Myanmar Web Pages Indexer with an Enhanced Dictionary

Pann Yu Mon† Yoshiki Mikami†
†Management and Information Systems Engineering Department
Nagaoka University of Technology, Japan
s065402@ics.nagaokaut.ac.jp, mikami@kjs.nagaokaut.ac.jp

Abstract

With the enormous growth of the World Wide Web, Search engines or Information Retrieval (IR) 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. A search engine which capable of searching the Web documents written in any other languages rather than in English is highly needed, especially when more and more Web sites are coming up with localized content in multiple languages. This paper reports an ongoing project of developing a Search Engine with the focus on Myanmar texts. In this case, one of the most important components of search engine is indexing of the Web documents.

In this study, the design and the architecture of a dictionary based indexer for Myanmar Web pages is proposed and which are expected to be used in search engines for Myanmar language.

Keywords

Myanmar, Indexing, Web search, Non-standard encodings

1. Introduction

Myanmar language being a member of the Tibeto-Burman language, which is a subfamily of the Sino-Tibetan family of language, is the official language of Myanmar. It is spoken by 32 million people as a first language while a second language by ethnic minorities in Myanmar. It is one of the minority languages on the Web having multi-encodings. These encodings are not given any name and can only be identified by their font names. Myanmar language Web pages are very few comparatively with English Web pages. Even within the available Web documents, most of them are not searchable and, hence not reachable due to the use of non- standardized multiple encodings. Web developers of such content hesitate to use any available standards like Unicode as the consequence of much delayed support of operation system and rendering the Myanmar scripts. Therefore, in order to search or process Myanmar language websites, the non-Unicode encoding should be able to be transcoded into standard one, and the user's queries be accepted in the same encoding format to index.

In this paper, the indexer for Myanmar Web pages with enhanced dictionary is discussed. Myanmar language is written in a syllabic system and there are no spaces always put between words or sentences. That is why word segmenting algorithm for Myanmar Language is needed not only for the indexer but also for various NLP processes. Very little research in different approaches has been published on segmenting sentences into words in Myanmar language.

Hla Hla Htay and et al., [1] used simple approach of removing the Myanmar stop words from the input sentence. It requires relatively small memory for database but the accuracy of word segmentation is only 65%. Tun Thura Thet and et al., [2] used dictionary based word segmentation approach, but their head words list does not include loan words, slang words and proper nouns which are essential for indexing in Myanmar Web pages.

In the rest of this article, more detail of the Myanmar word breaking algorithm and structure of the indexer will be described and some of the experiment that has been done will also be illustrated.

2. Characteristics of Myanmar Scripts

The Myanmar script is an abugida in the Brahmic family used for writing Myanmar. 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. And it is adopted from the Mon script, which is derived from Indian Brahmi flourished in the Indian subcontinent between 5th Century B.C and 3rd Century AD. Myanmar language has 33 consonants and 12 vowels according to traditional tones on grammar.

3. Motivation

As the Internet is becoming more and more popular and widespread, web sites launched in local language are coming up with huge volumes of information. So a Search Engine to which the keywords are given and the relevant pages being looked for are needed. Now with the multilingualism on the Web, i.e. Web content being written in different languages of the world has become important to have search engines that can search the documents written in different languages. To fulfill the goal of developing the search engine for Myanmar language, this study is essential.

4. Basic Architecture of Web Search Engines

The search engines have three major components:
(1) Crawler module,
(2) An Indexer module and
(3) Query processor module

Every engine relies on an indexer module to provide the most relevance information. The Indexer module sometimes called catalogue is to process the documents to be searched and to extract appropriate information. It extracts all the words from each page and records the URL where each word occurs. Text indexing of the Web poses special difficulties, due to its size and its rapid rate of change. In this study, it is mainly focused on dictionary based indexing of Myanmar language Web sites.

5. System Architecture

Overall processes of architecture are depicted in figure 1. Firstly it needs to collect Myanmar Web pages. Then, it goes to Natural Language Processing Tasks such as Parsing, Transcoding and Word breaking sometime called Tokenization. Finally the Indexing Files are created automatically. Each process will be explained in the next section in more detail.

