IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011
ISSN (Online): 1694-0814
www.IJCSI.org

Myanmar Language Search Engine

Pann Yu Mon[1], Yoshiki Mikami[2]

[1]Nagaoka University of Technology
Nagaoka, 940-3106, Niigata, Japan.

[2]Nagaoka University of Technology
Nagaoka, 940-3106, Niigata, Japan.

Abstract

With the enormous growth of the World Wide Web, search engines play a critical role in retrieving information from the borderless Web. Although many search engines are available for the major languages, but they are not much proficient for the less computerized languages including Myanmar. The main reason is that those search engines are not considering the specific features of those languages. A search engine which capable of searching the Web documents written in those languages is highly needed, especially when more and more Web sites are coming up with localized content in multiple languages. In this study, the design and the architecture of language specific search engine for Myanmar language is proposed. The main feature of the system are, (1) it can search the multiple encodings of the Myanmar Web page, (2) the system is designed to comply with the specific features of the Myanmar language. Finally the experiment has been done to prove whether it meets the design requirements.

Keywords: *Myanmar, Indexing, Web search, Non-standard encodings, Script, Tokenization.*

1. Introduction

Myanmar is a country having a population of about 56 million people comprising of 135 ethnic groups speaking their own vernacular languages or dialects. Among those, Myanmar language being spoken by more than 30 million people as their first language is the official language used in the administrative, judicial and commercial systems throughout the nation.

Myanmar language is also spoken by the rest minorities as their second language, and plays the role as the medium of easier and more comprehensive communications among the aforesaid ethnic groups. English, may be an alternative to solve all inconveniences resulting from having various ethnic groups, but is not widely understood beyond the intellectuals and those who can afford partial or whole English language education. General purpose search engines are designed mainly for English language. Non-English queries submitted to these search engines usually are not handled properly [1, 2, 3].

So using Myanmar in Web site is apparently more profitable for the Myanmar speaking people regardless of whether it is their first or second language. Besides, many Web users in Myanmar are not native English speakers, some even do not know English at all.
In that scenario, with the increase of Myanmar Web pages, the problem of efficient handling of Myanmar Web documents and user queries are becoming major issues.

1.1. Characteristics of Myanmar Scripts

The Myanmar script is an *abugida* in the Brahmic family used for writing Myanmar [4]. The characters are round in shape, because the traditional palm leaves used for writing on with a stylus would have been ripped by straight lines. It is written from left to right and no need to put white spaces between words. But the modern writing style contains spaces after each clause in order to enhance readability. Historically it was adopted from the Mon script, which further derived from Indian Brahmi script flourished in the Indian subcontinent between 5th Century BC and 3rd Century AD. Myanmar script has 33 consonants characters and 12 vowels characters.

1.2. Standardization of fonts and its encodings issues

Since 1988, Myanmar people have been trying to implement the Myanmar fonts to be used in computer system. The first Myanmar font for GUI environment (Mac) was developed in 1988 and the one for Windows system called Win-Inwa was developed in 1992. Actually it has been pretending to be English (technically Latin 1 or Windows Code Page 1252) fonts and is substituting Myanmar glyphs to English Latin glyphs. This means that they are using the code point allocated for the Latin alphabet to represent Myanmar characters.

Simultaneously, several alternative font families have been implemented by different groups of people such as MyaZedi, Zawgyi-One and MyMyanmar. Each of which has different mapping and none of these follows the Universal Coded-character Set (UCS) or Unicode standard. Those are partially followed the UCS/Unicode standard and they are not yet supported by Microsoft and other major software vendors.

Until round about 2000, there was no implementation of UCS/Unicode encoding for Myanmar language. At that time, Myanmar natural language processing task has been assigned to Myanmar Unicode & NLP Research Center. The efforts of that center are highly appreciate for a series of UCS/Unicode based fonts, 1st version called Myanmar1 and 2nd version called Myanmar2 and 3rd version called Myanmar3 have been implemented. Those fonts contain not only Unicode points and glyphs but also the Open Type Layout (OTL) logic and rules. Many Myanmar font makers have created several alternative UCS/Unicode encoding like Padauk and Parabaik.

