Myanmar Traditional Lexicographic Sorting

Myat Sapal Phyu[1], Thet Thet Zin[2], Win Win Thant[3]
University of Information Technology, Yangon, Myanmar
myatsapalphyu@gmail.com[1],ttzucsy@gmail.com[2], winwinthant@gmail.com[3]

Abstract

Lexicographic sorting is the task of arranging an unordered collection of words into increasing or decreasing order.The main objective of this paper is to develop Myanmar lexicographic sorting method in Myanmar traditional way. Myanmar lexicographic sorting is very important in indexing of search engine to optimize in the searching process. Myanmar lexicographic sorting is a difficult task in natural language processing since the nature of Myanmar script is complex rather than the English language. This paper proposes an efficient method for syllable segmentation and lexicographic sorting. Syllable boundary is determined by analyzing the possible combination in a syllable from Myanmar orthographic book. Lexicographic order is determined by comparing the each segmented syllable in Myanmar traditional way. The proposed system can handle not only normal words but also abnormal words including pali loan words, English loan words, kinzi and other complex words.

1. Introduction

Myanmar Language, also known as Burmese, is the official language of Republic of Union of Myanmar. Myanmar texts or words are composed of single or multiple syllables and does not have regular inter-word spacing. Myanmar language has no official rule to insert white space to specify words boundaries and hence syllable segmentation is essential prior step for lexicographic sorting. Syllabication is necessary for future progress in Natural Language Processing. Syllabication is the process of breaking words into syllables. Many syllabication methods are developed for different areas. In this research, syllable slicing method will be developed that is appropriate for Myanmar traditional lexicographic sorting rule. The main objective of the research is to develop Myanmar lexicographic sorting method. The proposed method can overcome the difficulties of segmenting and sorting complex structure of Myanmar words that is contained in Myanmar Orthographic book.Thus, types of word are analyzed in order to determine the Myanmar lexicographic Sorting order. Myanmar Orthographic book generally contains six types of word including normal words, English loan words, pali loan words, contractions and kinzi.

In section 2, some related works are described and in section 3, we introduced the background knowledge of Myanmar language and in section 4 analysis of Myanmar orthographic book is described. In section 5, the propose method and essential preprocessing tasks, and its algorithm are described. The expected experimental result is shown in section 6 and conclusion in section 7.

2. Related Work

Collation of Myanmar (Burmese) in Unicode **[**1**]** presented an algorithm for sorting text in the Myanmar language. It focuses on the “Spelling

Book Order". It collects these collation elements in ascending order for sorting. The author described that a complete implementation should take account of contraction and short form.

In paper [8] Rule-based Myanmar Syllable Segmentation, input text strings are converted into equivalent sequence of category form and compares the converted character sequence with the syllable rule table to determine syllable boundaries. It was tested on 32,238 syllables in the Myanmar orthography (Myanmar Language Commission 2006) and the experimental results show an accuracy rate of 99.96% for segmentation of regular words.

Syllabification, Normalization and Lexicographic Ordering of Myanmar Texts using Formal Approaches [5] can handle all multisyllabic words but cannot handle a limited number of irregular words. Finite State Transducer was used for syllable segmentation.

Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer [6] proposed a new method for Myanmar syllabification which deploys formal grammar and un-weighted Finite State Transducers (FST). It was tested on 11,732 distinct words contained in Myanmar orthography corpus.

These words yielded 32,238 syllables and are compared with correctly hand syllabified words. This method performs with 99.93% accuracy on FST (SFST) tools. This system will develop to handle both *regular* and *irregular* syllable structures of Myanmar with acceptable performance for future work.

Our proposed system intends to develop lexicographic sorting method by the detail analysis of Myanmar orthographic book. The proposed method can overcome the difficulties of segmenting and sorting complex structure of Myanmar words that is contained in Myanmar orthographic book.

3. Myanmar Syllable Structure

A Myanmar syllable consists of one initial consonant, zero or more medial, zero or more vowels, optional final and tone. However, some syllables can stand independently as stand-alone syllables such as independent vowels.

Table 1 shows the types of Myanmar alphabets and order codes are specified for each alphabet. C01 to C33 are assigned for 33 consonants and if no consonant is found, C00 is assigned. M01 to M11 are assigned for 11 medials and M00 is assigned if no medial is found. V01 to V11 is assigned for 11 vowels and V00 is assigned if no vowel is found.

