Retrieval of English-Myanmar Transliterated Words Based on Soundex Algorithm

Yin Yin Cho, Tin Myat Htwe
University of Computer Studies, Yangon
yinyincho.ucsy@gmail.com ;

Abstract

Transliteration is the process of mapping from one system of writing into another, word by word. This system presents an algorithm for English-Myanmar cross-language transliterated word retrieval. Soundex matching is used to identify strings that may be of similar pronunciation, regardless of their actual spelling. This is done by retrieving texts based on phonetic codes of keywords.

This paper presents a system that retrieves transliterated Myanmar words for entered English words by using the Soundex Algorithm.In this system, Soundex codes for vowels in phonetic sound are also included. The Myanmar consonants categorize into similar phonetic groups according to Soundex code.

1. Introduction

Transliteration is the process of mapping text written in one language in to another by means of a pre-defined mapping. It is useful when a user knows a language but does not know how to write its script. It is also useful in case of unavailability of a direct method to input data in a given language. Hence, transliteration can be understood as the process of entering data in one language using the script of the another language. In general, the mapping between the alphabet of one language and the other in a transliteration scheme will be as close as possible to the pronunciation of the word. English transliterated text has found widespread use with the growth of internet usage, in the form of chats, mails, blogs and other forms of individual online writing. This kind of transliterated text is often referred by the words formed by a combination of English and the language in which transliteration is performed.

Transliterating the two languages back and forth loses some information; two corresponding words may not be exactly matched. String Matching can be defined as the process of determining whether two strings are instances of the same string. Soundex is the best known phonetic encoding algorithm. It keeps the first letter and converts the rest into numbers according to an encoding table.

This system implements English word to a set of similar pronunciation of Myanmar words using a set of soundex algorithm and new modified encoding table. The system enables retrieval of words containing either the English keywords or the corresponding English-to-Myanmar transliterated words. Matching is used to identify strings that may be similar matching. Matching is used in applications such as word retrieval, where the spelling of a word is used to identify other strings that are likely to be similar matching.

This paper is organized as follows. Section 1 is the introduction, section 2 is related work. Soundex algorithms are presented in section 3. Section 4 is the proposed system design and section 5 is the system implementation and sample case study for Transliteration process. Section 6 is the conclusion and future work of the system.

2. Related Work

Nowadays internet usage is spreading all over the world. There are many web sites in Myanmar and it includes a lot of forums and blogs. Moreover, chatting is the most popular online applications. Most of Myanmar people, use English in Myanmar pronunciation in writing forums or talking online. It is a bit annoyed in reading those words. This system presents transliteration of English language (pronounced in Myanmar) into Myanmar Language. Bilingual dictionaries do not solve the problem since most of the transliterated words are not found in the dictionaries. It can be also used for to measure phonetic similarity between words. It is based on Soundex algorithm. In this paper, soundex codes are modified to fit into Myanmar Language.

There are previous works in transliteration process. The algorithms presented in [1] and [12] transcribe English and Japanese words into intermediate codes and use exact code matching during retrieval. Since transliterating the two languages back and forth loses some information, two corresponding words may not be exactly matched . Whereas the algorithm in [7] encodes each Katakana word into a phonetic string representation and uses partial matching with English words. Two words are considered to be in transliteration relation when the number of matched characters is more than a certain threshold. The algorithm uses a depth-first search which trends to take longer time than a straightforward matching so that some heuristics are incorporated to reduce search time. [8] presents an algorithm for encoding English word to a set of possible Thai sounds using a set of encoding tables and rules. The encoding tables are not fully elaborated and no details on effectiveness of the methods are reported.

The soundex algorithm developed by Russell was an early attempt at assigning a common phonetic code to similar sounding words and names. In [3] and [11], authors offer substantial enhancements to the original approach. In Metaphone [4, 5] and PHONIX [9], the input to these phonetic encodings or “sound-alike” algorithms is a word, and the result is an encoded key, which should be the same for all words that are pronounced similarly, allowing for a reasonable amount of fuzziness. The basic principle behind these phonetic matching schemes is to partition the consonants by phonetic similarity, and then use a single key to encode each of these sets. Strings that sound similar compare equal in their respective encoded form. For these particular algorithms, only the first few consonant sounds are encoded, unless the first letter is a vowel.

