Proposed Framework for Pali Words to Myanmar Text Translation

Phyu Phyu Khaing, Khine Zar Thwe
University Of Computer Studies Mandalay (UCSM)
phyuphyukhaing07@gmail.com, khinezarthwe.mdy@gmail.com

Abstract

The aim of this paper is to assist Buddhists and the new Buddhist students who are unfamiliar with some of the Pali words often used in the study of Buddhism. Moreover this can also assist the students of philosophy or orientalist who has to know the terminology of their field, which for common parlance is mostly not less 'unfamiliar' than are the words of the Pali language found in the dictionary. As Myanmar is Buddhism country, needs to deeper understand Buddha-Dhamma. In this study both building of Pali-Myanmar dictionary and Pali-Myanamr words segmentation to translate the given Pali-Myanmar scriptures to Myanmar text are proposed. So proposed system has two sections, first is building of Pali-Myanmar dictionary and second is translation for Pali words to Myanmar text in which Pali-Myanmar words is first needed the segmentation to provide translation of Pali-Myanmar words to Myanmar text . So words segmentation method is used with the combination of rule-based syllable segmentation and dictionary-based matching method for translation of Pali words to Myanmar text.

Index Terms*: Pali to Myanmar text, Pali-Myanmar dictionary, Pali segmentation, rule-based segmentation methods, Natural Language Processing.*

1. Introduction

Myanmar language is the official language of Myanmar and it is tonal, pitch-register and syllable-time language. Moreover the nature of Myanmar language is monosyllabic and analytic language. The sentence order is "subject-object-verb". It also known as Burmese and it is a member of the Lolo-Burmese grouping of the Sino-Tibetan language family. The language uses a Brahmic script called the Burmese script. The Burmese alphabet consists of 33 letters and 12 vowels, and is written from left to right. It requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. Besides the Burmese language, the Burmese alphabet is also used for the liturgical languages of Pali and Sanskrit.

Pali is the language used to preserve the Buddhist canon of the Theravada (ဓေထရဝါဒ) Buddhist tradition (is a branch of Buddhism that uses the teaching of the Pali Canon, a collection of the oldest recorded Buddhist texts, as its doctrinal core), which is regarded as the oldest complete collection of Buddhist texts surviving in an Indian language. Pali is closely related to Sanskrit, but its grammar and structure are simpler. Traditional Theravadins regard Pali as the language spoken by the Buddha himself, but in the opinion of leading linguistic scholars, Pali was probably a synthetic language created from several vernaculars to make the Buddhist texts comprehensible to Buddhist monks living in different parts of northern India. As Theravada Buddhism spread to other parts of southern Asia, the use of Pali as the language of the texts spread along with it, and thus Pali became a sacred language in Sri Lanka, Myanmar, Thailand, Laos, Cambodia and Vietnam. Pali has been used almost exclusively for Buddhist teachings, although many religious and literary works related to Buddhism were written in Pali at a time. So Pali is a spoken language, written in the script of the land where it is used: for example in Myanmar, it is written in Myanmar script. Strictly speaking Myanmar and Indic scripts are abugidas (alpha-syllabaries) and not alphabets [2].

2. Related Work

First attempt of an authentic dictionary of Buddhist doctrinal terms, used in the Pali Cannon and its Commentaries, this present will fill a real gap felt by many students of Buddhism. It provides the reader not with a mere superficial enumeration of important Pali terms and their Myanmar equivalents, but offers him precise and authentic definitions and explanations of canonical and post-canonical terms and doctrines, based on Sutta, Abhidhamma and Commentaries, and illustrated by numerous quotations taken from these sources, so that, if anyone wishes, he could, by

intelligently joining together the different articles, produce without difficulty a complete exposition of the entire teachings of Buddhism. This study intends for those who study the Buddhist teachings through the native Myanmar language but wish to familiarize themselves with some of the original Pali terms of doctrinal import. This paper assists the students of philosophy or orientalist who has to know the terminology of their field, which for common parlance is mostly not less 'unfamiliar' than are the words of the Pali language found in the dictionary.

The Pali language has not one script but many, the fact that there are so many scripts is hardly a pretext for learning none of them. The greatest number of books and manuscripts are found in Sinhalese, Burmese, Khmer-Muul, or closely related scripts of South-East Asia (Lao-Dhamma, Lanna, etc.). There are also some modern Indian publications that typeset Pali in Devanagari (i.e., the same script used for modern Hindi and Sanskrit), and, of course, the modern vernacular script of Thailand has been adapted to print Pali (although the classical tradition uses Lanna in Thailand's North-West and Khom throughout the rest of the country). Moreover Pali language had a significant influence on Myanmar language since ancient times. Buddhist monks and researcher study the Pali language for many religious works and to understand Buddhist Literature which is used in Pali language. So computerized Pali translator is needed.