F01 to F32 is assigned for 32 finals and F00 is assigned if no final is found. T01 to T03 is assigned for 3 tones and T00 is assigned if no vowel is found. IV01 to IV07 is assigned for 11 vowels and IV00 is assigned if no vowel is found. However, independent vowels are transformed into equivalent vowels to be more efficient for lexicographic sorting. These codes are specified by the order that is appropriate for Myanmar traditional sorting rule.

Table 1. Types of Myanmar Alphabets

Category	Name		Glyph	Code to Order
C	Consonant		က, ခ, ဂ,...,အ	C00->C33
M	Medial		ျ, ြ, ွ,...,ြွ	M00->M11
V	Vowel		ါ, ိ, ီ,...,ိုး	V00->V11
F	Final		က်, ခ်, ဂ်,..., အံ	F00->F32
T	Tone		့,း,္း	T00->T03
I	Independent Vowel	Stand-alone Independent Vowel	ဣ,ဩ	I00->I02
		Stand-alone/Combined Independent Vowel	ဤ, ဥ, ဦ, ဧ, ဩ	I03->I07

There are many possible combinations of Myanmar characters to become a syllable. Table

2 and 3 shows the possible combinations in a syllable.

Table 2. Possible Combinations in a Syllable with Independent Vowels

Character Sequence	Example
I	ဣ
IF	ဣၚ်
IV	ဥု
IT	ဦး
IFT	ဧည့်
IVFT	ဩဘာင်း

Table 3. Possible Combinations in a Syllable

Character Sequence	Example
C	က
CV	ကာ
CVT	ကား
CF	ကက်
CVF	ကုက်
CFT	ကင်း
CVFT	ကောင်း
CM	ကျ
CMV	ကျာ
CMVT	ကျား
CMF	ကျင်
CMFT	ကျင်း
CMVF	ကြောင်
CMVFT	ကြောင်း
CVTF	မား၏
CMVTF	ကျော့ချ်
CVTMF	ဒေ့ရှ်

4. Analysis of Myanmar Orthographic Book

In Myanmar orthographic book, Myanmar words can be classified into two groups, normal and abnormal words including kinzi, consonant stacking, contraction, great tha and English loan words. Table 4 shows the types of abnormal words that are contained in the Myanmar orthographic book[2]. Abnormal kinzi and abnormal stacking are not considered in later Myanmar orthographic book[3].

Table 4. Types of Abnormal Words

Types of Abnormal Words		Words	Orthographic Syllabification	English Meaning
Kinzi	Normal Kinzi	မင်္ဂလာ	မင်/ဂ/လာ	Bless
	Abnormal Kinzi	သင်္ဘော်	သင်း/ဘော	Ship
Consonant Stacking	Normal Stacking	ဒန်္သတ္တု	ဒန်/သတ်/တု	Aluminum
	Abnormal Stacking	မန္တလေး	မန်း/တ/လေး	Mandalay
Contraction		ယောကျ်ား	ယောက်/ကျား	Guy
Great Tha		သူရဿတီ	သူ/ရဿ/သ/တီ	Angel
English Loan Word		မားစ်ဂြိုလ်	မား၏/ဂြိုလ်	Mars

5. Research Methodology

In Myanmar traditional sorting method, words are collected based on the following order: consonant, vowel, final, and medial. Myanmar lexicographic sorting process contains the following steps:

- Word segmentation
- Normal word transformation
- Syllable segmentation
- Lexicographic sorting

5.1 Word Segmentation

In order to sort Myanmar words into ascending order, Myanmar input string of words is segmented into words by looking up the

punctuation marks. Lexicographic sorting process is mainly used for words rather than full sentence. In this study, word boundaries are determined by punctuation marks or other delimiters since the input text string is the sequence of words that is delimited by punctuation marks.

Table 5. Word Segmentation

Input Text	Word Segmentation Result
မင်္ဂလာ၊ မားစ်ဗြိုလ်၊ မုတ္တမ။	မင်္ဂလာ မားစ်ဗြိုလ် မုတ္တမ

5.2 Normal Word Transformation

Each of the segmented word may be normal or abnormal word. Abnormal words are difficult to determine syllable boundaries and sorting process.

