algorithm will be sorted after converting " " to " " and then compare the weights of syllables.

6. EXPERIMENTAL RESULTS

The proposed algorithm tested on different words of various sources such as newspapers, websites and, including simple words, loan words, compound words, and Pali words, as a total of 15200 words written in Zawgyi_One font. Firstly, we tested for simple words, loan words, compound words and the sorting result is shown in Figure 4. Secondly, we tested for the Myanmar Pali words and the result of sorting algorithm is shown in Figure 5. We achieved 99.73% average accuracy of both tests and we found 41 errors from these 15200 tested words. These errors are caused by sorting of words that include the independent vowels ,which is explained in section 3.2 ,such as " "," "," " , " "," "," "," " and the other vowels " "," "," ", "· ", since the proposed sorting algorithm is developed for variation of 12 basic vowels. For example, the words such as " " (desirable object of perception), " " (lead), " " (guest.) , which include " ", " "," " independent vowels and the algorithm will not be calculated the weight of these independent vowels and caused to sorting errors. The algorithm will be enhanced for these errors by mapping the independent vowels such as " ", " ", to " ", " ", vowels in order to correctly sort for the words that include these independent vowels in the future.

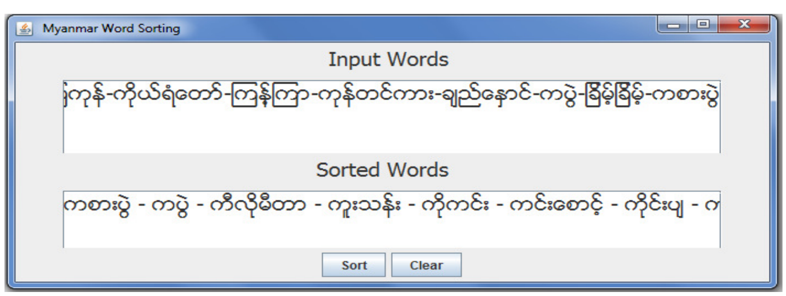

Figure 4. Result of sorting of Myanmar words

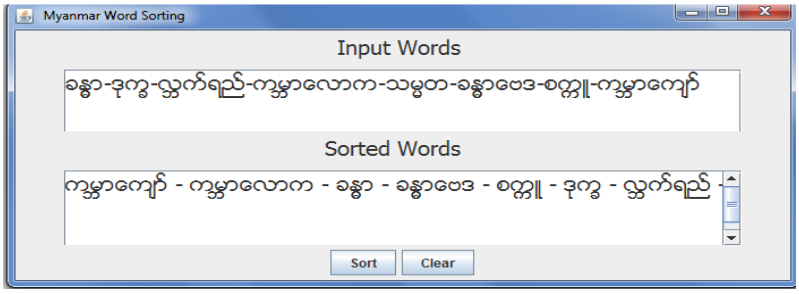

Figure 5. Result of sorting of Myanmar Pali words

7. CONCLUSION

The proposed sorting algorithm allows to sort different types of Myanmar words written by the user. Myanmar word sorting is an important step for search engine to search the sorted Myanmar keywords in indexing databases in order to speed up the searching process. This proposed algorithm is intended to develop the Zawgyi_One font for Myanmar words, since Zawgyi_One is mainly dominant on Myanmar Web documents and the sorting result of this algorithm will be used for indexing database of search engine in later in order to improve the performance of search engine. In future, the algorithm will be enhanced for independent vowels of Myanmar Language.

ACKNOWLEDGEMENTS

First of all, I would like to thank to U Myint Kyi “Takkatho Myat Soes”, member of MyanmarLanguage Commission, for his supports and encouragement to develop system. Moreover, I wish to express special thanks to Daw Myint Myint Kyi, a teacher of Myanmar, for her valuable guidance and suggestions to complete this paper.

REFERENCES

[1] Zin Maung Maung, Yoshiki Mikami , "A Rule-based Syllable Segmentation of Myanmar Text" ,Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 51–58.
[2] Hla Hla Htay, Kavi Narayana Murthy, "Myanmar Word Segmentation" Department of Computer and Information Sciences University of Hyderabad.
[3] Vlado Keselj, “Text :Ngrams” software, http://search.cpan.org/~vlado/ Text-Ngrams-1.8/
[4] "Myanmar Script", Southeast Asian Scripts.
[5] Myanmar-English Dictionary, Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar Language Commission, January 2013.
[6] Hla Hla Htay, Kavi Narayana Murthy, "Myanmar Word Segmentation using Syllable level Longest Matching", The 6th Workshop on Asian Language Resources, 2008.
[7] Zin Maung Maung "Identification of Adopted Pali Words in Myanmar Text", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012.
[8] U. H. Myint, Dictionary of Pali-derived Words, First Edition, Universities' Press, Yangon, Myanmar, 1986.
[9] Su Mon Khine and Yadana Thein, " Myanmar Web pages crawler " , Fourth International conference on Natural Language Processing , February 21~22 , 2015, Sydney ,Australia
[10] Di Jiang “The Current Status of Sorting Order of Tibetan Dictionaries and Standardization” , Academy of humanities, Shanghai Normal University, Shanghai, 200234 , Institute of Ethnology & Anthropology, CASS, Beijing, 100081
[11] Jiang, Di, Kang, Caijun: The Sorting Mathematical Model and Algorithm of Written Tibetan Language. Journal of Computer Science and Technology. (2004) Vol. 27: 4. 524-529

Authors

Su Mon Khine received M.C.Sc and B.C. Sc, in Computer Science, from University of Computer Studies, Yangon. She is now PhD candidate in Information and Technology and currently doing research at University of Computer Studies, Yangon. Her research interest includes web crawling, information retrieval, natural language processing and data mining.

Dr. Yadana Thein is now working as an Associate Professor in University of Computer Studies, Yangon (UCSY) under Ministry of Science and Technology, Myanmar. She is particularly interested in Optical Character Recognition, Speech Processing and Networking. She published over 40 papers in workshops, conferences and journals. Currently, she is working as a principal of University of Computer Studies, LoiCaw. She supervises Master thesis and PhD research candidates in the areas of Image Processing.