Moreover in this section, previous works of Myanmar language processing are [11]. In this study, a rule-based approach of syllable segmentation algorithm for Myanmar text is proposed. Segmentation rules were created based on the syllable structure of Myanmar script and a syllable segmentation algorithm was designed based on the created rules. Moreover a segmentation program was developed to evaluate the algorithm and simulates the results. [4] Proposed an efficient algorithm based on syllable rules matching. In order to evaluate the algorithm, a prototype has been developed to measure the accuracy of syllabification. In [10], focus on exploiting limited amounts of dictionary resource on low resource language such as Myanmar language, in an attempt to improve the segmentation quality of an unsupervised word segmenter.

For segmentation, three models are proposed in their study, generate model using set of dictionary, inserting n-gram model into the generated model and third is combination of first two models. The [7] reference study proposes an approach that will segment the input sentence to build meaningful words and tag these words with appropriate POS tags. Including experiments, this approach has the best performance for known words with a few ambiguous tags and the results show that their approach achieves high accuracy (over 90%) for different testing input. In paper [9] describes the use of two machine learning techniques, Naïve Bayesian classifier (NB) and transformation-based learning (TBL), to address the task of assigning function tags to Myanmar sentences. Function tagging is a process of assigning syntactic categories like subject, object, time and location to each word in the text document. It is an important step in Natural Language Processing. Function tags can help to improve the performance of Myanmar to English machine translation system. In this paper, we present a comparison of two methods in our experiments. The results showed that TBL was better and outperformed NB and there was a slight difference between the results.

3. Natural Language Processing

[1] NLP must ultimately extract meaning ('semantics') from text: formal grammars that specify relationship between text units' parts of speech such as nouns, verbs, and adjectives address syntax primarily.

Natural Language Processing is the ability of computational systems to deal with human language in spoken language or written form. One of the main is the important application of Natural Language Processing (NLP) is Machine Translating from one language into another language. The various form of knowledge is necessary to solve the ambiguities of semantic. The identification of Verb Phrase plays an important role in many natural language processing applications, such as partial parsing, information retrieval and machine translation.

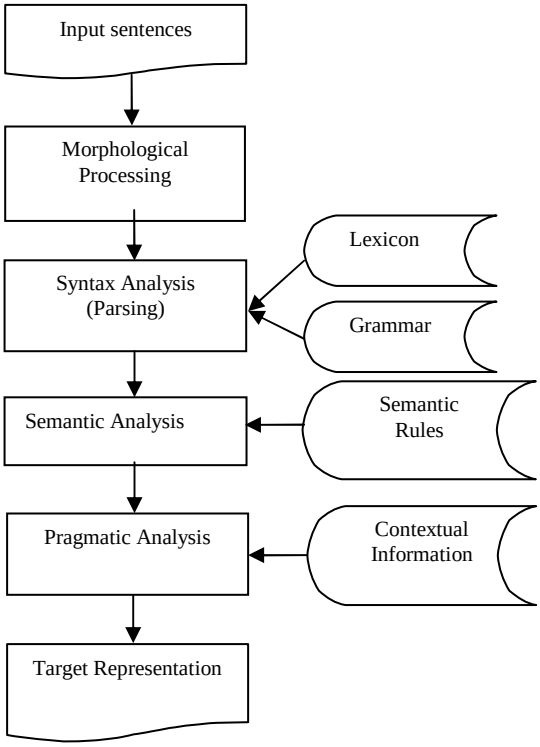
The preliminary stage which takes place before syntax analysis is morphological processing. In many languages such as Japanese and Myanmar, the morphology of words can be ambiguous in ways that can only be resolved by carrying out syntactic and/or semantic analysis on the input.

The output from the morphological processing phase is a string of tokens which can then be used for lexicon lookup. These tokens may contain tense, number, gender and proximity information (depending on the language) and in some cases may also contain additional syntactic information for the parser. The next stage of processing is syntax analysis.

Syntax and Semantics

A language processor must carry out a number of different functions primarily based around syntax analysis and semantic analysis. The purpose of syntax

analysis is two-fold: to check that a string of words (a sentence) is well-formed and to break it up into a structure that shows the syntactic relationships between the different words. A syntactic analyzer (or parser) does this using a dictionary of word definitions (the



lexicon) and a set of syntax rules (the grammar). A simple lexicon only contains the syntactic category of each word; a simple grammar describes rules which indicate only how syntactic categories can be combined to form phrases of different types.

