Phonetics Matching Approach for Converting System of Phonetics Transcriptions to Myanmar Text

Kyaw Kyaw Maung
kyawkyawmaung.ucsm@gmail.com

Abstract

Converting system of phonetics transcriptions to Myanmar text is intended to contribute as a portion of speech to text system of Myanmar language. In Myanmar language, the spoken language may be different with the written language in many cases. Phonetics transcriptions represent spoken language and Myanmar text represents written language. This paper proposed the phonetic string matching scheme that aligned between the phonetics transcriptions and the Myanmar text. Phonetics string matching is a research area of information retrieval systems, search systems, data mining, text and web-mining, speech recognition, and pattern recognition systems. The proposed system idea is come from the popular phonetic string matching algorithm called Soundex. The main difference between the proposed system and the original Soundex algorithm is that the proposed system is completely converted from text to phonetics symbols and then encoded of that symbol to relevant code points for accuracy.

1. Introduction

In a speech to text system, one of the important problems is mismatches between the spelling of words and the pronunciation of that word. In Myanmar language, the different tone level affects on the meaning of the word. According to the tone level, the meaning of the word may be different. There are many words with the same or nearly the same meaning, different spellings and the same or nearly the same pronunciations. These words are especially found in the names of person, places, and things and so on. These words may be viewed as same in phonetically but different in spellings. Basically, a syllable of Myanmar language can be formed by combining of consonants and vowels. There are thirty three consonants in Myanmar language. But they can be represented in phonetically by using twenty two phonetic symbols and it shows there are many consonants that are the phonetically the same in pronunciations with different spellings. In Myanmar language, the word boundary cannot be identified by using space. According to the nature of Myanmar language, the proposed phonetic matching system is a convenience approach for extracting of Information from large amount of text strings in Myanmar language.

Phonetic matching can be described as the process of identifying a set of strings are most likely to be similar in sound to a given query string. Phonetic matching is the process of identifying strings that have the same sound, after elimination of possible transmission or cognition errors. Transmission errors include sound-alike mistakes in data entry, mishearing of a spoken name on an imperfect transmission medium or errors like Chinese whispers errors. Cognition errors include mistaking a pronunciation for an expected word. [1]

This paper contributes the method for phonetic matching for Myanmar language that is based on the popular Soundex system and its improved versions. But the main different between the proposed system and the original Soundex is that these system is completely converted from phonetics representations of Myanmar Text for improve the matching. The code points of the proposed system are defined on phonetic representations of consonants, consonant clusters and vowels of Myanmar language. Then the system matches the source string and query string by using phonetic matching. The final result can be achieved by converting from phonetic to Myanmar text.

2. Phonetic Encoding Methods

Most of the phonetic encoding methods are converted from a string into a code according to the pronunciation of that strings. The phonetic encoding methods may be language dependent on how to define a code of a specific language. According to the code point specified standard, the matching results may be different even when the same algorithm is used. Phonetic matching has been developed mainly in English. Several techniques have been used for other languages, some European languages, Hindi, Marathi, Japanese, and Sindhi and so on. Phonetic matching is a European languages, Hindi, Marathi, Japanese, and

Sindhi and so on. Phonetic matching is a type of approximate string matching [1, 2, 3].

2.1. Soundex

Soundex system is a phonetic algorithm for indexing strings by sound. It is originally based on English language pronunciation. The algorithm mainly encoded consonants. Vowels will not be encoded if it is the first letter. The Soundex code for a string consists of a letter followed by three numerical digits. The letter is the first letter of a string. The digits encode the remaining consonants. Similar sound consonants are grouped into the same digits.

The Soundex indexing system can be described as follows:

1. Retain the first letter and drop all vowels of a, e, i, o, u and y, h, w.
2. Replace consonants with following code.
    - b, f, p, v → 1
    - c, g, j, k, q, s, x, z → 2
    - d, t → 3
    - l → 4
    - m, n → 5
    - r → 6
3. Remove all of the duplication of code numbers.
4. Continue until to get three code numbers. If it have more than three letters, retain the first three numbers. If it cannot assign three numbers, append with zeros until there are three numbers.

