Font Script Identification Based on N-gram Text Categorization

Kyaw Myo Than, Hla Hla Htay
University of Computer Studies, Yangon
kyawmyothan87@gmail.com, hlahlahtay123@gmail.com

Abstract

In this paper, we propose a method for identifying font scripts of Myanmar Language. Because of the unavailability of nationwide standardized encoding scheme in Myanmar font scripts, knowledge written in Myanmar language are scattered across internet pages. Font scripts Identifier are essential to merge those scattered knowledge into one for NLP application such as text categorization, information retrieval and text summarization. Our proposed method use N-gram based text categorization. A piece of text for 11 font scripts is taken for training. TF-IDF (Term Frequency-Inverse Document Frequency) weights of character N-grams for each font script are computed and stored as a profile for that particular font script. When a new text document is given to testify, TF-IDF weight is computed for that font script and cosine similarity is measured between the test and trained profiles. The highest similarity scored of the font script is taken as a result. 100% accuracy is obtained for testing of 11different font scripts by applying TF-IDF approach. Therefore, this method works well for Myanmar font script identification.

Keywords: Font, Font Script, Language Identification, Font Script Identification, N-gram, Text Categorization, TF-IDF Weights

1. Introduction

Information Communication and Technology (ICT) is developing incrementally and large amounts of information in Myanmar Languages are becoming available on the web. Companies, industries, supermarkets, government offices, banks, hotels and hospitals launch their own websites. Myanmar computer users use several kinds of Myanmar encoding to type documents and create websites. However, it is very difficult to computerize because of standard encodings. Thus it has a major bottleneck for many processes such as word segmentation, sorting, line breaking and so on.

The unavailability of Unicode Myanmar keyboard and the inconsistent and nonstandardized Myanmar fonts lead web developers to present information in English or in pdf, gif and jpg format if they try to present ethnic languages in Myanmar**.** We need font script identifier that can identify multiple Myanmar font scripts. Therefore, we can fetch information correctly and then convert it to standard font script.

Font script identification is considered as text categorization. Text categorization, also known as text classification, concerns the problem of automatically assigning given text passages into predefined categories. The task of text categorization is to automatically classify documents into predefined classes based on their content. A number of statistical and machine learning techniques have been developed for text classification, including regression model, K-nearest neighbor, decision tree, support vector machines, Markov model, comparison of Trigram Frequency vectors, N-gram based text categorization and so on. The proposed system will identify the script of Myanmar fonts by applying N-gram based text categorization and TF-IDF (Term Frequency-Inverse Document Frequency) Weights.

2. Nature of Myanmar Language Scripts

Myanmar language is the official language of the Union of Myanmar. It is spoken by 32 million as a first language and as a second language by ethnic minorities in Myanmar. Myanmar language is a tonal and analytic language using Myanmar script. Myanmar characters are rounded in shape and the script is written from left to right. No space is used between words but spaces are usually used to separate phrases. The Myanmar language still remains as one of the less privileged Asian languages in cyber space. Many people have put considerable effort into the computerization of the Myanmar script.

Because of the unavailability of nationwide standardized encoding scheme in Myanmar language/script and extension of Myanmar scripts (ethnic scripts), Myanmar people try to develop local version of True Type Myanmar fonts which can

easily run on Standard English version software. The first publicly used Myanmar font for Window platform, named *Shwe and Mya*, was developed around 1992. Some fonts were developed for Mac platform. But Mac is not a popular platform in Myanmar and these fonts are not widely used in Myanmar. The fonts are developed by using some graphic software and assign each character to the respective Latin key in ASCII code. These true type fonts can show the web pages correctly if they are downloaded and installed into the computer. But normally, the web sites use different fonts and just downloading one font is not enough for seeing all web sites written by ethnic languages in Myanmar. Normally, the font developers develop only the font of one script instead of developing font to use all ethnic languages (most ethnic languages use extended Myanmar scripts). In other way, the ethnic languages also have to develop their own fonts. The font that includes some ethnic scripts is *WinMyanmar* Font. Most Myanmar True Type Fonts use the Myanmar Typewriter keyboard style which is the one most Myanmar people are familiar with. But it is not widely used in creating ethnic language web pages.