Until now, the center keeps trying over Myanmar language processing tasks to cope well with all applications so as to complete all the tasks to cover the whole area which requires more endeavors.
Some example of Myanmar fonts and their encodings are listed in following table. These fonts use different encoding schemes and are different from one another.

Table 1: Various Myanmar Fonts and Their Encodings

Font Name	**Encoding Scheme**
Myanmar1, Myanmar2, Myanmar3, Padauk, Parabaik	UCS/Unicode: Follows Unicode 5.1.
MyaZedi, MyMyanmar, Zawgyi-One	Partially follows UCS/Unicode: Not fully follows UCS/Unicode, provided under proprietary licenses.
Popular, Win-Inwa	Graphic Encoding: Characters are decomposed into a set of graphic elements, which are mapped to ASCII code points, Provided under proprietary licenses.

2. Related Works

A few works have been done on language specific search engine to remedy the problems of specific nature of the language that in use. The following sub-sections describe some language specific search engines.

2.1. WebKhoj: Indian language IR from Multiple Character Encodings

Webkhoj is a search engine, which gives users an opportunity to search in 10 top Indian languages. It supports Hindi, Telugu, Tamil, Malayalam, Marathi, Kannada, Bengali, Punjabi, Gujarati, and Oriya [5]. In order to meet their goal of searching the relatively narrow segment of the Web, they used the language focused crawler. The language identification module they used returns the name of the language for a given web page. This module is aware of all the proprietary encodings and also uses a bag of words to recognize unknown encodings from META-tag information that might be found in an HTML page. For the ranking algorithm, they used the Term Frequency-inverse Document Frequency (TF-iDF) weight based retrieval technique.

2.2. Tumba! : The case for a Portuguese Web Search Engine

Tumba! is a Web search engine specially designed to provide better results to those searching information on the Portuguese Web. Tumba! is a public service since November 2002 (http://www.tumba.pt). It has a similar architecture and adopts many of the algorithms of global search engines, but its configuration data is much richer in its domain of specialization. It has a better knowledge of the location and organization of Portuguese Web sites. The concept of a Web crawler as a reusable software component was first proposed in Mercator: the crawler used by the Altavista search engine. It uses some of the best known algorithms for Web search and driving innovation. Tumba! provides better rankings than global search engines, as it makes use of context information based on local knowledge of the Portuguese Web and the handling of the Portuguese language.

The architecture of tumba! follows the pipelined model of high performance information systems. Information flows from Web publishing sources to end users through five successive stages such as the crawler, Web repository, indexing engine, ranking engine and presentation engine. The first four components are normal procedures of the search engines. Moreover, they have one more specific component which is presentation engine. That engine receives search results in a device-independent format and reformats them to suit multiple output alternatives, including Web browsers, mobile phones, PDAs and Web services. It can also cluster search results in multiple ways, giving users the possibility of seeing results organized by topic, instead of based only on the relevance of each matched page.

3. Motivation

Last five years, as more and more people accessing the World Wide Web in the country, the amount of Myanmar language contents has reached critical mass. A large number of journals have now online editions. There are also several magazines, information portals and personal blog sites written in Myanmar language.

The government has also introduced local languages in official web pages. While several information retrieval systems for the English language content are likely to be directly applicable to Myanmar language, but some modification should be added to meet with the specific features of Myanmar language for better performance. For some tasks, existing techniques need to be refined to suit Myanmar language. For certain features of the Myanmar language, new methodologies may have to be added.

Until now there is no specialized search tool for Myanmar language Web content for different encoding being used. Thus, information retrieval research for Myanmar language is a relatively new area.

Keeping the above things in mind, there seems to be a need of an efficient search engine that can search document in multi-encodings of the Myanmar Web pages. This research is a step towards that goal.

4. Specific Features of Myanmar Language

4.1. Various Encodings of the Myanmar Web Pages

Until 2000, most Myanmar Web pages used an image-based method to display Myanmar characters, often in GIF or JPEG. After invention the various Myanmar encodings, Web pages are created by various encodings. Among these, Zawgyi-One is widely used in Myanmar Web documents. Because Web developers in Myanmar of such content still hesitate to use UCS/Unicode as the consequence of much delayed support of operating systems and rendering the Myanmar scripts. Moreover, most of the Web developers are familiar with Zawgyi-One keyboard layout, and they do not want to learn unfamiliar ones. The other encodings are not given any names and can only be identified by their font names.

