Currency Word Translation from Pa-Oh to Myanmar

Nan Khin Pyone Myint, Dr. Myint Myint Khaing
University of Computer Studies (Pinlon)
Dawnankhinpyone@gmail.com

Abstract

In Natural Language Processing (NLP), Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. In Natural Languages of Myanmar and Pa-Oh language, word boundary identification is not easy between words with spaces. In this system, Pa-Oh to Myanmar language translation framework is proposed. Word tokenizing plays a vital role in most Natural Language Processing. Syllabification is also a important task in Pa-Oh. Working directly with characters does not help. It is therefore useful to syllabify texts first. The first step is entering the Pa-Oh words. And then the system syllabified the input word by looking up syllable files. After tokenizing the input words, each word examines whether they are in word list/dictionary or not quickly. Finally it can display correct currency words of Myanmar with the same meaning of Pa-Oh language.

1. Introduction

Nowadays, Natural Language Processing (NLP) is one of the important things for communication and understanding among people. There are many natural languages are used among people around the world according to their native groups. In computerized systems the translation processes are needed to communicate among different kinds of groups. The Pa-Oh also known as Thaungthu and Black Karen form an ethnic group in Myanmar, comprising approximately 600,000. The Pa-Oh form the second largest ethnic group in Shan State, and are classified as part of the “Shan National Race” by the government, although they believed to be of Tibeto-Burman stock, and are ethnolinguistically related to the Karen. They populate Shan State, Kayin State, and Kayat State[10]. Myanmar language, also known as Burmese, 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 (Ethnologue, 2005). Burmese is a tonal and analytic language using the Burmese script. Burmese characters are rounded in shape and the script is written from left to right [9].

The first approach of NL translation is word segmentation. Many of word segmentation ambiguities were resolved at the level of syllable segmentation. And then lexicon helps the meaning of words by integrating the definitions of vocabularies between two languages. The lexicon will give the definitions of the Pa-Oh words to Myanmar words. Therefore the arrangement of dictionary is Pa-Oh to Myanmar vocabularies. In this system, building of the Pa-Oh-Myanmar lexicon is the main aim of Natural Language translation.

The rest of the paper is organized as follows: related work is described in section2. Section 3 represented the syllabification using the longest string matching algorithm and corpus collection and collecting words list. Section 4 describes design and implementation of the Word Translation from Pa-Oh to Myanmar Language. Finally, conclusion gives in section 5.

2. Related works

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right regular inter-word spacing, although inter-spacing may sometimes be used. Myanmar characters can be classified into three groups: consonants, medials and vowels. The basic consonants in Myanmar can be multiplied by medials. Syllable or words are formed by consonants combining with vowels. However, some syllables can be formed by just consonants, without vowel. Other characters in the Myanmar scripts include special characters, numerals, punctuation marks and signs [5]. 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 is descended from the Brahmi script of ancient South India [6].

A Myanmar syllable has a base character, and may also have (or not) a pre-base character, a post-base character, an above-base character and a below-base character [7]. A syllable is a basic sound unit or a sound. A word can be made up of one or more syllables. Every syllable boundary can be a potential word boundary. In some cases, a word can include other words, in which case it is called a compound word. In Myanmar, a syllable is formed based on rules that are quite definite and unambiguous. A syllable can contain multiple consonants, multiple medials and multiple vowels [5].

In order to clarify the syllable structure, characters of the Myanmar script are classified into twelve categories. A Myanmar syllable consists of one initial consonant, zero or more medials, zero or more vowels and optional dependent various signs. Independent vowels, independent various signs and digits can act as stand-alone syllables. A finite state machine or finite state automaton (FSA) can be employed to demonstrate the syllable structure of Myanmar script [8].

3. Pa-Oh Writing System, Syllabification and Corpus Collection

As most of other languages Pa-Oh script is syllabic in nature, and written from left to right. It has no space between words and syllable segmentation represents a significant process in many NLP tasks such as word segmentation, sorting and so on. Syllable segmentation can be provided based on the created rules. Pa-Oh script is a writing system constructed from consonants, consonant combination symbols, vowel symbols related to relevant consonants and diacritic marks indicating tone level. Pa-Oh language consists of 33 consonants, consonants combination symbols, basic vowels and diacritic marks indicating tone level and Killer.

3.1 Consonant. In Pa-Oh script, the letters represent a consonant syllable. Both Pa-Oh and Myanmar languages are used same consonant basic letters. As Myanmar Language, Pa-Oh language is composed of 33 consonants script; the letters represent a consonant as shown in Table 1.