In the Myanmar languages, the text which is available on the web is difficult to use as it is because they are available in numerous encoding (fonts) based formats. Applications developed for Myanmar languages have to read or process such text. The glyphs are shapes, and when two or more glyphs are combined together form a character in the scripts of Myanmar languages. To view the websites hosting the content in a particular font-type then one requires these fonts to be installed on local machine. Languages were made and available in that form.

Consider for example the word “school” written in the Roman Script and the Myanmar Script.

A character of English language has the same code irrespective of the font being used to display it. Figure [1]. However, most Myanmar language fonts assign different codes to the same character. Figure [2].

Font	Arial	Times New Roman
Word	S c h o o l	S c h o o l
Underlying byte code	83 99 104 111 111 108	83 99 104 111 111 108

Figure 1. Illustration of glyph code mapping for English fonts.

Arial and Times New Roman are used to display the same word. The underlying codes for the individual characters, however, are the same and according to the ASCII standard.

Font	CE-CLASSIC	Zawgyi One Unicode
Word	၎ က ျ ာ င် း	၎ က ျ ာ င် း
Underlying byte code	0061 0075 0073 006D 0069 0066 003B	1031 1000 103A 102C 1004 1039 1038

Figure 2. Illustration of glyph code mapping for Myanmar fonts.

The same word displayed in two different fonts in Myanmar, CE-CLASSIC and Zawgyi One Unicode. The underlying codes for the individual characters are according to the glyphs they are broken into. Not only the decomposition of glyphs and the codes assigned to them are both different but even the two fonts have different codes for the same characters.

3. Identification of Font Script

The widespread and increasing availability of textual data in electronic form in various font encoded form in Myanmar languages increases the importance of using automatic methods to analyze the content of textual documents. Nowadays in Myanmar, many software companies and researchers exhibited many Myanmar fonts (encodings). Myanmar computer users use several kinds of Myanmar fonts to type text document and create websites. Therefore, many authors and publishers will choose what types of font they use to write their books, magazine, journals and newspapers. However readers do not know what types of font are used in writing papers, books which they read and texts on the web pages which people read. Different organizations in Myanmar use different types of font. Therefore, knowledge is scattered between the groups and Myanmar people. For these reasons, font script identifier is essential to identify multiple Myanmar font scripts.

In this paper, we consider font as language and identify many Myanmar font scripts. We can consider font scripts identification as a classification problem. The identification and classification of the text or text documents based on their content to a specific encoding type (specially font) are becoming imperative. Previous works were done to identify the language and later to identify the encodings also. Most of them N-gram based modeling technique. It may be helpful to make the difference clear here, the term refers a ‘glyph’ and the document refers the ‘font-data (words and sentences) in a specific font-type’. The term frequency-Inverse Document Frequency (TF-IDF) approach is used to weight each term in the document according to how unique it is. In other words, the TF-IDF approach captures the

relevancy among glyph-sequence, font-data and font type. Here the glyph-sequence means unigram (single glyph), bigram ("current and next" glyph) and trigram ("previous, current and next" glyph) etc.

3.1 N-Gram Based Text Categorization

An N-gram is a subsequence of n items in any given sequence. In computational linguistics N-gram models are used most commonly in predicting words (in word level N-gram) or predicting characters (in character level N-gram) for the purpose of various applications. A character N-gram is a set of n consecutive characters extracted from a word. Typical values for n are 2 or 3: these correspond to the use of bigrams to trigrams, respectively. Character based N-grams are generally used in measuring the similarity of character strings. Character "N-gram" matching for computing a string similarity measure is widely used technique in information retrieval, stemming, spelling and error correction, text compression, language identification, font script (encoding) identification, text search and retrieval.

Table 1. Different N-grams for the word 'computer'

n-gram	COMPUTER
2-grams	-C, CO, OM, MP, PU, UT, TE, ER, R-
3-grams	-CO, COM, OMP, MPU, PUT, UTE, TER, ER-
4-grams	-COM, COMP, OMPU, MPUT, PUTE, UTER, TER-

Therefore, an N-gram is a character sequence of length n extracted from a document. It is an n character slice of a longer string. In this work one word from the document was represented as the set of N-grams. Here also leading and trailing spaces were considered as the part of the word

3.2 TF-IDF (Term Frequency- Inverse Document Frequency) Weights

The TF-IDF Weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document.

