Myanmar Handwritten Recognition and Erratum Detection by using MICR

San Su Su Yee, Dr.Yadana Thein
University of Computer Studies, Yangon, Myanmar
suyebaby@gmail.com, yadanaucsy@gmail.com

Abstract

This paper contribute recognition and detection erratum for Myanmar handwritten compound words. MICR (Myanmar Intelligent Character Recognition) is used for the character recognition. Because this method is interesting algorithm to recognize Myanmar characters that has been developed recently in Myanmar. It contains statistical, semantic and the final decision is made by the voting system. Erratum Detection, it is a new contribution of detecting compound word in string texts. It dependent on the language set of string substitutions reflects the surface form of errors that result from cognitive, typographical mistakes, or mistyping. A robust erratum detection technique is needed to cover above all situation. The system index pair possible extended/medial code and then provides a pair code for detect complete compound word. In detection, it has three situations: twice the same extended, extended/medials pair not matching error and not compound word in real. The final output, Myanmar compound words will be produced editable text with highlight in each error word.

1. Introduction

There are different techniques that can be used to recognize characters. Among them, *O*ptical *C*haracter *R*ecognition and *I*ntelligent *C*haracter *R*ecognition are two basic techniques for character recognition. OCR typically involves the process of translating digitized images of text into a machine-readable format (such as ASCII or Unicode). But Myanmar characters and digits are round shapes in nature and have similar forms so that OCR occur error such as misrecognition, inconvenient. ICR can successfully overcome these problems.

Myanmar Intelligent Character Recognition (*MICR*) is a technique based on ICR. High speed recognition rates can be gained by using MICR. It can recognize both type-face and handwritten characters. It is used to recognize effectively hand-printed characters.

Writing is very important because it represent the language. In the world, many countries have their own language and native language writing system. The concepts of writing errors are a fuzzy one. The errors others make in Myanmar writing differ according to the characteristics of other language. In erratum detection system, needs to detect each complete compound word.

Familiar erratum detection approaches are often based on language knowledge, and mainly include rule-based method. Rule-based methods use rule sets, which describe some exact dictionary knowledge such as word or character frequency, etc.

Myanmar characters are complexity and widely can be seen in this observation that two of the most common reasons for miswriting are (1) the difference between writing representation and phonetic utterances and (2) phonetic similarity of Myanmar characters. So, we using more knowledge from language itself are required to develop Natural Language Processing.

The remainder of the paper is structured as follows: section 2 gives the background history of our mother language and characteristics of compound word. In section 3, we show up the implementation of the system and we give explanation for MICR method. In section 4, expresses about erratum detection system. In section 5, show the output. Experimental results and conclusion are in section 6 and 7, respectively.

2. History of Myanmar Language

The Myanmar language belongs to the Sino-Tibetan family of languages of which the Tibetan-Myanmar (Tibeto-Burman) subfamily forms a part. It has been classified by linguists as a monosyllabic or isolating language with agglutinative features. It is a tonal and analytic language. There are different types of language in Myanmar such as Myanmar, Karen, Rakhine, Chin, Mon, Shan, etc. But, Myanmar language is the mother language in Myanmar.

The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which derives from a Brahmi-related script borrowed from South India in about the eight century for the Mon language. The first inscription in Burmese dates from the following years and is written in an alphabet almost identical with Mon inscriptions. The earliest Myanmar and Mon language can be seen in MyaZeDi Stone inscription.

2.1. Myanmar Language Characteristics

Myanmar alphabet consists of (33) consonants, (12) vowels, (4) medials, (10) basic letters and (10) digits. In Pali alphabet consists of (41) letters, as shown in Figure 1.