Table1. Pa-Oh Consonants

Basic Consonants				
က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
ဟ	ဠ	အ		

3.2. Vowels. Pa-Oh language has two kinds of vowels. They are dependent vowel signs and independent vowel signs. There are eight dependent vowel signs and seven independent vowel sighs as shown in Table2 and Table 3.

Table2. Dependent vowel signs

Dependent vowel signs				
ေ–	–ု	–ူ	–ဲ	–ံ
–ိ	–ီ	–ာ	–ါ	

Table3. Independent vowel signs

Independent vowel signs				
ဣ	ဤ	ဥ	ဦ	
ဧ	ဩ	ဪ		

3.3 Basic Medial. Medials are known as “Byee Twe” in Pa-Oh language. There are 3 basic medials and 2 combined Table 3 and Table 4.

Table4. Basic Medial

Basic Medial		
–ျ		
ြ–		
–ွ		

3.4 Combined Medial. u (က), c (ခ) have special variant forms when used medially a modifiers of the syllable’s vowel. They combine with the preceding character, i.e. au (ကေ), uG (ကွ). It is also possible to fine two medials associates with a consonant MuG (ကြွ), usG (ကျွ). They are combining medial as shown in Table5.

Table5. Combined Medial

Combined Medial				
–ျ	+	–ွ		
ြ–	+	–ွ		

3.5 Killer, Diacritic and Kinzi. A killer sign (virama) sign nominally serves to cancel (or kill) the inherent vowel of the consonant letter to which it applied. There is one sign “ –်” which serves as a virama sign in Pa-Oh characters. Diacritics are defined as a sign that can be written above or below a letter to indicate a difference in pronunciation from the same letter when unmarked. There are five signs which serve as diacritics in Pa-Oh characters. Two are situated below “– ့” and “–ႇ” , the other one above “–ံ” and the last two are following “–း”, “–ႈ ” the consonant letter. Different vowel sign combinations can be combined again with one or any two of these diacritics. In addition to Myanmar Language, there are two diacritics in Pa-Oh language such as “–ႇ” (MAI NGA) and “–ႈ”(MAI PAK NGA) [3]. Kinzi is a special form of devowelised Nga (MYAMMAR

LETTER NGA) with the following letter underneath, i.e., subjoined. In this case, if the character after the second consonant is an Asat and the next character after Asat is an invisible Virama sign, then there should be no syllable before the second and third consonant. Kinzi also consists of two syllables but it is treated as one syllable in written form. The representation of Pa-Oh script can be seen in Table6.

Table6. Killer and Diacritic

Asat (Killer)	Diacritic					Kinzi
	anusvara	Dot	Visarga	Mai Nga	Mai Pak Nga	
်	ံ	့	း	ႇ	ႈ	င်္

3.6 Numerals

Pa-Oh numerals are decimal-based and Table7 shows zero to nine in sequence. No thousand separators are used; instead, spaces are sometimes used between digits for easy reading.

Table7. Numerals

Numerals				
၀	၁	၂	၃	၄
၅	၆	၇	၈	၉

3.7 Syllabification

As an initial attempt we use longest string matching algorithm for Pa-Oh text syllabification. Here it go from left-to-right in a greedy manner in figure 1[2].

```
Load the syllables from syllable -file
Load the sentences to be processed from
Store all syllables of length j in Nj where j=10…1
for-each sentence do
  length → length of the sentence
    Pos →0
while (length >0) do
for j = 10…1 do
  for- each syllable in Nj do
      if string-match sentence (pos,pos+j) with syllable
      Syllable found Mark syllable
        pos →pos+j
      length →length-j
    End if
  End for
 End for
End while
Print syllabified string
End for
```

Figure1.Longest string matching algorithm

3.8 Corpus Collection

Development of lexical resources is a very tedious and time consuming task and purely manual approaches are too slow. We collect Pa-Oh currency corpus about 500 syllables. It uses Wynnpao TRUE TYPE FONT to collect these syllables. In order to use this system the user must install Wynnpao font.

3.9 Collecting word list

Word lists and dictionaries in electric from are of great value in computational linguistics and NLP. Here we describe our efforts in developing a word list for Pa-Oh. Word lists and dictionaries in electronic form are of great value in computational linguistics and NLP.

3.9.1 N-gram

N-gram models are a type of probabilistic model for predicting the next item in a sequence. N-grams are used in various areas of statistical natural language processing and genetic sequence analysis. An n-gram is a subsequence of n items from a given sequence. The items in question can be phonemes, syllables, letters, words or base pairs according to the application. An n-gram of size 1 is referred to as a “unigram”: size 2 is a “bigram” (or, less commonly, a “digram”); and size 4 or more is simply called an “n-gram” [9].