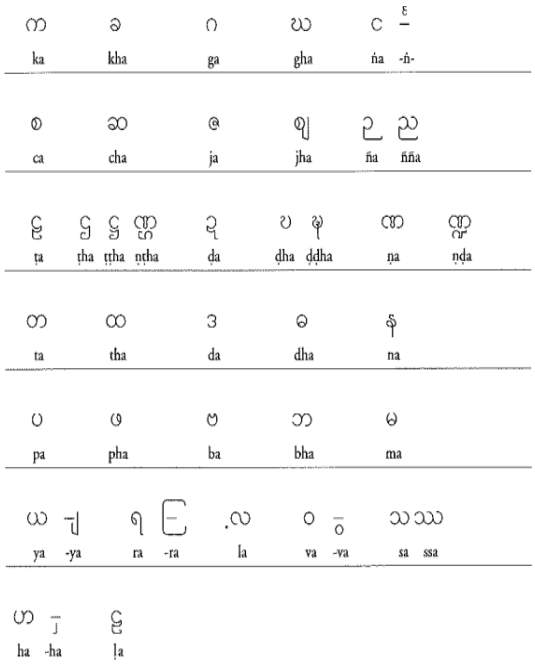
Figure 1. The logical steps in Natural language Processing

Semantic analysis is the term given to the production of this formalized semantic representation.

In order to carry out semantic analysis the lexicon must be expanded to include semantic definitions for each word it contains and the grammar must be extended to specify how the semantics of any phrase are formed from the semantics of its component parts. After semantic analysis the next stage of processing deals with pragmatics. Unfortunately there is no universally agreed distinction between semantics and pragmatics.

3.1. Nature of Pali Words

Pali-Myanmar words are formed by spelling of Pali into Myanmar words. So Pali-Myanmar words are



used in Myanmar consonants, vowels. The modern Myanmar script designated "round Myanmar script". The consonants of Pali-Myanmar are

Consonants subscript and superscript consonants used in writing double consonants are given only if there is some variation from the usual form of the consonant. The main difference between Pali and Myanmar is the transliteration of the sign အ. This sign docs not designate a sound in itself, hut indicates the inherent vowel a or serves to write the vowels used in connection with consonants. This means that there are two ways to write initial vowels other than a and â in Myanmar. By using an inverted comma (') to designate a in Myanmar words, the transliteration makes it possible to return to the original script. The inverted comma is not used in Pali texts or words as this would lead to confusion because the Romanization of Pali has been long established. The letters involved are: အိ(i) or (i), အီ (î), or ဤ (î), အု (u) or ဥ(u), အူ (û) or ဦ (û) အေ(e) or ဧ (e) and အော(o) or ဩ (o). [2]

Pali-Myanmar words need to decide syllable and word boundaries. Each syllable can be constructed multiple characters.

Example: သိက္ခာ (က္ခာ has two consonants)

The system has the ability to recognize Pali words with conjunct consonants and Pali words without conjunct consonants. A conjunct consonant compose of two consonants letters and also called double consonant.

In Myanmar script, conjunct consonants are added in a subscripted consonant form. Some Pali-Myanmar words are composed of without consisting conjunct consonants and these are same structure as Myanmar text.

3.2. Stacked consonants

The practice of putting smaller consonants underneath the syllabic ending to start off another syllable is known as stacked consonants. Since Burmese has merged all of its consonant endings into a glottal stop (like the uh in "***Uh*** oh"). Thankfully, stacked consonants are confined to loan words, usually Pali.

Example: မေတ္တာ (တ္တာ has two same consonants)

Myanmar phonological tones have basically four tones as in following example.

အ {a.} -- the creaky tone is extra-short [ă]

အာ {a} -- the medial [a] is of medium length

အား {a:} -- made up of two segments [a] and [:] (or [a]) is long and empathic. The so-called "tone #4" is a rime where a checked vowel is followed by a "killed" (or an {a.thût}) consonant,

အက် {ak} = [æk]

3.3. Encoding of Pali Text

[6]The basic consonants and vowels of Myanmar words are relatively obvious in how they are encoded, by examining the character charts. In Unicode system, characters are stored in the order in which they are read.

Example: ကြောင့် is stored in Unicode system of cyber is က+ြ + ေ + ာ+င + ့+်

Unicode standard states that vowels are stored after the consonant, according to how they are pronounced. Moreover Syllable Chaining, Devoweliser, how kinzi is represented in Unicode, stacked characters and Asat rules are need to follow the syllable segmentation of Myanmar text encoding system.