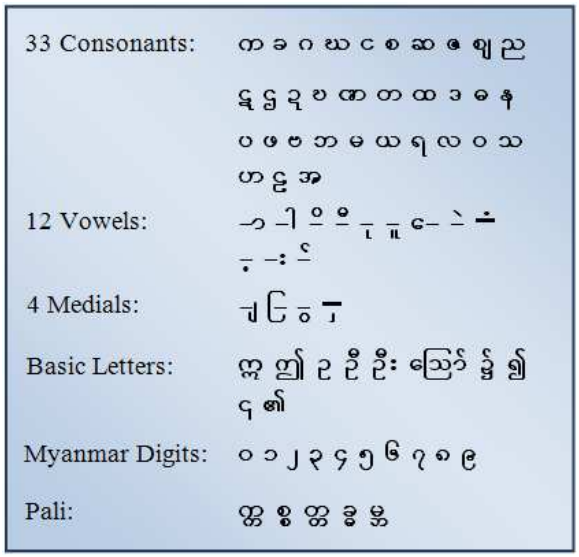
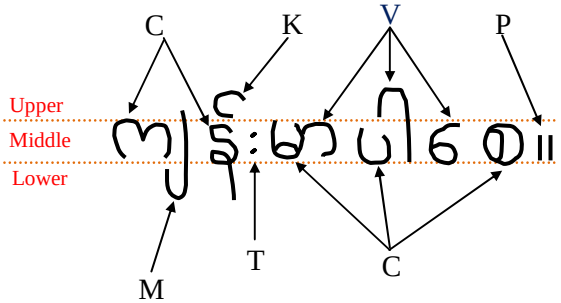


Figure1. Patterns of Myanmar alphabet

Some Myanmar characters can stand only one (U C) or combined with other extended characters to become meaningful words (U? aWm). Myanmar script is written from left to right, as shown in Figure 2. The rounder forms were without tearing the writing surface of the leaf. There are no spaces between words or between syllables, although informed writing developed to permit writing on palm leaves often contains spaces after each clause.

Myanmar syllable can be composed of multiple characters. Each consisting of two or more stems joined together is known as a compound word. We distinguish mainly (7) kinds of compound words in Myanmar writing system. Table 1: shows the structure of compound words.



C - Consonant **M - Medials**

V - Vowels **P - Punctuation**

K - Killer **T - Tones**

Figure 2. Overview of Myanmar language

Table1. Structure of Compound Word

Symbols Compound Word		
NO.	**Compound Words**	**No. of Characters**
1	ကေ	2
2	ညို့	3
3	ဆော်	4
4	ငြိမ်	5
5	ဗျိုက်	6
6	နှောင့်	7
7	မြှောင့်	8

3. System Framework

The basic architecture of the proposed system in this paper is shown in Figure 3.In this system includes five stages: Data acquisition, Pre-processing, MICR method, Erratum detection and Output.

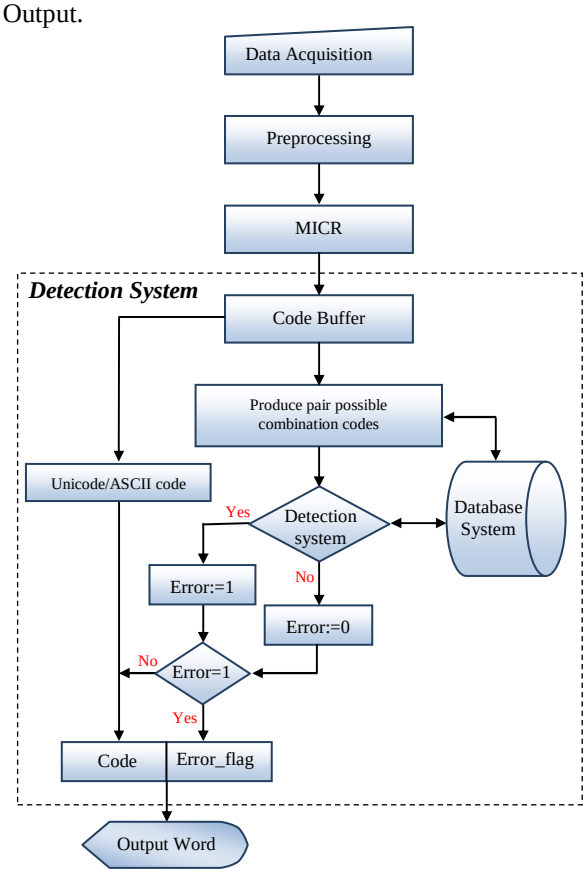


Figure 3. Proposed system design

Two different types of data input method: *online* and *offline*. In the stage of *Data Acquisition*, the proposed system can handle on only online data input by users. Isolated characters are needed to process the image.