In the Daith-Makotoff Soundex system, it partitions the set of letters into seven disjoint sets, assuming that the letters in the same set have similar sound. Each of these sets is given a unique key, except for the set containing the vowels and the letters h, w, and y, which is considered to be silent and is not considered during encoding.

An advantage of Soundex is the small table size and simplicity of the letter-by-letter algorithm, which can provide significant speedup over the other phonetic methods.

3. Soundex Algorithms

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The Soundex algorithm can be used for identifying names that sound alike but are spelled different. It is the best known phonetic encoding algorithm. It keeps the first letter and converts the rest into numbers according to an encoding table. The idea of indexing information is how it sounds, rather than alphabetically was born.

Soundex codes begin with the first letter of the surname followed by a three-digit code that represents the (first three) remaining consonants. Zeros will be added to names that do not have enough letters to be coded. Soundex algorithm consists of following steps.

- Soundex algorithm steps are initially retaining the first letter of the string in step 1.
- Step 2 removes all occurrences of (a, e, i, o, u, h, w, y) unless they are first letter.
- Step 3 is assigning the remaining letter (b, f, v, p) for 1, (c, g, j, k, q, s, x, z) for 2, (d, t) for 3, (l) for 4, (m, n) for 5 and (r) for 6 respectively.
- If two or more letters with the same number were adjacent in the original name (before step 1), then omit all but the first.
- Step 4 fill with 0 if there are < 3 digits or otherwise drop the rightmost digits.
- Step 5 is returning the first four characters.

Phonetic categories of English letters are as follows:

- a, e, i, o, u, y = The vowels ("oral resonants").
- b, f, p, v = The labials and labio-dentals.
- c, g, k, q, s, x, z = The gutterals and sibilants.
- d, t = The dental-mutes.
- l = The palatal-fricative.
- m = The labio-nasal.
- n = The den to or lingua-nasal.
- r = The dental fricative.

3.1. Variations of Soundex Algorithms

There are several rules for variations of Soundex Algorithm shown as below.

American Soundex system is an improvement of Soundex algorithm. It handles some multi-character n-grams and maintains relative vowel positioning, whereas Soundex does not.

Henery code, devised by Louis Henry, is based on the Russell Soundex method but is adapted for the French language and classifies each name as a three-letter code. Like the Russell Coding Technique, the Henry method can also result in completely different names being brought together as with the French names Mireille, Marielle and Merilda which are all given the code MRL. A second form of the Henry algorithm returns codes that are not limited to three letters; however, it is still prone to mismatching as it often modifies the phonetic structure of the names, resulting in it sometimes missing good links or linking completely different names.

Daitch-Mokotoff Soundex (D-M Soundex) was developed in 1985 by genealogist Gary Mokotoff and later improved by genealogist Randy Daitch because of problems they encountered while trying to apply the Russell Soundex to Jews with Germanic or Slavic surnames (such as Moskowitz vs. Moskovitz or Levine vs. Lewin). The improvements of this system are: Information is coded to the first six meaningful letters rather than four. Initial letter is coded rather than kept as is. When a letter or combination of letters may have two different sounds, it is double coded under the two different codes. A letter or combination of letters maps into ten possible codes rather than seven.

4. Proposed System

This system presents the displaying in Myanmar language if user types in English with Myanmar pronunciation. It is mainly based on the Soundex algorithm. In this system, vowels play the main role since 'man' and 'min' are pronounced differently even though they have the same Soundex code. Therefore, this system modifies the Soundex code table according to Myanmar language, for example, H and HL have different Soundex code for their pronunciation. Ha is pronounced as (ဟ) and Hla is pronounced as (လှ). It uses Soundex codes and transaliteration is performed using Soundex rules, explained in section 5.1.