5.1. First Step: Fetching Myanmar Web Pages

World Wide Web is the most convenient existing source of linguistic data providing the users abundance of texts in various types in a large number of languages.

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) [3] was used. 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. After downloading, the downloaded Web pages are passed to NLP tasks phrase.

5.2. Second Step: NLP Task Processing

Natural Language Processing includes three main parts: Removing of HTML tags from downloaded Web pages, converting of non-Unicode encodings to Standard Unicode and Word breaking. Each of which is explained in following section in more details.

HTML Parsing: After downloading the web pages, in order to get only text data, it needs to remove the HTML tags from it. In that case, some Myanmar Web Pages are made by using the *Mixture Encoding Style format*. For example သ&#4150; လ&#4156; င&#4153; အ&#4141; ပ&#4153; မက&#4153. Those are coded in decimal value. After converting it into hexa value and see in Unicode table, it can get full meaning of Myanmar words. The result of this example is သံလွင်အိပ်မက်. The Web page publishers typed simple Myanmar words but the Front page editor software automatically encoded that Myanmar words to *Mixture Encoding Style* format. For those kinds of Web pages, it needs to be solved by converting decimal value to hexa value and replaced that value with Myanmar characters according to the Unicode table.

Transcoding: Since Myanmar language content is being published in multiple encodings on the Web, transliteration of encodings to a popular standard such as Unicode [4] is needed. If the Web pages are encoded in Unicode, then the work becomes easier. Myanmar Web documents are using various types of non-Unicode encodings. But in that system, it targets to convert only one non-Unicode encoding (ZawGyiOne) to Unicode because most of the Myanmar Web documents use that encoding.

In the transcoding process, more than one-to-one mapping is necessary. The mapping of consonant conjuncts which must be done with one-to-many mapping will be done first. Then, one to one mapping of codes is done by the replacement of codes. In mapping process we used font map file from [1]kanaung converting process web site.

Word Breaking: In digital form, the languages such as Chinese, Japanese or Arabic including Myanmar represent a greater challenge as words are not clearly delineated by white space. To process text, words have to be determined 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 indexed words of the documents to return search results. Word breaking is therefore an essential pre-requirement in applications where data and information are to be processed.

A very common approach to word segmentation is to use variation of the *longest matching algorithm*, frequently referred to as the *greedy algorithm* [5]. In this study, that algorithm is used to find the word on the input data. It normally starts at the first character in a text using a heard word list and attempts to find the longest word in the list. If such a word is found, the longest-matching algorithm marks a boundary at the end of the longest word, and then it repeats from the beginning to search for following match. If no such match is found in the word lists, the character is simply segmented as a word.

In this program, all of the Myanmar head words that included in [2]Myanmar–English Dictionary 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 it may include other forms of words such as proper nouns, slang words and loan words. More words including 265 of proper nouns, 399 of cities names, 400 of slang words, 350 loan words and 460 of Myanmar proverbs are collected and added in the head words lists. In this case, cities names are collected from the Myanmar@a glance Multimedia CD produced by Ministry of Travel and Tourism which include all of the cities and capital name of Myanmar. And all of the Myanmar proverbs are collected from the Myanmar Idioms book by [3]Hla Thamein. The proper nouns, slang words and loan words are manually collected. Even those 30,374 words are not actually sufficient to index the content of Myanmar Web Pages correctly. It needs to collect more words.

All of those head words are stored in the database in reverse order of syllable length to compare with the input data. 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 is largely depends on the head word lists.