The term frequency in the given document is simply the number of times a given term appears in the particular document.

$$tf_i = n_i / n_k$$

with n_i being the number of occurrences of the considered term, and the denominator n_k is the number of maximum occurrences of term in that document.

The document frequency is the number of documents where the considered term has occurred at least once.

$$|\{d \in D \mid t \in d\}|$$

The inverse document frequency is the logarithm of the number of all documents divided by the number of documents containing the term t).

$$\log \frac{|D|}{|\{d \in D \mid t \in d\}|}$$

$|D|$ is the total number of documents in the document set; $|\{d \in D \mid t \in d\}|$ is the number of documents containing the term *t*.

The term frequency –inverse document frequency weights is

$$W = tf.idf$$

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

Using the cosine similarity between document d_j and query q can be calculated as:

$$sim(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

3.3 Overview of the system

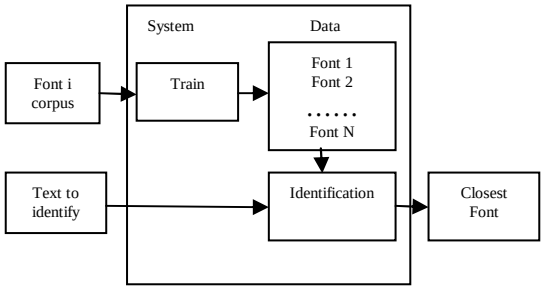

Figure 3. General architecture of the system

When we want to identify a text document, we build the n-gram frequency profile for this text and compare it with each font profile we have computed when training the system. And then, we will match the profiles and will calculate similarities between

test font and training fonts by applying TF-IDF (Term Frequency-Inverse Document Frequency) Weights. The system calculates the similarity from the profile of the unclassified text to each profile of the known fonts and chooses the nearest similarity font.

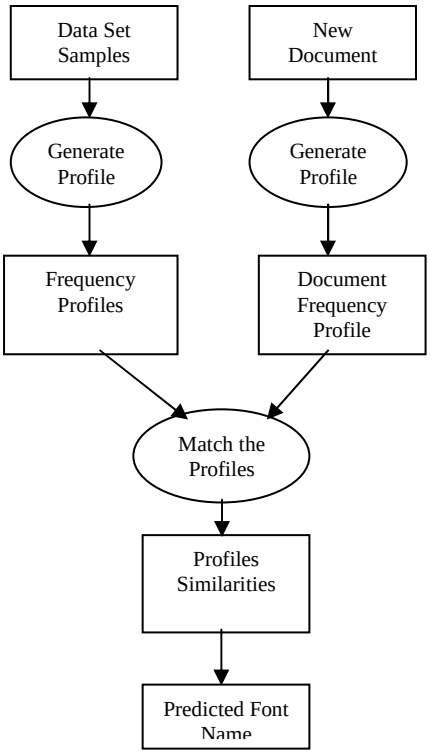

Figure 4. Design of the system

3.4 Modeling and Identification

***Data Preparation*:** For training we need sufficiently enough data of that particular type. And here we have collected and used more than 3000 unique sentences per font-type. it is collected and prepared manually. We tried nearly 10 different fonts of Myanmar language and one English font.

***Modeling*:** Generating a statistical profile for each font type using these TF-IDF weights is known as modeling the data. For modeling we considered four different types of terms. They are 2-gram, 3-gram, 4-gram and 5-gram.In raw text modeling the term refers the glyph based 2-gram or 3-gram or 4-gram or 5-gram.

A piece of text for 11 font scripts is taken for training. We would generate N-gram terms (2-gram, 3-gram, 4-gram and 5-gram) and TF-IDF weights of character N-grams for each font script are computed and stored as a profile for that particular font script.

The procedure for building the profiles is: we have taken all the provided data at once. And also we have considered four different kinds of terms for building profiles.