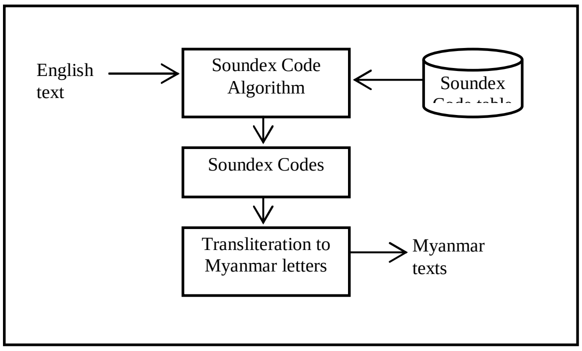
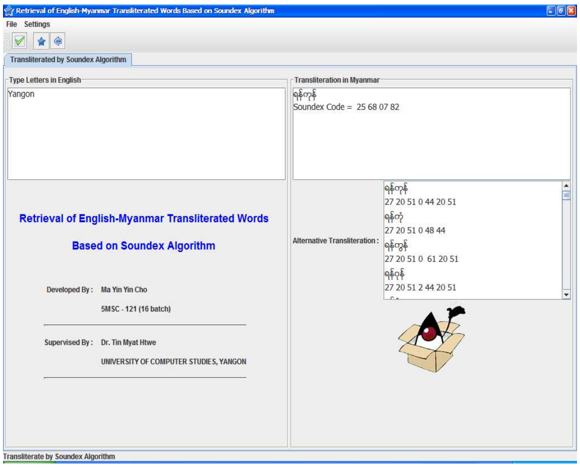


Figure 1: Proposed System Design

5. System Implementation

This system is implemented using Java programming language. It is implemented as windows based program. When user enters English word (in Myanmar pronunciation) and it will output in Myanmar Language. The implementation is shown in Figure 2.


Retrieval of English-Myanmar Transliterated Words Based on Soundex Algorithm
File
Settings
Transliterated by Soundex Algorithm
Type Letters in English
Transliteration in Myanmar
Alternative Transliteration :

Yangon

ရန်ကုန်
Soundex Code = 25 68 07 82

ရန်ကုန်
27 20 51 0 44 20 51
ရန်ကုံ
27 20 51 0 48 44
ရန်ကွန်
27 20 51 0 61 20 51
ရန်ဂုန်
27 20 51 2 44 20 51

Retrieval of English-Myanmar Transliterated Words

Based on Soundex Algorithm

Developed By : Ma Yin Yin Cho
5MSC - 121 (16 batch)

Supervised By : Dr. Tin Myat Htwe
UNIVERSITY OF COMPUTER STUDIES, YANGON

Transliterate by Soundex Algorithm

Figure 2: System Implementation

5.1 Myanmar Soundex Code Rules

In this paper, Soundex algorithm is modified to fit for Myanmar Language. It has following rules.

- Words are coded into Soundex codes first.
- In this system there are two types of Soundex codes, consonants and vowels (consonant code for ကာ၊ ခ and vowel code for အဲ၊ အိ)
- In the conventional Soundex system, number of codes are fixed.
- In this system, number of codes are not fixed, it depends on the length of the word.
- There are two digits for each code, if it is a consonant code, code digit number is less than 50 and if it is a vowel code, it is more than 50.
- When vowels (A, E, I, O, U) are started in the word, both consonant code and vowel codes are used for single word combination, otherwise, for the start of the word, only consonant code is used.
- For the vowels inside the word, word combination, character behind the vowel is used to get the vowel code. For example, phonetics of (AT and AM) are different.
- When adjacent letters have the same code number, they are coded as one sound.
- Several letters and letter combinations pose the problem that they may sound in one of two ways. The letter and letter combinations, (HN, HL), are assigned according to Soundex code table.
- One Soundex code may have alternations in Myanmar words. For example g can be pronounced as ကာ၊ ဂ၊ ဃ.

5.2 Soundex Encoding Table

Soundex algorithm is modified to code for vowels for the phonetic sounds. This system keeps codes for both consonants and vowel as shown in following tables. In the table, soundex codes with same pronunciations are grouped together.

Table 1: Soundex Code for Main Sound (Consonants)

Soundex Code	**English**	**Myanmar Tone**	**Alternatives**
00	A, E, I, O, U	အ	
01	B, BH	ဘ	ဗ
02	CH	ချ	ခြ
03	CY, CI, S, X	စ	ဆ