Another problem is that some Myanmar Web Pages are made by using the so-called *Mixture Encoding Style format*. It is the mixture of UCS/Unicode code points and HTML-entity, like သံ လြ င္ အိ ပ္ မက&#4153 **(**သံလွင်အိပ်မက်**)**. Some of the Web pages publishing software automatically encode Myanmar words in *Mixture Encoding Style* format. Special treatment should be done on that kind of Web pages.

4.2. Compound Words

Myanmar language is rich in compound words [6]. Two or more simple words combine together and form a compound word. For example; two simple words, "ခဲ" (lead) and "တံ" (rod, stick etc.) combine together to form a compound word "ခဲတံ" (pencil). Similarly, "ကျန်းမာ" (healthy) combines with "ပျော်ရွှင်" (happy) to form "ကျန်းမာပျော်ရွှင်" means (healthy and happy). Although compound words are widely seen in every language and not a specific feature of Myanmar language, the point here is that there is no space between components of compound words in Myanmar, while there is a space between components of words in many other languages, English for example. And this difference creates a specific problem in Myanmar language search.

In the case of English language search, it is the common practice of search engines to return those pages which contain a sub-set of the words in the user's query as well. For instance, when a user puts in words "chocolate ice-cream", the search engine returns not only the Web pages that include exact word "chocolate ice-cream" but also returns those pages with the word "chocolate" and "ice-cream" alone. It is made possible by tokenization of the input query.

In contrast to this, when searching the Myanmar words in general search engine, it behaves like a "phrase search" in English language search. That is, putting double quotes ("") round the query, telling search engines to consider the exact words in exact order without any changes. For example; document "A" contains a Myanmar compound word XYZ. And another document "B" contains every components of the word XYZ in non-consecutive manner, like "X…Y…Z". If user searches a query XYZ, search engine retrieves the document A, but does not retrieve the document B, because word segmentation is not done by the search engine. That's why special treatment is needed for Myanmar language search engine. It requires word segmentation process both in indexing stage and in input keyword processing stage.

4.3. Stacked Words (subscript form)

Myanmar writing system has been strongly influenced by the Pali and Sanskrit. In ancient times, words were written on a piece of stone in subscripted-form because of the

limitations of space availability. Later, some of subscripted-words had been changed to expanded-form. But it can be also written in subscripted form. Everybody can write freely for those words as their convenient.

Myanmar script has 33 letters to indicate the initial consonant. It can be arranged as consonant groups by 7 rows and 5 columns. Those consonant letters can be differentiated into two groups, as first 5 rows can be defined as 'Wag' (ဝဂ်) group and the last two rows as 'A-Wag' group (အဝဂ်) as shown in table 2. There are 25 consonant letters in 'Wag' group and 8 in 'A-Wag' group.

Table 2: Myanmar consonant letters

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	

'Wag' group

'A-Wag' group

For writing of subscripting consonant, there are some rules.

Rule1: The consonant from first, third and fifth column (except the consonant letter "င") of the 'Wag' group and only three consonant letters ("ယ", "လ", "သ") of the 'A-Wag' group can be subscripted by itself.

Rule2: The consonant letters from the first and third column of the 'Wag' group can be subscripted by the consonant letters from the second and forth column. In that case, the letters from first and third column should always be in the upper position.

Rule3: The consonant letters from the fifth column of the 'Wag' group can be stacked with any other consonant letters from the same row and they should always be in the upper position.

Possible combinations of Myanmar stacked consonant are listed in table 3.