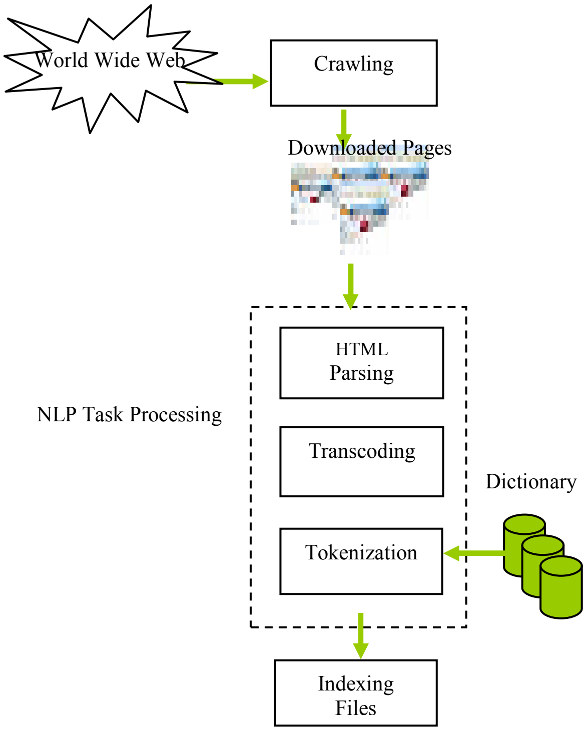


Figure 1. Step by step construction process of the system

5.3. Third Step: Indexing Files

An index is a mechanism for pointing a given term in a document. 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. In this approach two main index files: Inverted Index and Forward Index are implemented.

Inverted Index: One of the cruxes of all information retrieval and database systems is the inverted index file [6]. It provides a critical shortcut in the search process. Many search engines construct an inverted index file to quickly locate the documents containing the words in an input query and then calculate the ranking of these

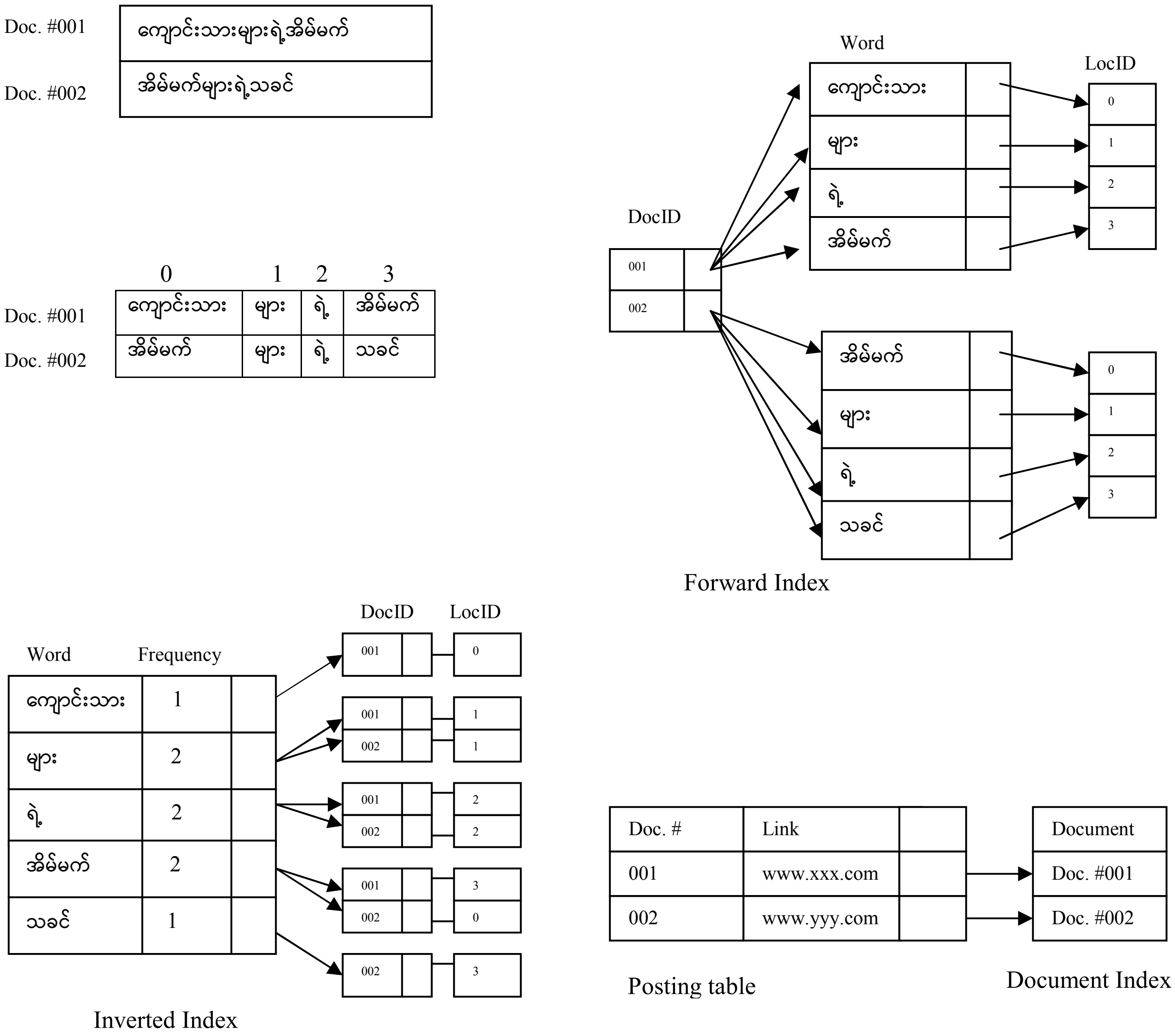
[1] *http://kanaung.googlecode.com/svn/trunk/python/fontmap.json*
[2] *Myanmar-English dictionary produced by Department of the Myanmar Language Commission*
[3] *http://www.mmproverb.com/*