04	C, K, KH, Q	က	ခ
05	D	ဒ	ဓ
06	HP, PH, F	ဖ	ပ
07	G, K	က	ဂ ဃ
08	GI, J, GY	ဂျ	ဂြ ကျ ကြ
09	H	ဟ	
10	HL	လှ	
11	HN	နှ	
12	KY	ကျ	ကြ
13	L	လ	
14	M	မ	
15	MY	မြ	မျ
16	N	န	ဏ
17	NY	ည	
18	Oo	ဦ:	
19	P	ပ	ဖ
20	PY	ပြ	ပျ
21	QU, KW	ကွ	
22	HT T	ထ	တ
23	V	ဗ	
24	W	ဝ	
25	Y	ရ	ယ
26	Z	ဇ	ဈ
27	GW	ဂွ	ကွ
28	MW	မွ	
29	NW	နွ	
30	LW	လွ	
31	SH, SIA	ရှ	ယှ
32	TH	သ	ထ
33	R	ရ	
34	HM	မှ	
35	BRU	ဘရ	
36	NG	င	
37	MH	မှ	
38	YW	ရွ	
39	DW	ဒွ	ဓွ
40	PHY	ဖြ	
41	PW	ပွ	
42	LY	လျ	

Table 2: Soundex Code Table for Vowels

Soundex Code	**English**	**Myanmar Tone**	**Alternatives**
50	AY, AYE	ေ -း	ေ -
51	AB, AC, AD, AF, AG, AK, AT	- က်	-တ်
52	A, EH, AH, UH	-	- ာ
53	ARR, ARE	- ာ	- ား
54	AI, EI, I, AING, INE	-ို င်	-ို င်း
55	ORE	-ိုး	
56	AW, AU	ေ -ာ်	ေ - ာ
57	AUNG, AUN	ေ-ာင်	ေ -ာင်း
58	EE, E, I	-ီ	-ီး၊ -ိ
59	EW	-ုး	
60	IB, IT, IP	- စ်	
61	IN	- င်	- င်း
62	INT	- င့်	
63	OO, U	-ူး	
64	OH, O, IO	-ိုး	
65	OW	ေ - ာင်း	
66	AN, AND, ND, EN, AM, UM, AN	-န်	-မ်၊ -ံ
67	OHN, OWN	-ုန်း	-ုံး
68	OUT, OGK, OK, OAT, AUK	ေ - ာက်	ေ- ာတ်
69	OE	-ိုး	
70	AL, EL, E`	-ဲ	-ယ်
71	WAL, WEL	-ဲ	
72	OT, UT	-ုတ်	-ွတ်၊ ေ - ာက်
73	ATE, EIK, IGHT	-ိတ်	-ိပ်၊ -ိက်
74	ON, ONG	-ုန်	-ွန်၊ -ုံ
75	AIK	-ိုက်	
76	ANT	-န့်	
77	OI	-ွိုး	
78	WE	ေ -ွ	-ီး
79	EIN, AIN	-ိန်း	-ိမ်း

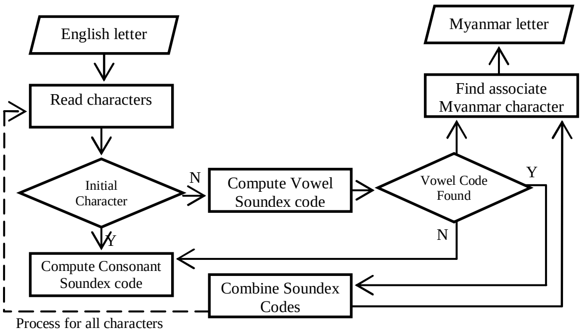

English letter
Read characters
Initial Character
N
Y
Compute Vowel Soundex code
Vowel Code Found
Y
N
Myanmar letter
Find associate Myanmar character
Compute Consonant Soundex code
Combine Soundex Codes
Process for all characters

Figure 3: System Flow Chart

System Flow Chart is shown in Figure 3. When user enters an English letter, it is decomposed into single characters. Then each character and character pairs are matched against Soundex codes and encoded. Then related Myanmar characters are extracted and paired into Myanmar letter. System Input and output are shown as below:

- System input – English Language, for example 'Yangon'
- System process – Soundex algorithm (Soundex Code), Transliterated to Myanmar pronunciation and Myanmar letter generation
- System output – Myanmar Language, 'ရန်ကုန်'

5.3 Case Study

In this system, there are variations for the encoded soundex codes. Case Study 1 presents the encoding system without variations, i.e., there is only one result for the transliteration and Case Study 2 presents the encoding system with variations, i.e., there is more than one pronunciation for one soundex code.