Table 3: Possible combinations of Myanmar consonant stacked and examples

Possible combinations	Example
က္က, ကၠ, ဂ္ဂ, ဃ္ဃ,	တက္ကသိုလ် (University)
စ္စ, ဆ္ဆ, ဇ္ဇ, ၛ္ၛ, ဈ္ဈ, ည္ဆ, ဉ္ဇ, ဉ္ဈ	သစ္စာ (loyalty)
ဋ္ဌ, ဋ္ဌ, ဍ္ဍ, ဏ္ဍ, ဏ္ဎ, ဏ္ဏ	ဝဋ္ဌ (a circle)
တ္ထ, ထ္ထ, ဒ္ဒ, ဒ္ဓ, န္ဒ, န္ဒ, န္ဓ, န္ဓ, န္နံ	သတ္ထိ (power ability)
ပ္ပ, ပ္ဖ, ဗ္ဗ, ဗ္ဘ, မ္ပ, မ္ဖ, မ္ဗ, မ္ဘ, မ္မ,	ကပ္ပိယ (one who makes a thing allowable)
ဿ, လ္လ	ဝိဿာ ("viss" , a traditional Burmese unit of weight measurement)

For writing the subscript words, it can be divided as follow.

Omitted virama sign: In Myanmar language, some words can be written in short form by omitting virama sign. For example "တက်ကသိုလ်" can be presented in stacked form as "တက္ကသိုလ်" (University) by obeying **Rule1**. The word "ယောက်ခမ" can be abbreviated as "ယောကၡမ" (mother in law) by omitting the virama sign by **Rule2**. Another word "မန်တလေး" can be written in stacked form as "မန္တလေး" (a city in Myanmar) by **Rule3**.

Abnormal stacked consonant: Some Myanmar words can be written in different forms without omitting any character in both forms even though they are not belonging to the same consonant group. Two forms have exactly the same meaning, the same pronunciation but different representation. An example is "ထွင်း" in subscripted form. It can be expanded as "ထမင်း" (rice). Similarly, the word for "သမီး" is sometimes abbreviated as "သျှီး" (daughter). Those words are not found in native Myanmar words, except for the purpose of abbreviation.

Consonant letter becomes medial: Some words can be changed to short form by changing "consonant letter + virama sign + consonant letter" to "consonant letter + virama + medial". For example "လက်ယာ" as "လက်ျာ" (right hand side). In that case, the consonant letter "ယ" becomes medial "ျ".

Khinzi form: When the consonant letter "င" is stacked in the other four consonant letters in the same row, it becomes khinzi. Khinzi can be written in short form by changing "consonant letter "င" <nga> + virama sign" to khinzi. For example; "သိင်္ဂီ" as "သိၤဂီ"(gold). In that case, the consonant letter "င" stacked to the consonant letter "ဂ" and the former must be put in upper position by **Rule3**.

Therefore, a search engine designed for Myanmar language must consider above mentioned features of the language.

5. System Architecture

Language specific search engine should take into account the specific features of the target language. Efficient indexing and searching are one of the basic process of every information retrieval (IR) system. Moreover, this search engine takes into account specific features of Myanmar language for improving the effectiveness of retrieving the Myanmar language document. Overall processes of architecture are depicted in Figure 1. Each process will be explained more detail in the following sub-sections.

5.1. Fetching Myanmar Web Pages

In order to download Myanmar Web pages, it needs efficient crawler that can collect only Myanmar Web pages selectively from the World Wide Web. In this research, the Language Specific Crawler (LSC) [7] was used. This crawler designed to download only the targeted language. LSC runs concurrently with language identifier which indicates whether a downloaded Web page is written in the target language or not and collects Myanmar Web pages efficiently.

Here language identifier is heart of language specific crawlers. Especially, Myanmar language Web pages are using different encodings and cannot specify by checking HTML's META declaration or some others language detection tools such as Mozzilla charset detector.

In this crawler, n-gram based algorithm is chosen for language identification, because of its high accuracy in identification, its resilience to typographical errors, and its minimal data requirement for Training Corpus (TC). That language identifier is called G2LI (Global Information Infrastructure Laboratory's 2nd version Language Identifier) [8]. This method first creates n-bytes sequences of each training text. Then it checks the n-bytes sequences of the collected page with those of the training text. The language having the highest matching rate is considered as the language of the Web page.

In order to identify correctly all Myanmar fonts and encodings, it is required to train the G2LI using sufficient collections of the TC. While the existing TC contains more than 300 language translations of the Universal Declaration of Human Rights (UDHR), nine newly created UDHR texts are added to the TC for this purpose, which correspond to almost all Myanmar fonts as shown in table 1.