The major drawback of the Soundex algorithm is that it retains the first letter, and it may cause any error or variation at the beginning of a string generates a different Soundex code [2]. In some language like Hindi and Marathi there is an ambiguity to match when using Soundex algorithm exactly. For example, the two strings as ‘sandy’ and ‘sandyha’ for phonetic matching, the Soundex generate the same code as ‘s530’. By look at the code, both strings are phonetically matching, but both the strings are phonetically different [4, 5].

2.2. Metaphone

Metaphone is a phonetics algorithm for indexing words by English pronunciations. It improves on the Soundex algorithm by using information about variations and inconsistencies in English spellings and pronunciations to produce a more accurate encoding to match words and names with similar sounds. Original metaphone algorithm has many errors and the improved version of metaphone is developed as the Double metaphone algorithm. And both of these algorithms is improved as Metaphone3 which corrects thousands of miscoding that will produced by the first two versions. [7]

2.3. NYSIIS

The New York State Identification and Intelligence System (NYSIIS) is a phonetics algorithm devised in 1970. It features an accuracy increase of 2.7% over the traditional Soundex algorithm. Its transformation rules are similar with other Soundex based algorithms, but it returns a code that is only made of letters. [9, 2]

3. Proposed System Architecture

The proposed system is the intended to solve the problem of mismatches between the spoken language and the written language in Burmese. Myanmar Dictionary and English-Myanmar Dictionary show that there are many Myanmar words that have mismatches between the spoken language and the written language. Some pronunciations are not matched in some writing words.

Myanmar Consonants and Phonetics Symbols

Character groups	Myanmar Characters and Phonetics symbols representations				
က group	က k	ခ k□	ဂ □	ဃ □	င ŋ
စ group	စ s	ဆ s	ဇ z	ဈ z	ည/ညျ □
ဋ group	ဋ t	ဌ t□	ဍ d	ဎ d	ဏ n
တ group	တ t	ထ t□	ဒ d	ဓ d	နံ n
ပ group	ပ p	ဖ p□	ဗ b	ဘ b	မ m
non-group	ယ j	ရ □	လ l	ဝ w	သ θ/ð
		ဟ h	ဠ l	အ ə/a	

The system is firstly converted from Myanmar text into phonetics transcriptions. Both Myanmar text and Phonetics transcriptions are based on Unicode fonts. Myanmar3 and Charis SIL Phonetics Fonts are

used for these systems. Phonetics symbols are followed by International Phonetics Associations Standards (IAP) and these symbols are show in Table 1.

Table 1 presents the consonant and its corresponded phonetics symbols. In thirty-three Myanmar consonants, there are twenty-two speech sounds because some of the consonants have the same pronunciation. So, the phonetic symbols have same representation in some consonants [10, 11].

Table 1. Non-nasalized Vowels

No	Tone 1 (သံတို)	Tone 2 (သံရှည်)	Tone 3 (သံလေး)
1.	အ a□/ə	အာ a¯	အား aˆ
2.	အိ i□	အီ i¯	အီး iˆ
3.	အု u□	အူ u¯	အူး uˆ
4.	အေ့ e□	အေ e¯	အေး eˆ
5.	အဲ့ □□	အယ် □¯	အဲ □ˆ
6.	အော့ □□	အော် □¯	အော □ˆ
7.	အို့ o□	အို o¯	အိုး oˆ

Table 2. Nasalized Vowels

No	Tone 1 (သံတို)	Tone 2 (သံရှည်)	Tone 3 (သံလေး)
1.	အန့် ã□	အန် ã¯	အန်း ãˆ
2.	အင့် □̃□	အင် □̃¯	အင်း □̃ˆ
3.	အိန့် e□̃□	အိန် e□̃¯	အိန်း e□̃ˆ
4.	အုန့် o□̃□	အုန် o□̃¯	အုန်း o□̃ˆ
5.	အိုင့် a□̃□	အိုင် a□̃¯	အိုင်း a□̃ˆ
6.	အောင့် a□̃□	အောင် a□̃¯	အောင်း a□̃ˆ
7.	အွန့် □̃□	အွန် □̃¯	အွန်း □̃ˆ