In terms of the Unicode standard, Pali-Myanmar text is encoded. Unicode for Pali-Myanmar text is divided into groups they are

1. Consonants

Examples: က (1000), အံ(1021 1036), အး(1021 1038)

2. Independent Vowels

Example: သြ (1029)

3. Dependent Vowels

Example: ာ(102C) , ေ်(1031 102C 103A)

4. Conjuncts

Example: ္ထြ (1000 1039 1010 103C 102D)

5. Finals

Example: င်္ (1004 103A 1039 1002 103A)

Proposed system syllable segmentation must follow the Unicode encoding system for Pali-Myanmar text to get machine understanding.

4. Proposed System Architecture

Proposed System has two main part first is computerized Pali-Myanmar translation and second is computerized creation of Pali-Myanmar dictionary. Pali-Myanmar dictionary is not fixed dictionary system user can add new words and edit existing words or delete the existing words.

4.1. Syllable Word Segmentation

Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language. There are seven word segmentation methods as follows;

1. translation with human translators segmentation
2. translation with character breaking
3. syllable breaking
4. syllable breaking and Maximum matching
5. unsupervised word segmentation
6. syllable breaking, maximum matching and unsupervised word segmentation
7. supervised word segmentation

Most of the Myanmar syllable segmentation are Normal Form or Backus-Naur Form, rule based segmentation and dictionary-based segmentation

method. In this proposed system, there are two main steps as in the following figure.

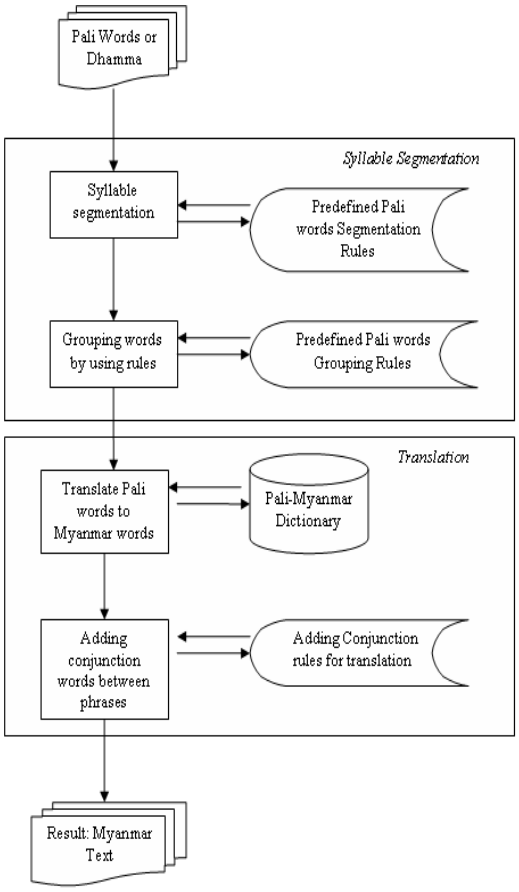


Figure 2. System Flow Diagram for Proposed System

In the Syllable segmentation block is first step of proposed system and it intends to segment Pali words by using predefined Pali words segmentation rules. Pali-Myanmar words are made up of more than one consonant and it can use stacked consonants like မေတ္တာ. The segmentation process, needs to defined Pali words segmentation rules depends on the nature of Pali word and structure of Pali words. And then system segment the Pali words as the following;

Example:

Input Pali words: ဗုဒ္ဓဂုဏောအနန္တော

After segmentation: ဗုဒ္ဓ ဂုဏော အနန္တော

Second section of the proposed system is translation from Pali words to Myanmar text. Translation use Pali-Myanmar dictionary by using dictionary based matching and statistical approach using N-gram of syllables were combined. Moreover conjunction words need to add at suitable places to get equivalent meaningful Myanmar text.

Example: Segmented words: ဗုဒ္ဓ ဂုဏော အနန္တော

After Search and comparing with Pali-Myanmar dictionary, pair the Pali word with its meaning like below;

ဗုဒ္ဓ = မြတ်စွာဘုရား

ဂုဏော = ဂုဏ်တော် ကျေးဇူးတော်

အနန္တော= အဆုံးမရှိ ကျယ်ဝန်း သော

4.2. N-gram Based Phrases Extraction from Corpus

Pali-Myanmar text does not place space between words. Thus, system needs to segment the input Pali words. System creates Pali words by N-gram method for input segmented Pali sentence to search the corpus in the dictionary. In this situation each segmented word is defined one word. There is no standardized corpus for Pali-Myanmar language and this proposed system attempt to be more comprehensive. So the proposed system used dictionary based matching and statistical approach using N-gram of syllables were combined.