In language specific crawler, it is used the breadth first search (BFS) searching algorithm. BFS uses a queue data structure for first in first out (FIFO) purpose. Two queues called *seedURLqueue* and *outgoingURLqueue* are used in LSC. The former is for seed URLs (starting point) and later is for outgoing URLs.

After downloading, the downloaded Web pages are passed to next step.

5.2. HTML Parsing

After downloading the web pages, in order to get only text data, it needs to remove the HTML tags.

5.3. NLP Tasks Processing

The downloaded Myanmar pages are passed through different natural language processing tasks before they stored in the index. NLP task includes two main parts: converting of non-Unicode encodings to standard Unicode and tokenization.

Transcoding: Since Myanmar language content is being published in multiple encodings on the Web, converting of encodings to a popular standard such as Unicode is needed. If the Web pages are encoded in Unicode, then the work becomes easier. In the transcoding process, mapping of codes between two encodings takes various forms; a single code to a single code, a single code to a sequence of codes, or a sequence of codes to a single code is done. In mapping process, third party tools have been used, such as the Unify Converter Engine from Prince Kanaung group [9]. Some Myanmar Web Pages are made by using the *Mixture Encoding Style format* as mentioned earlier. Those are coded in decimal value. For those kinds of Web pages, HTML-entity should be converted to UCS/Unicode point by converting decimal value to hexa-decimal value.

Tokenization: This step is relatively straightforward in English, but non-segdmented languages such as Chinese, Myanmar pose additional challenges as words are not clearly delineated by white space. Before constructing the index files, word should be segmented first. For example, search engines require documents to be indexed by words. When a query is submitted to a search engine, key words of the query are compared against the words that are stored in index files to return search results. Tokenization is therefore an essential pre-requirement in information retrieval process. A very common approach to tokenize is