Input: 'Yangon'
Processing:

- Soundex code for initial letter, 'Y' = 25 (consonant code)
- Soundex code for 'AN' = 66 (vowel code)
- Soundex code for 'G' = 07 (consonant code)
- Soundex codes for 'ON' = 74
- 25 = ရ, 66 = - န်, 07 = က and 74 = -ုန်

Output: ရန်ကုန်

Input: 'Dataw'
Processing:

- Soundex code for initial letter, 'D' = 05 (consonant code)
- Soundex codes for 'AT' = 51 (vowel code)
- Next character pair is 'AW', which is only in vowel codes, therefore, last word of previous word pair 'T' is processed. Soundex code for 'T' = 22
- Soundex codes for 'AW' = 56
- 05 = ဒ၊ဓ, 51 = - က်၊ -တ်, 22 = တ၊ထ and 56 = ေ -ာ်၊ေ -ာ

Output: ဒက်တော်၊ ဓက်တော်၊ ဒတ်တော်၊ ဓတ်တော်၊ ဒက်ထော်၊ ဓက်ထော်၊ ဒတ်ထော်၊ ဓတ်ထော်၊ ဒက်တော၊ ဓက်တော၊ ဒတ်တော၊ ဓတ်တော၊ ဒက်ထော၊ ဓက်ထော၊ ဒတ်ထော၊ ဓတ်ထော။

The first answer is the default answer, therefore, in this case, ဒက်တော် is the output for the input 'Dataw'. Sometimes the first answer may not be correct since the Soundex codes store for all general cases. The length of the word is not limited in this system. User can also enter white spaces between words. If there are white spaces, the input string is tokenized into multiple words and each and every word is transliterated. It generates output, with nearly similar pronunciation in Myanmar. It does not include double words, for example, input word 'Mandalay' will get the output as 'မန်ဒလေး'.

6. Conclusion

This paper presents the English-Myanmar cross-language transliterated word retrieval. The retrieval is done by using phonetic codes retrieval based on a Soundex coding rather than searching for the word themselves. It enables retrieval of corresponding Myanmar words from English language transliterated words. It assists the user easily to get the information that requires the Myanmar words. By using this system, the user can easily get corresponding Myanmar word and can reduce time consuming to type the Myanmar words reviewing information and can get their related information.

7. References

[1] A. Kumano, "Building a technical term dictionary with Katakana-English Matching", Gengoshorigakai - Annual Conf. of the Japanese Association for Natural Language Processing, (in Japanese) Japan, March, pp.221-223.

[2] D. E. Knuth, The Art of Computer Programming, Vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts, 2nd edition, 1982.

[3] J. Celko. Joe Celko's SQL For Smarties: Advanced SQL Programming. Morgan Kaufmann Publishers, Inc., 1995.

[4] L. Phillips, "Hanging on the Metaphone", Computer Language, 7(12), 1990.

[5] L. Phillips, “The Double Metaphone Search Algorithm”, C/C++ Users Journal, 18(6), June, 2000. Also available online at http://www.cuj.com/documents/s=8038/cuj0006philips/.

[6] O. Htun, S. Kodama and Y. MikaMi, "Measuring Phonetic Similarities in Myanmar IDNs", IDNA – Internationalizing Domain Names in Applications, http://www.icann.org/en/topics/idn/first-track/idna-protocol-en.htm

[7] N. Collier, A. Kumano, and H. Hirakawa, “Acquisition of English-Japanese proper nouns from noisy-parallel newswire article using Katakana matching”, Proc. of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, Dec. 2-4, pp. 309-320.

[8] S. Ongroongruang, R. Prongsirivattana, and V. Jantarasukree, “English to Thai Word Retrieval Using Sound Index”, Proc. 2nd SNLP’95, Bangkok Thailand, Aug. 2-4, 1995, pp. 407-413.

[9] T. N. Gadd, “PHONIX: The Algorithm”, Program, 24(4), pp. 363-366, 1990.

[10] The Soundex Algorithm, available online at http://www.archives.gov/research_room/genealogy/census/soundex.html.

[11] U. Pfeifer, T. Poersch, N. Fuhr. Searching Proper Names in Databases.

[12] Y. Matsuo and S. Shirai, “Using pronunciation to automatically extract bilingual word pairs”, Shizengengoshori, (in Japanese), November, pp.101-106.