Therefore, it is necessary to transform abnormal word to normal one as the preprocessing task for syllable segmentation and lexicographic sorting. It is transformed according to the following rules [2][3]:

Table 6. Normalization Rules

No	Rule	Example
1	If the word contains "ိင်", replaces with "ငိ".	မင်္ဂလာ => မင်ဂလာ
2	If the word contains "ိန်" and other phonetic rules, replaces with "ငိး". [2]	သင်္ကြန် => သင်းကြန်
3	If the word contains "သသ", replaces with "သသ်". [2][3]	ပိသသာ => ပိသ်သာ မနုသသ => မနုသ်သ
4	If the word contains "ု" and if not "ိ", replaces with "ိ".	အတ္တ => အတ်တ
5	If the word contains "ု" and other phonetic rules and if not "ိ", replaces with "ိး". [2]	မန္တရား => မန်းတရား
6	If the word contains independent vowels, replaces with associated vowels as shown in Table 6.	သြ => အော

It can reduce the complexity of syllable segmentation and lexicographic sorting process. Rule 2 and Rule 5 are omitted in later editions, it was considered in Myanmar orthographic book, first edition. We can also omit these rules if we use later editions of Myanmar orthographic book.

Table 7. Vowel Transformation

Independent Vowel	Equivalent Vowel
ဣ	အိ
ဤ	အီ
ဥ	အု
ဦ	အူ
ဦး	အူး
ဧ	အေ
သြ	အော
ဪ	အော်

Table 8. Normal Word Transformation

Abnormal Words	⟶	Normal Words
မင်္ဂလာ	⟶	မင်ဂလာ
မုတ္တမ	⟶	မုတ်တမ

5.4 Syllable Segmentation

Syllable boundaries of each word are determined by checking the sequence of possible character in a syllable. Syllable segmentation is the essential prior step for lexicographic sorting process. After transforming abnormal words to normal words, syllable boundaries of each word are determined by the proposed algorithm.

Table 9. Syllable Segmentation

Words	1st Syllable	2nd Syllable	3rd Syllable
မင်ဂလာ	မင်	ဂ	လာ
မားစ်ဗြိုလ်	မားစ်	ဗြိုလ်	-
မုတ်တမ	မုတ်	တ	မ

Figure 1. Lexicographic Sorting Algorithm

Algorithm: Lexicographic Sorting
1. Syllable_Segmentation(Normal_Word)
2. pos1<— ϕ
3. syllable<— ϕ
4. assign order code of each character for C, M, V, F, T
5. **repeat**
6. if(Normal_Word$_{pos}$ is consonant)
7. syllable+=C
8. pos++
9. if(Normal_Word$_{pos}$ is medial)
10. syllable+=M
11. pos++
12. else
13. syllable+=M
14. if(Normal_Word$_{pos}$ is vowel)
15. syllable+=V
16. pos++
17. else
18. syllable+=V
19. if(Normal_Word$_{pos}$ is final)
20. syllable+=F
21. pos++
22. else
23. syllable+=F
24. if(Normal_Word$_{pos}$ is tone)
25. syllable+=T
26. pos++
27. else
28. syllable+=T
29. if(Normal_Word$_{pos}$ is medial)
30. syllable+=M
31. pos++
32. else
33. syllable+=M
34. if(Normal_Word$_{pos}$ is final)
35. syllable+=F
36. pos++
37. else
38. syllable+=F
39. **return** syllable
40. **until(**Normal_Word.length**)**
41. **compares** code sequence of each each syllable
42. **print** sorted result

Figure 1 shows the lexicographic sorting algorithm. In this algorithm, it assigns order code of each character for consonant, medial, vowel, final and tone that are specified in table 1. the algorithm determines the syllable boundary of each word according to Myanmar syllable structure. The maximum possible combination to become a syllable is consonant, medial, vowel, final and tone. However, English loan words can be constructed from more than this sub-syllable. Although Myanmar orthographic book does not contain much English loan words, we consider the possible combination for these words. It can be combined with extra final or medial and final(eg; ကျာ့ရှ်, ဒေ့ရှ်). Therefore, the algorithm checks consonant, medial, vowel, final, tone, medial and final. It firstly checks the 1st position of syllable whether it contains consonant or not. If the consonant is found in the 1st position, it will return the character code of this consonant. And then it will check the 2nd position whether it contains medial or not. If the medial is found, it will return the character code of this medial and checks 3rd position. Otherwise, it will return character code and checks 2nd position whether it contains vowel or not. Then, it will check the remaining possible character in a syllable. After checking these sequence, it can determine the syllable boundaries of each word. This process will be performed until the end of each word. Then, it will determine the lexicographic order by comparing the code sequence of each syllable.