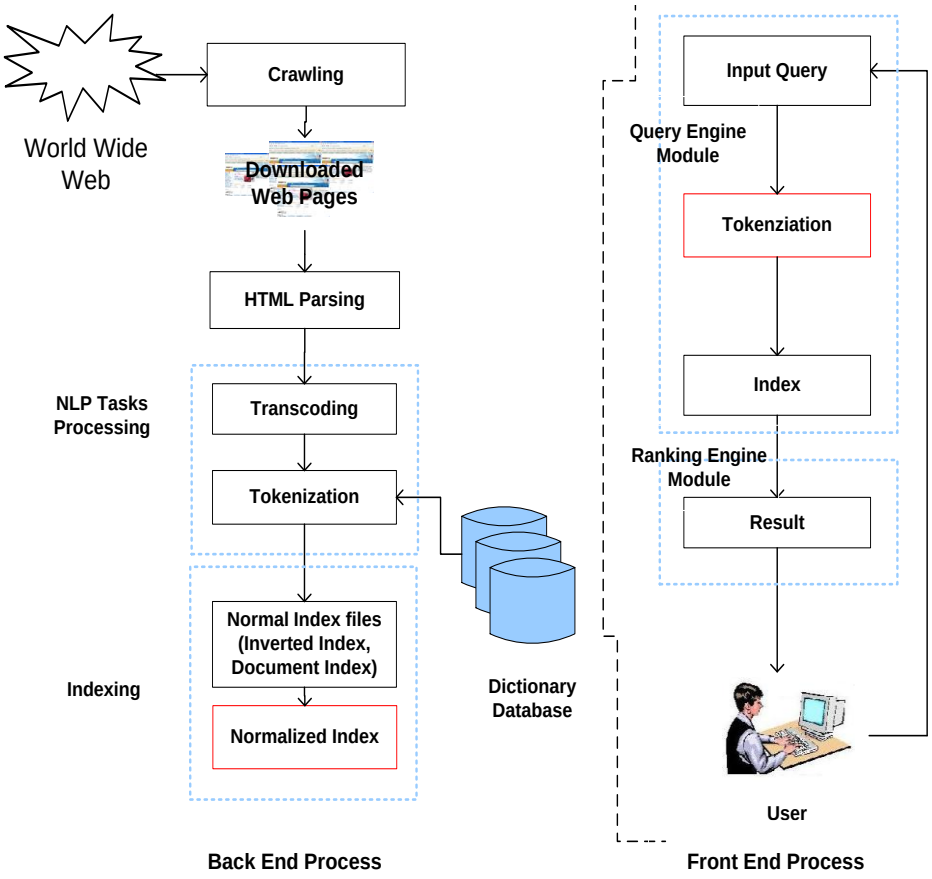

Fig. 1 Architecture of Search Engine

to use *longest matching algorithm*, frequently referred to as the *greedy algorithm* [10]. In this study, that algorithm is used to find the word on the input sentence. It normally starts at the first character in a text using a head word list and attempts to find the longest word in the list.

In this program, all of the Myanmar head words included in Myanmar–English Dictionary [11] are used as indexed file which includes about 28,500 Myanmar words. But those words alone are not sufficient to index Myanmar Web pages efficiently, because of the existence of other forms of words such as proper nouns, slang words and loan words. Additional words including 265 of proper nouns, 399 city names, 400 slang words, 350 loan words and 460 Myanmar proverbs are collected and added to the head word lists. In this process, city names are collected from "Myanmar at a glance" Multimedia CD produced by Ministry of Travel and Tourism which includes all the cities and capital name of Myanmar. And all the Myanmar proverbs are collected from the Myanmar Idioms book by Hla Thamein [12].

The proper nouns, slang words and loan words are collected from an online source called "Myanmar Orthography" [13]. In total, 30,374 words are collected.

All of those head words are stored in the database in reverse order of syllable length. If the input word is matched with one of the head word, the program will retrieve that word. If the input word does not match with the head word lists, the program cannot retrieve the word correctly. Thus the accuracy of this algorithm largely depends on the head word lists.

5.4. Index Term Extraction

Documents retrieved by the crawler are stored in a repository of Web pages. The objective of constructing the indexer is to optimize speed and performance in finding relevant documents for an input query. Without the benefit of an index file, the search engine would scan every document in the database, which would spend considerable time and computing power. The indexer also calculates scores, such as the term frequency and document frequency of each word, which can be used for search

result ranking or further processing. In this system three main index files: Inverted Index, Normalized Index and Document Index are implemented.

Inverted index: One of the cruxes of all information retrieval and database systems is the inverted index. The inverted index can determine whether a word contains in which specific document, and it also stores information regarding the frequency of each word. When query got from the user, the searcher module can retrieve the relevant document easily by referencing the inverted index file. This index holds the segmented words, frequency of those segmented words within a document and document ID.

Normalized Index: This index file handles most of the Myanmar language unique features. As mentioned earlier, in Myanmar writing system, there are many alternative forms with the same meaning but different representation. Those kinds of words are stored in normalized index file with same Normalized Word ID as shown in Table 4. Therefore, if the short form of word is given as a query, the query engine module will look for the normalized index file to retrieve the words that have same Normalized Word ID. By this way, all words that have same meaning but different representation will be retrieved properly.

Table 4: Some Example Words in Normalized Index

Word	Normalized Word ID
သမီး	1
သျှီး	1
ထမင်း	2
ထွင်း	2
ယောက်ကျား	3
ယောကျ်ား	3

Document index: document ID and name of the document are saved.

5.5. Query Engine Module

This module accepts search queries from users and the query will be tokenized. In this case, tokenization is needed to be broken down the compound words. And then the result pages go to the Ranking Engine Module.

5.6. Ranking Engine Module

A query may contain more than one word. In this case, documents containing all the words (i.e., corresponding to Boolean AND function of the words), would be first listed, and followed by document containing a subset of the query words (corresponding to OR function).
If the query contains only single word, ranking based on frequency count will be used. After all the processing, the relevant documents will be sent back to users.

6. Experimental Results

The system is tested with some selected queries and a small test data consisting of 100 Web documents to check whether it meets the design requirements.