Various *preprocessing* operation are: Gray Scale Converting, Noise Filtering, Binarization and Extraction. Firstly, convert the incoming original image into gray level image and then filtering the noise of the image result from gray scale conversion of image. If conversion of a gray-scale image into a binary image, we extract row and column for each character recognition. And then, labeling scheme is used in this system for the one character lonely.

3.1. MICR (Myanmar Intelligent Character Recognition)

MICR system is based on **ICR** (Intelligent Character Recognition). MICR was trained in both typeface and handwritten characters, so it can recognize both online and offline characters. But it is more convenient for noise free images and isolated characters to improve accuracy rate. This system used statistical and semantic approach to collect information. That information includes the data of width and height ratio, horizontal and vertical black stroke count, number of loops, end point, open direction, histogram values and character type, etc.

After collecting this required information for each character, we put them on the properties array to record them. Properties of each character are compared with Pre-Defined Database: *Basic characters* (B-database), *Extended characters* (E-database), *Medials* (M-database). When the incoming character matches with the database, the voting system is used to make the final decision of the image on that information, as shown in Figure 4. Then, the output code numbers are stored in the code buffer.

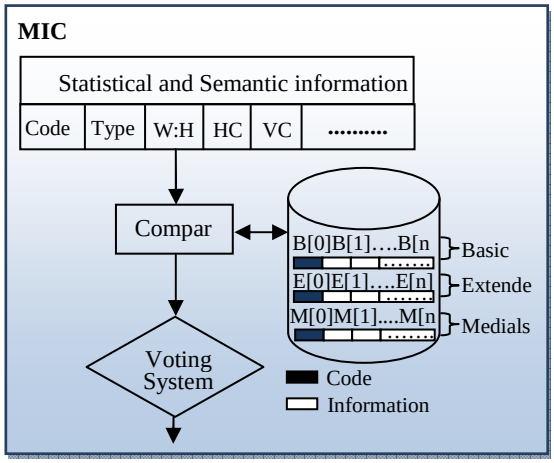


Figure 4. Architecture of the MICR

3.2. Successful Applications of MICR

MICR has been successfully applied in a lot of application such as:

Using MICR

- Speed limited road signs recognition
- Car license plate reader
- Recognition of Myanmar basic characters and compound words(vowels)
- On-line Handwritten Myanmar Pali Character recognition
- Online Myanmar Medial Hand-Printed Characters Into Machine Editable Text
- Handwritten English Characters to Machine editable text by applying MICR
- Converting Myanmar Portable Document format to machine editable text with format

Using both MICR and MVM

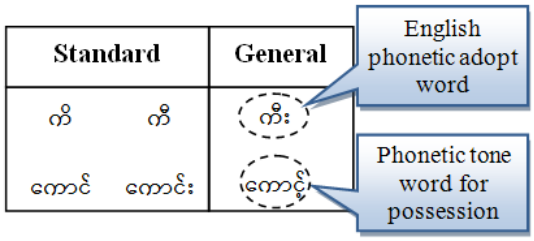
- Voice production of Handwritten Myanmar Compound Words
- Enhancing the Myanmar Pali Recognition based on MVM

4. Erratum Detection System

Firstly, 'why erratum (printing or typing error) is becoming' is presented. Erratum can be of many types, such as typographical error, cognitive error, etc. Myanmar language writing breaks down possible human typing errors into two classes, typographical error and cognitive errors. Typographical errors (e.g., misspelling 'ee' instead of 'ee;') generally occur due to people's mistakes while typing. Cognitive errors (e.g., misspelling 'Ai' instead of 'bi') are caused by writers who do not know how to spell the word.

In Myanmar syllable structure, syllables or compound words are formed by consonants combining with vowels or medials. However, some syllables can be formed by just consonants, without any vowel (e.g., rr 00). Myanmar writing system can be distinguished into two parts: standard writing system and general writing system, as shown in Figure 5. *Standard writing system*, it has 1864 words and it is exist in Myanmar writing language as real compound words. In *General writing system*, there are lots of words which are described in published Myanmar dictionaries. And then, it has much type of words: adaption words, phonetic tone words, dialect words, etc. In this paper, the system can detect all form.