documents by relevance. Figure 2. shows simplified illustration of indexing files of this system. 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.

Forward Index: Forward index stores a list of words per document. The simplified form of the forward index is shown in the following. It includes a list of pairs of a document and a word.

Document Index: In the Document table all of the Web pages are saved.

Posting Table: In the Posting table the link information for each document is saved.



	0	1	2	3
Doc. #001	ကျောင်းသား	များ	ရဲ့	အိမ်မက်
Doc. #002	အိမ်မက်	များ	ရဲ့	သခင်

Forward Index

Word	Frequency	DocID	LocID
ကျောင်းသား	1	001	0
များ	2	001 002	1 1
ရဲ့	2	001 002	2 2
အိမ်မက်	2	001 002	3 0
သခင်	1	002	3

Inverted Index

Doc. #	Link
001	www.xxx.com
002	www.yyy.com

Posting table

Document
Doc. #001
Doc. #002

Document Index

Figure 2. Indexing files for sample documents

6. Experiment Result

The experimental data presented in this section was obtained by downloading the Web pages by LSC. As the target is to handle large number of data sets, entire of this system has been implemented in Python Language which is an interpreted, interactive, object-oriented programming language.

7. Evaluation

In this section, an experimental evaluation of the approach is performed by considering the two main following parameters: Evaluation on time consuming and Evaluation on accuracy of word segmentation.

7.1. Evaluation on Processing Time

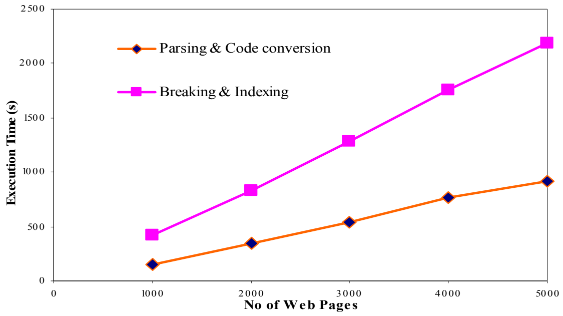


Figure 3. Processing time of each step on 5,000 Web documents

Above figure shows processing time of parsing and code conversion, word breaking and constructing the inverted index file on 5,000 Web documents. The parsing of HTML tags step and code conversation process are combined in this approach. It took 915s to complete on 5,000 documents. And it can clearly be seen that word breaking and indexing process took 2,211s.

7.2. Evaluation on accuracy of word segmentation

The performance of indexer is measured by the accuracy of word segmentation on input sentence. The accuracy of our word segmentation algorithm is discussed in this section.

In this test data of 321 Kbytes, 109,064 characters excluding punctuation marks, numerals and English words are found. In terms of words, we identified total 7,403 Myanmar words. But 246 words (3%) were not indexed. Detail results are shown in Table 1. As described in section 5.2, it needs to collect more Myanmar words. Those words can be obtained from the un-indexed word of this experiment.