Table 5: Experimental Result with Four Query Words

CASE	Query word	The number of pages returned by the system (recall rate)	The number of pages containing query word (s) found through the manual check process
1	baby (ကလေး)	40 (100%)	5 (encoded in UCS/Unicode) 34 (encoded in Zawgyi-One) 1 (encoded in Win-Inwa) 40 (TOTAL)
2	I (ကျွန်မ)	9 (100%)	2 (encoded in UCS/Unicode) 0 (encoded in Zawgyi-One) 7 (encoded in Win-Inwa) 9 (TOTAL)
3	daughter (သမီး)	38 (92%)	7 (contains subscripted-form, သျှီး,) 34 (contains expanded-form, သမီး,) 41 (TOTAL)
4	healthy and happy (ကျန်းမာပျော်ရွှင်)	27 (84%)	6 (contains compound form “healthy and happy”, ကျန်းမာပျော်ရွှင်) 13 (contains only “healthy” ကျန်းမာ) 13 (contains only “happy” ပျော်ရွှင်) 32 (TOTAL)

The test data are the collection of the first 100 pages collected by Language Specific Crawler (LSC) starting from three seed URLs (http://www.linhtet.com, http://gtchmawbi.ning.com and http://burmese.bnionline.net). These 100 pages are found to be encoded in three encoding schemes; namely UCS/Unicode (14 pages), Zawgyi-One (79 pages) and Win-Inwa (7 pages). As already stated in the previous section, these three encodings are the typical encodings used in Myanmar Web documents, and seem to be appropriate for testing multiple encoding capability of the system. Total size of the tested data is 652 Mbytes. Four query keywords; "baby" (ကလေး), "I" (ကျွန်မ), "daughter" (သမီး) and "healthy and happy" (ကျန်းမာပျော်ရွှင်) are chosen. The word "daughter" is chosen because it has both subscripted-form (သ္မီး) and expanded-form (သမီး). The word "healthy and happy" is chosen because it is a typical Myanmar compound word consisting of two Myanmar words "healthy" and "happy". These two words are expected to work as verifiers for testing the normalization function of the system.

Then the authors manually checked how many documents in the test data contain these query words. Manually checked counts are shown in the right hand column of Table 5.

Finally the authors run the system and counted how many pages are returned by the system for each query word. The query words are given in UCS/Unicode. The result is given in Table 5. The recall rate given in parentheses in the third column is calculated as the hit page count by the system divided by the page count as verified by manual checking process. Major findings from this experiment can be summarized as follows.

1) The results of case 1 and 2 prove that the system returns 100% of target documents successfully regardless of encodings used. If a search engine works in single-encoding basis, only a part of the target documents can be returned.

2) Case 3 (daughter) result shows that those pages with both subscripted-form word and expanded-form word are successfully retrieved.

3) Case 4 (healthy and happy) result shows that the system works properly for tokenization of the compound word and returns those pages which contains only elements of the compound word.

Though the scope of the experiment is limited to only four cases, its results are satisfactory and the system works as a search engine to meet the design objectives described in the section IV of this paper.

7. Conclusion

While many languages have been represented in the cyber space, several languages still have not. Myanmar language is one of those languages that have a little representation in the World Wide Web. In addition, there are no search tools for Myanmar language Web pages because of several reasons, which are specific to Myanmar language, such as existence of multiple encodings used in Myanmar Web documents including *Mixture Encoding Style*, existence of alternative writing practices for the same word and a special treatment needed for searching Myanmar compound words.

While proposed search engine is composed of regular search engine components, such as an indexer, index files, query word parsing engine, ranking engine, etc., authors introduced a set of new architectures in the proposed system. The new architectures include; (1) A language specific crawler which is capable of collection Web documents encoded in major encoding schemes used for Myanmar Web documents, (2) A transcoding (code-conversion) module which transforms all collected documents into normalized form (UCS/Unicode), and (3) A special indexer (Normalized Indexer) which transforms various alternative presentations of words into normalized form. The last component makes it possible to handle stacked words and alternative spellings of words in proper manner.

Author have implemented these Myanmar-specific architectures into the search engine and evaluated its searching performance using a set of test Web document data and sample queries. Evaluation results show that the proposed search engine performs its searching tasks at satisfactory level. As measured by recall rates, the search engine retrieves 100% of the target documents in multiple encoding capability test, 92% of the target documents in form-normalization test (searching capability of alternative writing forms like subscripted form), and 84% of the target documents in compound word searching capability test. Although the latter two aspects of performance are not perfect at all, but still far better than the results expected in the case of normal search engines.