Figure 5. Example words of Myanmar



writing system

After the MICR method, the next step is to detect erratum the incoming compound words. This system consists of three main parts: produce pair codes, detection system, and database system.

4.1. Produce possible pair codes

The architecture of the produce pair codes design as shown in Figure 6. In this part, the code numbers of compound words are arranged because it needs to index the consonant code number to produce pair code for extended/medials codes. Then, extract the extended/medials code. According the sequential code number result that got this stage, possible pair code will be produced. By producing pair code number, it uses possible combination of extended/medials database. There are (6) rules that need to follow by producing each compound word code for compound words. Figure 7 is illustrated by using (4) rules.

4.1.1. First rule. In this rule, it indexes the first pair extended/medials characters codes or consonant/ extended character code (e.g., ၎) to produce pair code. These codes are compared the database. If these codes are really combined in the writing system, the pair code is produced for error detection. When the two codes aren't combined, the system returned the incoming two codes. And then, these codes are stored in the final code buffer.

4.1.2. Executing another rule. It indexes the previous rule produce pair code and next extended/medial code and combines them to perform new pair code.

4.2. Detection system

In this system, it detects three situations: twice the same extended error, extended/medials pair not matching error and not compound word in real.

4.2.1. Twice the same extended error. This error is performed when the writer producing text containing the same extended character or medials character by twice. But, this system allows the consonant twice the same, as shown in Figure 8.

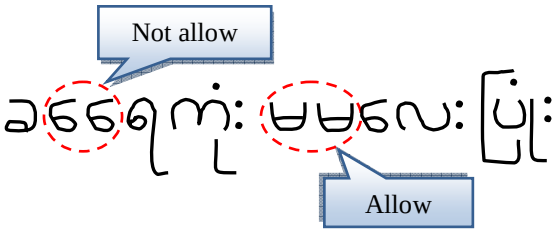


Figure 8. Detect error for twice the same extended error

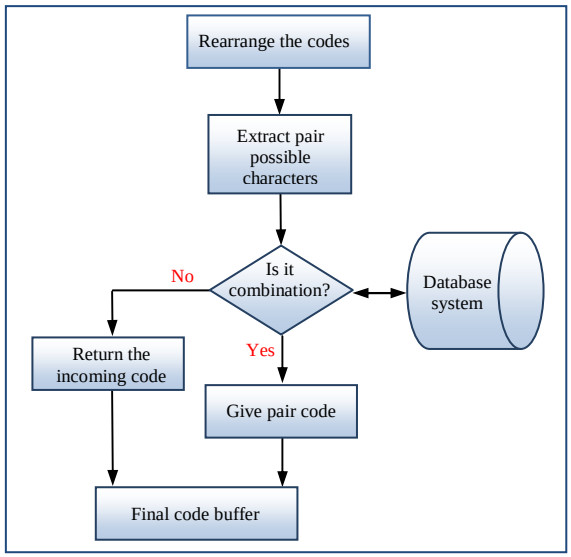


Figure 6. Producing pair code design

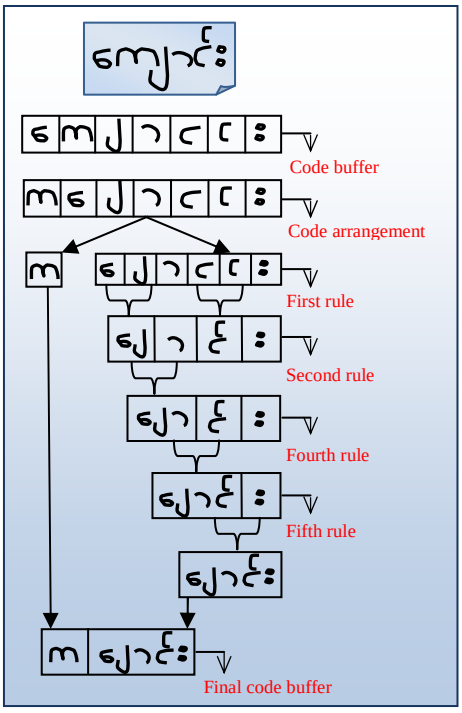


Figure 7. Step by step execution by using 4-Rules for 7 characters compound word

4.2.2. Extended/Medials pair not matching error. When the extended or medials character can't combine each other, this error is happen, as shown in Figure 9.