3.9.2Using Bi-grams in text categorization

Text categorization is a fundamental task in Information Retrieval. The standard approach to text categorization has so far been using a document representation in a word-based space, i.e. as a vector in some high dimensional Euclidean space where each dimension corresponds to a word. This method relies on classification algorithms that are trained in a supervised leaning manner [4].

3.9.3Syllable N-grams

Pa-Oh language uses a syllabic writing system unlike English and many other western languages which use an alphabetic writing system. Interestingly, almost every syllable has a meaning in Pa-Oh language. It have been developed scripts in Perl to syllabify words using the list of syllable as a base and their generate n-gram statistics. Almost all monograms are meaningful words. Many bi-gram are also valid words and as it move towards longer n-grams, it generally get less number of valid words [1] [2].

4. Design and Implementation of the System

There are four steps in the Word Translation from Pa-Oh to Myanmar Language in Figure2. The first step is entering the Pa-Oh words. And then the system syllabified the input word by looking up syllable files and tokenized each word look up word list or dictionary quickly. Finally, system gives correct words of Myanmar currency with the same meaning of Pa-Oh language.

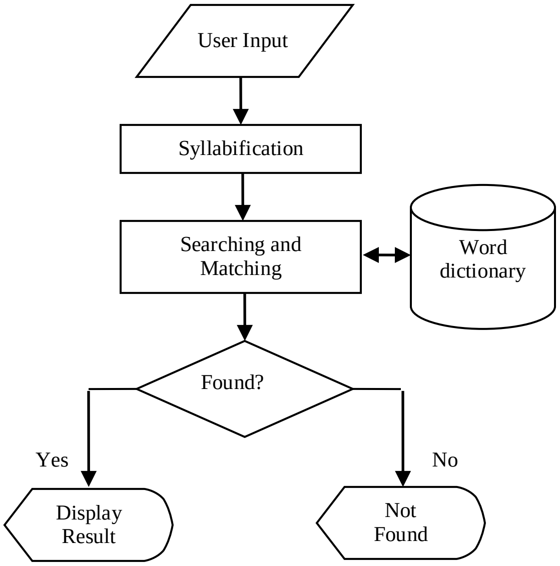


Figure2. Design and Implementation of the System

4.1 Sample of Word Translation from Pa-Oh to Myanmar language

Input sample sentence – တစွဲးဟားခန်

Syllabification- တ | စွဲး | ဟား | ခန်

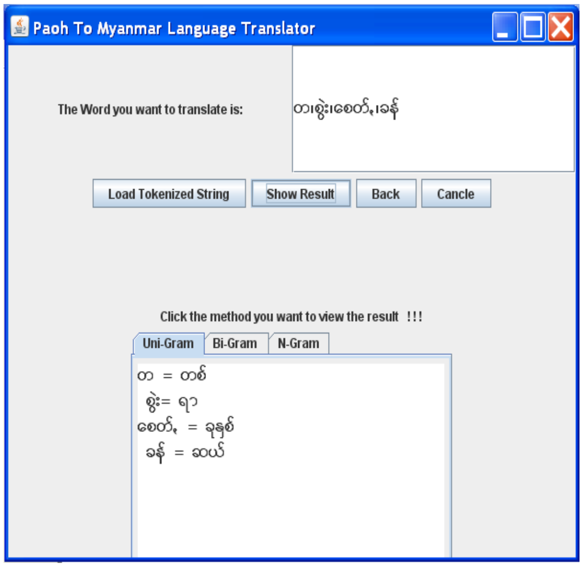
Checking with dictionary- describe by unigram, bi-gram and n-gram as shown in Table 8.

Table 8 Sample different kinds of syllables

Uni-gram 1-syllable	Bi-gram 2-syllables	N-gram 4 0r more syllables
တ (တစ်)	တစွဲး (တစ်ရာ)	တစွဲးဟားခန်
စွဲး (ရာ)	ဟားခန် (ငါးဆယ်)	
ဟား (ငါး)		
ခန် (ဆယ်)		

4.2 Implementation of the system

The input Pa-Oh words are segmented by longest string matching algorithm and then checking with dictionary which described unigram, bi-gram and n-gram as shown in figure 3, 4 and 5.