Table 1. Resutl of word breaking

	N_{seg}	N_{non}	N_{total}	N_{seg}/N_{total}
Head Words	7226	165	7371	98%
Loan Words	124	58	182	68%
Proper Nouns	19	5	24	79%
Slang Words	34	18	52	65%
Total	7403	246	7629	97%

N_{seg} : the number of words correctly segmented by the program on input text
N_{non}: the number of words not segmented by the program on input text
N_{total} : the number of total words varified manually

By seeing the word breaking result, it can be considered to add more loan words and slang words in the index file. The errors resulted from the incorrect spelling in the original text, undefined headwords and incorrect description of syllable length in the database. Moreover, some error resulted from the words ending with some characters such as "့" (Myanmar Sign Dot Below) and ambiguity in word segmentation. Some examples of errors are listed in following Table.

Table 2. Some examples of errors

Types of Errors	Errors	Correct Words
1. Incorrect Spelling	ဓတုဇီဝလက်နက်	ဓာတုဇီဝလက်နက်
	သြဇာ (သ+ြ)	ဩဇာ
2. Lacking of Character ့ (Myanmar Sign Dot Below)	ခိုင်မာတဲ	ခိုင်မာတဲ့
	နေထိုင်ကြတဲ	နေထိုင်ကြတဲ့
	ပြည်သူ	ပြည်သူ့
	မည်သည်	မည်သည့်
3. Not listed in Headwords	ကျောက်ချော	ကျောက်ချောကတ္တရာ
4. Not listed in Headwords (proper noun)	ဘုရင့်နောင်	
5. Not listed in Headwords (place name)	နဝဒေး	
6. Not listed in Headwords (slang words)	အငမ်းမရ	
7. Not listed in Headwords (loan word)	တီဗွီ	
8. Ambiguity in Word Segmentation	အဝေးပြေး + စ + ခန်းမ	အဝေးပြေး+စခန်း +မ

8. Conclusion and Future Work

In this paper, the efficient word level indexing over a collection of Web pages based on longest string matching algorithm and Inverted Indexing file are addressed. Also the experimental results for each process and error analysis are presented.

The proposed algorithm performs segmentation work well and provides itself to be used as a practical word segmentation engine for various NLP applications, including Myanmar search engine. Statistical data generated by this program is useful as background information for designing various Myanmar NLP applications including input system etc. It is expected that this ongoing research will yield benefits for our Myanmar search engine development task. We also intend to experiment with indexing over larger collections of the link structure of the Web.

And this research is an ongoing stage of developing Language specific search engine for Myanmar. Several challenges are expected to encounter within near future. Thus the eminent work to be done is listed in detail as follow.

Stop Words Removal & Stemming: Like English, Myanmar Language also has stop words which do not contain important information to be used in Search queries. Usually these words should be removed from input search queries because they return vast amount of unnecessary information. In order to get better efficiency in indexing and searching process, it needs to remove stop words. Stemming also carries similar pros and cons. Reducing inflated words to their stemmed form will save space and processing.

Query Processor Module: This module is the third part of a search engine. It filters through the millions of pages in the repository to find the matches to a search queries and rank them in order to provide the most relevant information.

Acknowledgement

We acknowledge and highly appreciate the kind assistance and help given to us by Myanmar Unicode & NLP Research Center. We would like to express our thanks to Dr. Daw Myint Myint Than and U Ngwe Tun as they kindly provided us with the data we were in need of.

References

[1] Hla Hla Htay and et al., “Myanmar Word Segmentation using Syllable level Longest Matching”, Proceedings of the 6th Workshop on Asian Language Resources (ALR6), Hyderabad, India, January 2008.

[2] Tun Thura Thet and et al., “Word Segmentation of the Myanmar Language”, Journal of Information Science, Vol. 34, No.5, pp 688-704. 2008.

[3] Pann Yu Mon, Chew Yew Choong, Yoshiki Mikami, “Language Specific Crawler for Myanmar Pages”, Proceedings of the 11th International Conference on Humans and Computers (HC 2008), Nagaoka, Japan, November 2008.

[4] F.Yergeau. UTF-8, a transformation format of ISO 10646. RFC Editor, United States, 2003.

[5] “Handbook of natural language processing” by Robert Dale, Hermann Moisl, H. L. Somers

[6] “Understanding search engines Mathematical Modeling and Text Retrieval Book” by Michael W. Berry, Murray Browne.