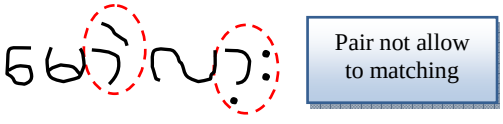


Figure 9. Detect error for pair not matching error

4.2.3. Not real word error. Humans often make errors during communication, in either spoken or written language. In this error includes typographical errors and cognitive errors. The typographical errors involve regular forms of mistyping rather than cognitive errors. Some compound words not exist in language writing system, as shown in Figure 10.

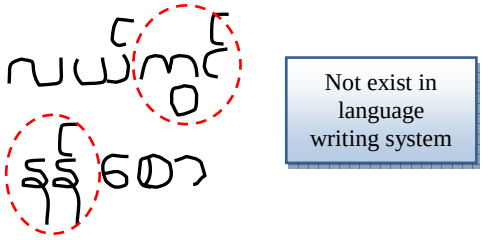


Figure 10. Detect error for not real compound word error

4.3. Database system

In this system, it collects all of the databases for produce pair code and detection system. Produce pair code system, it is used 6 databases for 6 rules. In Figure 11, it shows by using two possible pair code database to produce pair code. In detection system, it uses 7 databases for each writing system (standard writing and general writing) to compare compound word.

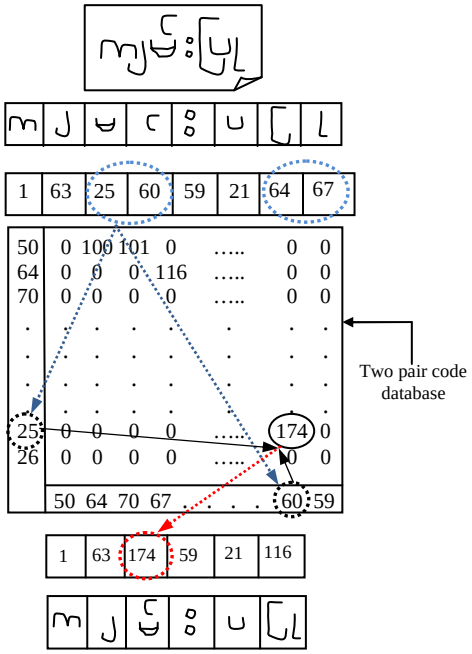


Figure 11. Produce pair code by using two possible pair code database

5. Output

After that, the recognized combined words are produced as output. This output can be shown in the Microsoft Word file as the editable text format and incorrect compound words with color.

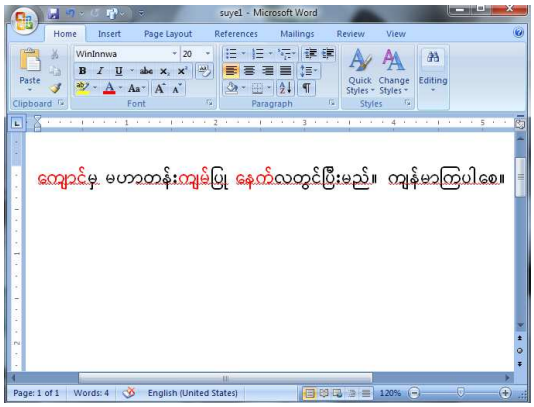


Figure 12. Output data with color

6. Experimental Result

In this paper, MICR method is used for the handwritten character recognition. But MICR can recognize not only handwritten but also type-face. By using MICR method, we can achieve the high accuracy rates and the high recognition rate. The performance of the detection is totally depending on the MICR.