Although the scope of evaluation experiment is limited, it proves that this search engine works properly to handle the specific features of Myanmar language.

8. Future Works

A search engine is a complex system and much remains to be done. First immediate goals are to improve search efficiency and to handle approximately one million Web pages. Some improvements for result relevancy include page ranking algorithm, for search efficiency include stemming, stop words removing are also be listed.

As this is a dictionary-based approach search engine, the list of the words that contained in dictionary is very important. The more words especially cities name and proper nouns will be collected in order to handle the all possible Myanmar words.

Acknowledgments

For help and advice of various sorts, I owe thanks to many people who helped me during this endeavor lasted. Without their kind assistance and consideration, this paper would not likely have matured. It is cordially noted and acknowledged the assistance of Myanmar Unicode & NLP Research center given to us is highly appreciated. Moreover, deepest thanks have been extended to Dr Daw Myint Myint Than and U Ngwe Tun for giving us their valuable data essentially in need of.

References

[1] Haidar Moukdad, "*Lost In Cyberspace:How Do Search EnginesHandle Arabic Queries?".* The 12th International World Wide Web Conference, Budapest, Hungary, May 2003.
[2] Haidar Moukdad and Hong Cui, "*How Do Search Engines Handle Chinese Queries?".* Webology, 2 (3), Article 17 (2005). Available at: http://www.webology.ir/2005/v2n3/a17.html.
[3] Judit Bar-Ilan, and Tatyana Gutman, "*How do Search Engines Handle non-English Queries?* - A case study". In Proceedings of the Alternate Papers Track of the 12th International WWW Conference, Budapest, Hungary, 2003.
[4] Peter T. Daniels and William Bright, "The World's Writing Systems". Oxford University Press, 1996.
[5] Prasad Pingali, Jagadeesh Jagarlamudi, and Vasudeva Varma, "WebKhoj: Indian language IR from multiple character encodings". In Proceedings of the 15th International Conference on World Wide Web Edinburgh, Scotland, May 2006.
[6] Okell, John, "A reference grammar of colloquial Burmese" London: Oxford University Press, 1969.
[7] Pann Yu Mon, Chew Yew Choong, Yoshiki Mikami, "Language Specific Crawler for Myanmar Pages". In Proceedings of the 11th International Conference on Humans and Computers (HC 2008), Nagaoka, Japan, November 2008.
[8] Chew Yew. Choong, Yoshiki Mikami, Marasinghe Chandrajith Ashuboda and S. T. Nandasara, "Optimizing n‐gram Order of an n‐gram Based Language Identification Algorithm for 68 Written Languages". The International Journal on Advances in ICT for Emerging Regions 02 (02): 21 – 28, 2009.
[9] Prince Ka Naung – Unify Converter Engine available at http://kanaung.googlecode.com/svn/trunk/python/fontmap.json (Accessed Date 2010, January).
[10] David D. Palmer, "A Trainable Rule-based Algorithm for Word Segmentation", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, 321 – 328, 1997.
[11] Myanmar-English dictionary produced by Department of the Myanmar Language Commission.
[12] The list of Myanmar proverbs available at http://www.mmproverb.com/ (Accessed Date 2010, January).
[13] "Myanmar Orthography" source available at http://myanmarwords.pikay.org/2008/12/ah.html (Accessed Date 2010, April).

Pann Yu Mon got her bachelor's degree, B.C.Sc, in Computer Science, from the University of Computer Studies Yangon, Myanmar, and her master's degree, ME, in Management Information System from the Nagaoka University of Technology, Japan. She is currently a Ph.D part 3 candidate in the Management and Information System Engineering Department of Nagaoka University of Technology. Her research interests include Web crawling, Web archiving and Web searching.

Yoshiki Mikami is a professor of management and information science at Nagaoka University of Technology, Japan. He has initiated several international collaborative projects, such as Language Observatory Project, the Asian Language Resource Network project, and the Country Domain Governance Project. He also serves as a chairman of Joint Advisory Committee for ISO registry of character codes. Mikami received a BE in mathematical engineering from the University of Tokyo and a Ph.D from the Graduate School of Media and Governance at Keio University.