Table 3. Glottal Stop Vowels

No	Glottal Stop Vowels
1.	အတ်/အပ် □□
2.	အစ် □□
3.	အိတ်/အိပ် e□□
4.	အက် □□
5.	အုတ်/အုပ် o□□
6.	အိုက် a□□
7.	အောက် a□□
8.	အွတ်/အွပ် □□

In Myanmar vowels, there are three types of vowels namely, (1) non-nasalized Myanmar Vowels (2) Nasalized Myanmar Vowels and (3) Glottal stop Myanmar Vowels. In Non-nasalized Myanmar Vowels and Nasalized Myanmar Vowels, there are three types of different tone level that shown in Table 2 and Table 3. There are 22 non-nasalized vowels, 21 nasalized vowels and 8 glottal stop vowels [10, 11].

When converting phonetics text to Myanmar text, the problem of the mismatches between the spoken language and the written language may be occurred. Examples are show in Table 5.

Table 4. The mismatches examples between writing and pronunciation in Myanmar

No	Writing in Myanmar	Pronunciation in Myanmar
1.	ကျော့+ကွင်း □□□+kwiˆ	ကျော့+ဝင်း □□□+□wiˆ
2.	ထိုင်+ခုံ t□a□̃¯+ k□o□̃¯	ထိုင်+ဂုံ t□a□̃¯+ □o□̃¯
3.	လေး+တင်း+ခွဲ leˆ+t□̃ˆ+k□□ˆ	လေး+ဒင်း+ဂွဲ leˆ+d□̃ˆ+□□ˆ
4.	သက်+တော်+စောင့် θ□□+t□¯+sa□̃□	သက်+တော်+ဇောင့် θ□□+t□¯+za□̃□

3.1. Phonetics Matching for Myanmar Language

Proposed phonetics matching scheme is done on phonetic transcriptions of Myanmar language. The goal is for encoded to the same representation on both writing and spoken language of Myanmar language. Some consonant may be changed into another consonant form when change between the written and the speaking. Also some consonants have the same pronunciations. These consonants are kept into the same group. The proposed phonetics string matching for Myanmar language is the following:

- Retain the first consonant letters of every word and drop all occurrences of nasalized vowels, non-nasalized vowels and glottal stop vowels.
- Replace consonants with digits are as follows(after the letter):
  - k, k□, □, ŋ, □ → 1
  - s, z → 2
  - t, t□, d, n → 3
  - p, p□, b, m → 4
  - j, □, l, w, θ, ð → 5
  - h → 6
- Remove all of the duplication of code numbers.
- Continue until to get three code numbers. If it have more than three letters, retain the first three numbers. If it cannot assign three numbers, append with zeros until there are three numbers.

3.2. Phonetics Matching for Myanmar Language

The proposed system is intended to reduce of differences between the written language and the spoken language of Myanmar language. For example, one of the name of city from Myanmar namely YANGON can be write in the written language of Myanmar is ရန်ကုန် /jã˗ ko□̃ˆ/. But the pronunciation (spoken language) of these words is ယန်ဂုန် /jã˗ □o□̃ˆ/. In this case, the spoken consonant sound ဂ /□/ sound has changed in the written language as က /k/. Vowels sounds အုန် /o□̃ˆ/ are the same in the spoken language and the written language. As another instance, the city name TAUNGGYI can be written in the correct spelling of written language is - တောင်ကြီး /ta□̃˗ kiˆ/. These spelling has change in speaking as တောင်ဂြီး /ta□̃˗ □iˆ/. In this case the consonant sound က /k/ has changed into ဂ /□/ sound. Another examples are already shown in Table 5. But In Myanmar language, every words may not be always changed between the writing and the speaking. For example, ခရမ်းသီး /k□ə jãˆ θiˆ/ is the same in both reading and writing.

In phonetics string matching scheme, the possible changes of consonant sounds are categorized into a same group, and that group is defined as a specific number like 1, 2, 3, and so on. In a speech to text system, the input of the speech signals is represented as the spoken language and it is need to adjust between the spoken language and the written language. In this case, the Soundex like phonetic string matching scheme can be adjust between the spoken language and the written language. The step by step of the phonetics matching can be described the following case study:

Problem: To match the spoken language of ခေါင်း ပေါင်းစ /k□a□̃ˆ paˆ sə/ and the written language of ဂေါင်းဗေါင်းဇ /□a□̃ˆ paˆ zə / is the same or not?