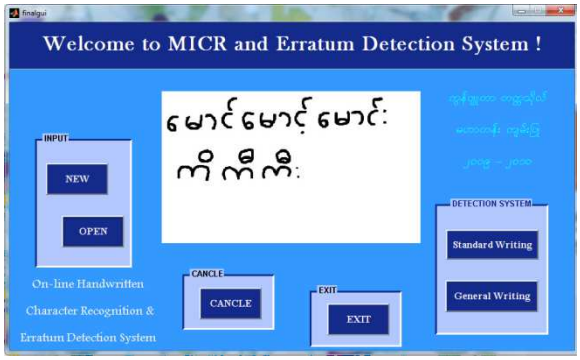
Table 2. Results on the types of compound words

Types of compound words	No. of words	No. of characters	Recognition accuracy rate
Two characters compound word	10 words	20 characters	98.2%
Three characters compound word	10 words	30 characters	97.8%
Four characters compound word	10 words	40 characters	96%
Five characters compound word	10 words	50 characters	95%
Six characters compound word	10 words	60 characters	93%
Seven characters compound word	10 words	70 characters	90%
Eight characters compound word	10 words	80 characters	88.9%

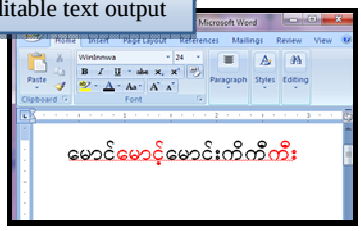
Table 3. Recognition and detection result for handwritten character

Handwritten		
Sample	Recognition accuracy rate	Erratum detection accuracy rate
10 words	98%	97%
30 words	96.10%	95%
50 words	94%	92.60%
70 words	92.4%	90%
over 70 words	90%	88.60%

Table 2 shows MICR recognition accuracy rate on the types of compound words. Table 3 also shows the erratum detection accuracy rate for handwritten characters. In this system, it can detect two types of writing system (Standard writing and General writing), as shown in Figure 13.



Standard writing editable text output



General writing editable text output

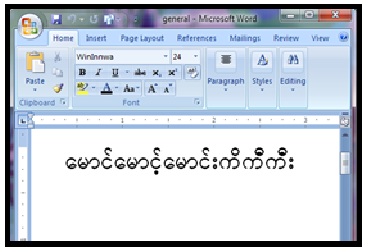


Figure 13. Erratum detection result for standard writing and general writing

7. Conclusion

This paper represented the erratum detection of Myanmar handwritten compound words applying MICR and Detection system for all compound words. The character recognition MICR method was successfully developed for Myanmar characters and was found to perform reasonable well with sufficient accuracy. In this paper, the system can detect irregular form of all compound words in Myanmar writing system.

8. References

[1] Tun Thura Thet; Jin-Cheon Na, Wanna Ko Ko, "Word Segmentation of Myanmar Language", Journal of Information Science JIS, 2nd October,2007

[2] Zaw HTUT (Mr.), "Features of Myanmar Language Document Styles", Executive Committee Member, MCSA Myanmar Computer Federation (MCF)

[3] E.E.Phyu, Z.C.Aye, E.P.Khaing, Y.Thein and M.M.Sein, **"**Recognition of Myanmar Handwritten Compound Words based on MICR", the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008

[4] Yin Mon Aung, Ei Kay Khine, Khaing Wai Myo, Dr. Yadana Thein, "Writer Independent Online Myanmar Medial Hand-Printed into Machine Editable Text", 30th Asian Conference on Remote Sensing (ACRS), China,

[5] Ei Theingi, Ei Kay Khine, Thu Wai Kyaw Kyaw, Dr. Yadana Thein, "Enhancing the handwritten Myanmar characters recognition system for pali", 30th Asian Conference on Remote Sensing(ACRS),China, 2009

[6] Naushad UzZaman, Mumit Khan, "A Comprehensive Bangle Spelling Checker" , Center for Research on Bangla BRAC University, Bangladesh

[7] Gerhard B. van Huyssteen, Menno M. van Zaanen, "Learning Compound Boundaries for Afrikaans Spelling Checking", North-West University (South Africa) & University of Tilburg (The Netherlands), Centre for Text Technology, North-West University, Potchefstroom, 2531, South Africa