5.4 Lexicographic Sorting

Myanmar words are sorted based on syllable. A Myanmar syllable encoded in Unicode can be

broken into five parts for sorting, consonant, medial, vowel, final and tone. In order to sort each word in traditional way the resulting sequence has five levels by order of priority as consonant, medial, final, vowel and tone. Final and vowel are switched from their encoded order. [1]

The purpose of this study is to sort Myanmar words in traditional way. In Myanmar traditional way, Myanmar lexicographic order is determined by (a) consonant order, (b) vowel order, (c) medial order and (d) final order. In order to sort each word, syllable boundaries of each word is firstly determined. After segmenting word into syllable, lexicographic order is determined by proposed lexicographic sorting algorithm. The algorithm compares the code sequence of each syllable. If the order cannot be determined by first syllable, it will be determined by second syllable. This process will be performed until the end of each word.

Table 10. Lexicographic Sorting

Sorted Words	Compare Code Sequence of 1st Syllable						
	C	M	F	V	T	M	F
မားစ်ြိုလ်	25	00	00	01	02	00	06
မင်ဂလာ	25	00	05	00	00	00	00
မုတ်တမ	25	00	16	04	00	00	00

6. Expected Result

Table 10 shows the expected syllable slicing accuracy compared with other methods. Although the given experiment results percentage are only slightly different, in paper [7], these rules only consider for normal words. In paper [5], it can solve almost all normal and abnormal words. Our approach can also solve both normal and abnormal words. In our approach the lexicographic order is determined by comparing each syllable while [5] determines by comparing each character. Our approach can reduce the comparison time for character level.

Table 11. Exprimental Reults

Method	Source Data	No of Syllables	Accuracy (%)
Formal Approaches[5]	Myanmar Orthographic Book	32,283	99.93%
Six Rules-based Syllable Segmentation [7]	16 documents	32,567	100%
Rule-based Syllable Segmentation[8]	Myanmar Orthographic Book	32,283	99.96%
Our approach	Myanmar Orthographic Book	32,283	100%

7. Conclusion

Myanmar lexicographic sorting process is supported by Natural Language Processing tasks, syllable segmentation and sorting. In this study, syllable segmentation is the pre-processing task for lexicographic sorting. The proposed system is effort into computerization of Myanmar script. Lexicographic sorting process is performed by the combination of syllable segmentation and Myanmar lexicographic sorting algorithm.

References

[1] K.Stribley, "*Collation of Myanmar in Unicode*", technical report June 17,2007.

[2] Myanmar Language Commission."*Myanmar Orthography*", first Edition, University Press Yangon, Myanmar, 1978.

[3] Myanmar Language Commission."*Myanmar Orthography*", Second Edition, University Press Yangon, Myanmar, 2006.

[4] Tin Htay Hlaing and Yoshiki Mikami. (2011). "*Collation Weight Design for Myanmar Unicode Texts*". In the proceedings of International Conference on Human Language and Technology, Alexandria, Egypt, pp. 1-6.

[5] Tin Htay Hlaing "*Syllabification, Normalization and Lexicographic Ordering of Myanmar Texts using Formal Approaches*", August, 2014.

[6] Tin Htay Hlaing and Yoshiki Mikami. (2013). "*Automatic Syllabification of Myanmar Texts using Finite State Transducer*", International Journal on

Advances in ICT for Emerging Regions, Volume 6, Number 2.

[7] T.T.Thet,J.C.Na, W.K.Ko "*Word Segmentation of Myanmar Language*". Journal of information science JIS.2nd October 2007.

[8] Zin Maung Maung, Mikami Yoshiki. (2008): "*Rule-based Syllable Segmentation of Myanmar Texts*". In Proceedings of the 6 thWorkshop on Asian Language Resources,January 11-12, Hyderabad, India.

[9] 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.

[10] http://www.myanmarlanguage.org/unicode/myanmar-fonts-which-follow-unicode-rules

[11] http://unicode.org/charts/PDF/U1000.pdf