To Solve these problem, firstly, the writing word ခေါင်း ပေါင်းစ /k□a□̃ˆ paˆ sə/ is reduced into k□ps by removing all vowels. Second step is replace all consonants with digits, so, k□→1, p→4 and s→2. So, the final result for the writing of ခေါင်း ပေါင်းစ /k□a□̃ˆ paˆ sə/ is 142.

The spoken word ဂေါင်းဗေါင်းဇ /□a□̃ˆ paˆ zə/ if first reduced into □pz by removing all vowels. And then replace all consonants with digits, so, □→1, p→4 and z→2. So, the final result for the spoken language of ဂေါင်းဗေါင်းဇ /□a□̃ˆ paˆ zə/ is equal to 142.

So, the system decided that the two strings ခေါင်း ပေါင်းစ /k□a□̃ˆ paˆ sə/ and ခေါင်း ပေါင်းစ /k□a□̃ˆ paˆ sə/ are the same strings.

4. Future work

The original Soundex system's drawback has been shown in section 2.1. The new Soundex based algorithms like [3, 4, 5] has try to solve the weakness of original algorithm by adding some rules to original Soundex algorithm. Instead of directly used of the Soundex algorithm, these researches changes to original algorithm to match their specific language. The original Soundex algorithm is an approximate string matching algorithm. So, it cannot be exact match between strings. The rate of exact match is rely on how correctly defined the consonants that are assigned in the same group. In Myanmar language, the inconsistent of the speaking and writing is exists not only on the consonants but also in the vowels. For the consonant sound changes, the proposed system can be solved successful. But when the vowel sound changes, the proposed system cannot be solved. In the future, the proposed system can be solved these problems by

defining which vowels can be changed in reading and writing and evaluate the vowels into matching steps. According to the experimental results of some research [2], the phonetics matching techniques are generally much faster than other pattern matching techniques.

5. Conclusion

The proposed system is aligned between the written language and the spoken language of Myanmar language for matching. Because this method is one of the approximate string methods, the result may not be matched in exact. The future work of this research is to build a complete converting system between phonetics and Myanmar language. The language model is needed to develop for the proposed system.

References

[1] J. Zobel and P. Dart, "Phonetic String Matching: Lessons from Information Retrieval", *In Proceedings of the 19th ACM International Conference on Information Retrieval* (*SIGIR '96*), 166–172.

[2] P. Christen, "A Comparison of Personal Name Matching: Techniques of Practical Issues", *Technical Report, Department of Computer Science, The Australian National University, Canberra, Australia*, September, 2006.

[3] M. Yasukawa, J.S. Culpepper, and F. Scholer, "Phonetic Matching in Japanese", *In Proceedings of SIGIR 2012 Workshop on Open Source Information Retrieval, August 16, 2012, Portland, Oregon, USA.*

[4] S. Chaware and S. Rao, "Rule-based Phonetic Matching Approach for Hindi and Marathi", *In Proceeding of Computer Science & Engineering: An International Journal (CSEIJ)*, Vol1, No.3, August 2011, pp. 13-24.

[5] S. Chaware and S. Rao, "Analysis of Phonetic Matching Approaches for Indic Language", *In Proceeding of International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 1, Issue 2, April 2012, pp. 110-116.

[6] C. Snae, and M. Bruckner, "Novel Phonetic Name Matching Algorithm with a Statistical Ontology for Analysing Names Given in Accordance with Thai Astrology", *Issues in informing Science and Information Technology*, *Volume 6*, 2009, pp. 497-515.

[7] http://en.wikipedia.org/wiki/Metaphone

[8] L. Philips. "Hanging on the Metaphone", *Computer Language, Vol. 7, No. 12 (December)*, 1990.

[9] http://en.wikipedia.org/wiki/New_York_State_Identification_and_Intelligence_System

[10] The Myanmar Language Commission, Ministry of Education, Union of Myanmar, "MYANMAR-ENGLISH Dictionary", *Issue of 10th Edition, 2011*,

[11] T. Tun, "Acoustic Phonetics and the Phonology of the Myanmar Language", *Issues of 1st Edition*, 2007.