Table 1. Pali-Myanmar word syllable word

No	1-gram	2-gram	3-gram
1	တေ	ဒေသ	မုတ္တမံ
2	မေ	သီလ	စရဏ

Finally, proposed system has to add suitable conjunction words for the suitable places, so the final result is as below.

"မြတ်စွာဘုရား ၏ ဂုဏ်တော် ကျေးဇူးတော် သည် အဆုံးမရှိ ကျယ်ဝန်း ၏"

Inserting conjunction words is determined by predefined rules for suitable adding conjunction words.

4.3. Inserting conjunction words

After translating each segmented Pali words to Myanmar words, system needs to add conjunction words between Myanmar words to get smooth meaning of given Pali words. For this, system need to determine how to add conjunction is depend on the proposed building predefined set of the rules for adding conjunctions.

So the system proposed Pali segmentation and segmentation rules and adding conjunction rules.

4.4. Myanmar Text Database for Pali Words

The proposed system needs to build Pali-Myanmar dictionary which will use in translation step

of proposed system. It is also computerized Pali-Myanmar dictionary and new terms for Pali-Myanmar can easily add, edit and delete. Pali-Myanmar dictionary is used to translate each segmented Pali word is converted to Myanmar word. Pali-Myanmar dictionary needs maintain correct corpus of the Pali-Myanmar words. To improve accuracy of translation, the constructed Pali-Myanmar dictionary needs to keep the concise and accurate pair of pali and Myanmar words. The proposed system intends to provide validation for Pali words for preparing Pali-Myanmar dictionary. Moreover the proposed system will validate input pali words for spelling checking for wrong typing like "ကာ်" instead of "ကျ်". So Myanmar language spelling checking rules also defined for that validation process. The accuracy of the translation is also depending on the completeness of the Pali-Myanmar dictionary and N-gram based syllable combining method.

5. Conclusion

The proposed system attempts to be effective for studying of the Buddhist literature and the study of Buddhist scriptures in Myanmar. Moreover proposed system can dynamically add new terms and can search given Pali word in time and accurately. Segmentation of Pali words is based on predefined rules of Pali segmentation. There are two types of Pali words, such as Pali-Latin and Pali-Myanmar. In this paper, proposed system intends for Pali-Myanmar words translation system. Moreover encoding method for Pali-Myanmar text uses Myanmar Unicode system. The proposed framework for Pali-Myanmar translation system intends to automatic understanding of Pali-Myanmar text and automatic translation of Pali-Myanmar text to Myanmar text.

References

[1]. "A Toolkit for Natuaral Language Interface Construction"

[2]. Charles Duroiselle, "Practical Grammar of Pali Language", Third Edition 1997

[3]. Ei Phyu Phyu Soe, "Grapheme-to-Phoneme Conversion for Myanmar Language", the 11th International Conference on Computer Applications (ICCA 2013)

[4]. Hafiz Musa, Rabiah A.Kadir, Azreen Azman, M.Taufik Abdullah, "Syllabification Algorithm based on Syllable Rules Matching for Malay Language"

[5]. M.H.Bode, "The Pali Literature of Burma" The Royal Asiatic Society of Great Britain and Ireland", first published 1909, reprinted 1966

[6]. Martin Hosken, "Representing Myanmar in Unicode Details and Examples Version 3"

[7]. Phyu Hninn Myint, Tin Myat Htwe and Ni Lar Thein, "Bigram Part-of-Speech Tagger for Myanmar Language"

[8]. Thet Thet Zin, Khin Mar Soe, and Ni Lar Thein "Myanmar Phrases Translation Model with Morphological Analysis of Statistical Myanmar to English Translation System", University of Computer Studies, Yangon, Myanmar.

[9]. Win Win Thant, Tin Myat Htwe and Ni Lar Thein, "Comparative Study of Naïve Bayesian Classifier and Transformation-Based Learning for Myanmar Function Tagging"

[10]. Ye Kyaw Thu, "Integrating Dictionaries into an Unsupervised Model for Myanmar Word Segmentation"

[11]. Zin Maung Maung, Yoshiki Mikami, "A Rule-based Syllable Segmentation of Myanmar Text"

[12]. Zin Maung Maung, "Identification of Adopted Pali Words in Myanmar Text" , University of Computer Studies Mandalay, Myanmar, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012