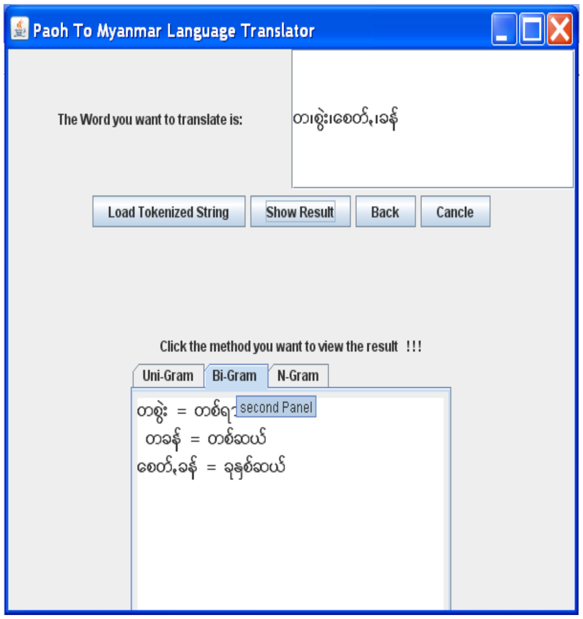


The Word you want to translate is: တ၊စွဲး၊စေတ့်၊ခန်

Click the method you want to view the result !!!

တ = တစ်
စွဲး= ရာ
စေတ့် = ခုနှစ်
ခန် = ဆယ်

Figure3.Display Uni-gram result

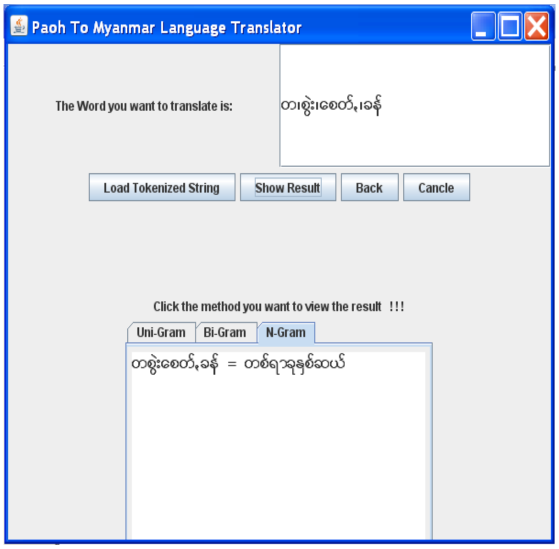


The Word you want to translate is: တ၊စွဲး၊စေတ့်၊ခန်

Click the method you want to view the result !!!

တစွဲး = တစ်ရ
တခန် = တစ်ဆယ်
စေတ့်ခန် = ခုနှစ်ဆယ်

Figure4.Display Bi-gram result



The Word you want to translate is: တ၊စွဲး၊စေတ့်၊ခန်

Click the method you want to view the result !!!

တစွဲးစေတ့်ခန် = တစ်ရာခုနှစ်ဆယ်

Figure5. Display N-gram result

5. Conclusion

Syllables are building blocks of words and syllable segmentation is essential for the language processing of Pa-Oh script. In this paper, the system has described the need and possible techniques for segmentation on Pa-Oh script. The segmentation rules were created based on the characteristics of Pa-Oh syllable list. This system can be applied in other NLP applications such as Information Retrieval Systems, Information Extraction Systems and Machine Translation. The statistical construction of machine readable dictionaries has many advantages in sorting and word segmentation. A test corpus containing 500 Pa-Oh syllables was tested in the program. A complete syllabification algorithm for Pa-Oh script can be further implemented by applying this algorithm.

6. References

[1] Hla Hla Htay, Kavi Narayana Murthy,"Myanmar Word Segmentation using Syllable level Longest Matching", Department of Computer and Information Sciences University of Hyderabad, India

[2] Hla Hla Htay, G.Bharadwaja Kumar, and Kavi Narayana Murthy, "8.5 Statistical Analyses of Myanmar Corpora"

[3] Michael Everson and Martin Hosken, "Proposal for encoding one additional Myanmar character for Pa'o Karen in the UCS".

[4] RON BEKKERMAN, JAMES ALLAN, "Using Bigrams in Text Categorization", Department of Computer Science University of Massachusetts Amherst, 01003 USA, {ronb|allan}@cs.umass.edu, December 27, 2003

[5] Tun Thura Thet; Jin-Cheon Na , Wunna Ko Ko,"Word Segmentation for the Myanmar Language".

[6] Unicode Consortium, The Unicode Standard 4.0: Southeast Asian Scripts (Addison Wesley, California, 2004).

[7] Zaw Htut, Myanmar-Thai Co-workshop on Myanmar Language Implementation, Input Methods and Basic Encoding in Myanmar Language. Available at: http://www.myanmars.net/unicode/doc (accessed 1 January 2007).

[8] Zin Maung Maung and Yoshiki Mikami, "A Rule Based Syllable Segmentation of Myanmar Text".

[9] N-gram
http://en.wikipeda.org/wiki/N-gram

[10] Pa-O
http://www.en.wikipedia.org/wiki/Shan_State