(i)First step is to generate the n-gram terms (2-gram, 3-gram, 4-gram and 5-gram)

(ii)Second step is to calculate the term frequency for the term like the number of time that term has occurred divided by the maximum number of term in that specific type of data.

$$tf_i = n_i / n_k$$

(iii)Third step is to calculate document frequency like in how many different data type that specific term has occurred.

$$|\{d \in D \mid t \in d\}|$$

(iv)Fourth step is to calculate inverse document frequency like all data types divided by the document frequency.

$$\log \frac{|D|}{|\{d \in D \mid t \in d\}|}$$

(v)Fifth step is to compute TF-IDF which is calculated like term frequency * inverse document frequency.

$$W = tf.idf$$

The common terms get zero values and other terms get non-zero values depending upon their term frequency values. From those values the profiles for each data type (font type) is generated.

Identification: While identifying the name of font script of input text document, the system first generates the terms (like 2-gram, 3-gram, 4-gram and 5-gram) of the text document, get the TF-IDF weight of each term from the profiles and calculate cosine similarity. Finally, the system calculates the similarities from the profile of the unclassified text to each profile of the known eleven fonts and the highest similarity scored of the font script is taken as a result.

$$sim(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

Table 2. 3-gram of the Myanmar word“ မြန်မာနိုင်ငံ ”of the Myanmar 3 Unicode font as document 1(d_1)

3-gram	Encoding number	f	tf	idf	w
- မြ	\u1056\u1019\u103B	1	1	Log2=0.3	0.
မြန	\u1019\u103B\u1014	1	1	Log2=0.3	0.
ြန်	\u103B\u1014\u1039	1	1	Log2=0.3	0.
နှ်မ	\u1014\u1039\u1019	1	1	Log2=0.3	0.
်မာ	\u1039\u1019\u102C	1	1	Log2=0.3	0.
မာန	\u1019\u102C\u1014	1	1	Log2=0.3	0.
ာန်	\u102C\u1014\u102	1	1	Log2=0.3	0.
နှ်	\u1014\u102D\u102F	1	1	Log2=0.3	0.
ိုင	\u102D\u102F\u1004	1	1	Log2=0.3	0.
်ငံ	\u102F\u1004\u1039	1	1	Log2=0.3	0.
ငံ	\u1004\u1039\u1004	1	1	Log2=0.3	0.
်ံ	\u1039\u1004\u1036	1	1	Log2=0.3	0.
ံ -	\u1004\u1036\u1056	1	1	Log2=0.3	0.

Table 3. 3-gram of the Myanmar word“ မြန်မာနိုင်ငံ ”of the WinInnwa font as document 2(d_2)

3-gram	Encoding number	f	tf	idf	w
- ြမ	h1Ah4Eh72	1	1	Log2=0.3	0.3
ြမန	h4Eh72h65	1	1	Log2=0.3	0.3
မန်	h72h65h66	1	1	Log2=0.3	0.3
နှ်မ	h65h66h72	1	1	Log2=0.3	0.3
်မာ	h66h72h6D	1	1	Log2=0.3	0.3
မာန	h72h6Dh45	1	1	Log2=0.3	0.3
ာန်	h6Dh45h64	1	1	Log2=0.3	0.3
နှ်	h45h64h6B	1	1	Log2=0.3	0.3
ိုင	h64h6Bh69	1	1	Log2=0.3	0.3
်ငံ	h6Bh69h66	1	1	Log2=0.3	0.3
ငံ	h69h66h69	1	1	Log2=0.3	0.3
်ံ	h66h69h48	1	1	Log2=0.3	0.3
ံ -	h69h48h1A	1	1	Log2=0.3	0.3

Table 4. 3-gram of the Myanmar word“ မြန်မာ” as user query (q) or test document

3-gram	Encoding number	f	tf	idf	w
- မြ	\u1056\u1019\u	1	1	Log3/2=0.1	0.18
မြန	\u1019\u103B\	1	1	Log3/2=0.1	0.18
ြန်	\u103B\u1014\	1	1	Log3/2=0.1	0.18
နှ်မ	\u1014\u1039\u	1	1	Log3/2=0.1	0.18
်မာ	\u1039\u1019\u	1	1	Log3/2=0.1	0.18
မာ	\u1019\u102C\	1	1	Log3/1=0.4	0.48

The similarity between document 1 and query test document is

$$sim(d_1,q)=\frac{(0.3*0.18)+(0.3*0.18)+(0.3*0.18)+(0.3*0.18)+(0.3*0.18)+(0*0.48)}{\sqrt{0.3^2+0.3^2+0.3^2+0.3^2+0.3^2+0^2}*\sqrt{0.18^2+0.18^2+0.18^2+0.18^2+0.18^2+0.48^2}}$$

$$= 0.8$$

The similarity between document 2 and query test document is

$$sim(d_2,q)=\frac{(0*0.18)+(0*0.18)+(0*0.18)+(0*0.18)+(0*0.18)+(0*0.48)}{\sqrt{0^2+0^2+0^2+0^2+0^2+0^2}*\sqrt{0.18^2+0.18^2+0.18^2+0.18^2+0.18^2+0.48^2}}$$

$$= 0$$

Therefore, the user query font is Myanmar 3 Unicode font.

4. Experiment Results

Testing Criteria: The system can identify font script of text documents. Therefore, the system is tested for many documents. While testing for the given ‘X’ input documents which are identified and calculates the identification accuracy in (%) as given below.

$$Accuracy = \frac{N}{M}$$

Where *N:* number of correctly identified tokens and *M:* total number of tokens.

***Testing and Results*:** For the given ‘X’ number of different input documents we are identifying the closest profiles and from them we are calculating the accuracy as explained above. It is done (repeatedly) for various (2-gram, 3-gram, 4-gram and 5-gram) categories. The accuracy of identification is 100%.

***Diversity Assumption*:** The accuracy of a font encoding identifier depends on the number of encodings from which the identifier has to select one. This is about how many encodings are assumed to be in the world. In practical terms, this is reflected in the number of encodings for which the system has been trained.

Font-Type Identification: The testing is done for 500 unique sentences per font-type. We have added English data as also one of the testing data set, and is referred to as English-Text. Even we type 100 characters (two or three sentences), the name of font script of that characters are identified. Some Myanmar Unicode fonts use the same code point for the same character. Therefore, we use more characters to identify for Unicode font scripts. The similarity scored between input test font script and trained font scripts is described in Figure 5.

However, most Myanmar language fonts assign different codes to the same character and Some Myanmar Unicode fonts use the same code point for the same character. Therefore, the similarity scored between the same encoding fonts is near and the similarity scored between the different encoding fonts is far.

Figure 5. Similarities between the input text and trained fonts

5. Conclusion

Nowadays in Myanmar, many computer users use several kinds of Myanmar encoding to type documents and create websites. Different organizations in Myanmar use different types of font. Therefore, knowledge written in documents are scattered between the groups and Myanmar people. For these reasons, font script identifier is essential to identify multiple Myanmar font scripts. Therefore, we can fetch information correctly and convert it to standard encoding (font script). And then we can collect knowledge sources. We can apply our system in Search Engine and Machine Translation. Automatic identification of a script in a given document facilitates many important applications such as automatic archiving of multilingual documents, searching online archives of document and for the selection of script specific text understanding system in a multilingual environment. In this paper, we have discussed N-gram based text categorization and the new TF-IDF weights based approach for font identification. Even we have one page of A4 paper as trained data per font-type, we will identify font script of the input text document. The proposed system can be used for other language scripts as well with minimal modification. As we can identify font scripts (encodings), we will identify languages such as Myanmar language, Japanese language, Indian language and so on. Therefore, we will implement language identifier to identify many languages for our country and we will convert multiple encodings (font scripts) to standard encoding as further extension.

References

[1] Munirul Mansur, Naushad UzZman and Mumit Khan, “Analysis of N-gram based Text Categorization for Bangala in a newspaper corpus,” Proc. of 9th International Conference, 2006.

[2] Muntsa Padro and Lluis Padro, “Comparing methods for Language Identification,” Proceedings of the XX Congreso de la Socieded, 2004.

[3] P.Kishore, “Identification and Conversion on font-data in Indian Languages,” In International Conference on Universal Knowledge, 2007.

[4] http://www.myanmarnlp.net.mm

[5] http://en.wikipedia.org

[6] http://www.mcf.org.mm