Myanmar Word Stemming and Part-of-Speech Tagging using Rule Based Approach

Kyaw Htet Minn, Khin Mar Soe
University of Computer Studies, Yangon
{kyawhtetminn, khinmarsoe}@ucsy.edu.mm

Abstract

Myanmar language is spoken by more than 33 million people and use itas an official language of the Republic of the Union of Myanmar in bothverbal and written communication. With the rapid growth of digital content in Myanmar Language, applications like machine learning, translation and information retrieval become popular and it required to obtain the effective Natural Language Processing (NLP) studies.The main objective of this paper is to study Myanmar words morphology, to implement n-gram based word segmentation and to propose grammatical stemming rules and POS tagging rules for Myanmar language. So, this paper proposed the word segmentation, stemming and POS tagging based on n-gram method and rule-based stemming method that has the ability to cope the challenges of Myanmar NLP tasks. The proposed system not only generates the segmented words but also generates the stemmed words with POS tag by removing prefixes, infixes and suffixes. The proposed system provides 80% to 85 % accuracy. The data are collected from several online sources and the system is implemented using Python language.

Keywords: *Natural Language Processing, segmentation, n-gram, rule-based, stemming, POS tagging*

1. Introduction

With the rapid growth of the information technology, the availability of the research projects on language processing have been increased. Most of the developed countries' language have a various successful research in NLP. But for the developing countries' language such as Myanmar language still need to study and research.

The proposed research is evaluated on segmenting, stemming and POS tagging. Myanmar language are written from left to right. It is syllabic alphabet and written in circular shape. It has sentence boundary marker and follows the subject-object-verb (SOV) order. In particular, preposition adjunctions can appear in several different places of the sentence. Myanmar language users normally use space as they see fit, some write with no space at all. There is no fixed rule for word segmentation. Many researchers have been carried out Myanmar word segmentation in both supervised and unsupervised learning. We use Myanmar Word Segmentation with N-gram matching approach in the proposed system. Word stemming is another important feature supported by present day indexing and search systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. Part-of-speech (POS) tagging is the especially crucial for the disambiguation of a word. In this paper, we tackle the POS tagging with stemming in same framework by matching the clues word.

The rest of the paper is organized as follows: The Section (2) gives a brief overview of existing stemmers in literature. All the theories used in this thesis is discussed in Section (3) and the Section (4) discusses about our proposed system. The experimental result and performance of the proposed algorithm is analyzed in Section (5) and the paper is concluded in Section (6).

2. Related Work

Stemmer has an important role to improve the efficiency as well as performance of the IR systems. Lovin's Stemmer [7] was the very first stemmer, which was published in 1968. In this stemmer, 260 rules are defined. Dawson [6] employed another rule-based stemming method and covers a list of 1200 suffixes. Then in July 1980, Martin Porter [8] from the University of Cambridge developed the "Porter Stemmer", which is a rule based stemmer with five steps using rules up to 60. Later, few other stemmers were developed by Paice & Husk [2], Dawson and Krovetz [13]. Most of the above stemmers were rule based stemmers which followed the Suffix Stripping approach. Among these, the Porter Stemmer has proved to be an extremely useful resource to

researchers who work on stemmers also it has been applied to languages other than English.Rule based stemmer for Asian languages such as Japanese, Thailand, Malay, Indian and Nepali [1] are also found.

Natural Language Processing research in Myanmar Language started in the year (2006) at University of Computer Studies, Yangon (UCSY) with the release of the first English-Myanmar translation project. Corpus building and annotation for Myanmar, word segmentation, Text-To-Speech System, digitized Myanmar dictionary and many other researches also got started in the followed years [5]. These works motivated researcher to do further research on Natural Language Processing functions for Myanmar language. This proposed stemmer is one of the motivated work in Myanmar language which has both prefixes, infixes and suffixes stemming capabilities.

3. Background Theories

In this research, N-gram matching algorithm is used for segmentation. Stemming is executed by rule based stemmer and POS tagging is executed in the same framework of stemming.

N-gram

This N-gram approach is a string similarity approach that involve the system manipulating a measure of similarity between an input data and each of the definite words in the training corpus or predefined database. An n-gram is a set of 'n' successive characters extracted from a word. Typical values for n are two or three, these corresponding to the use of di-grams or trigrams, respectively. It is depend on the research work that how much the 'n' will be defined. For example, the word "COMPUTER", the results in the computation of the di-grams

C, CO, OM, MP, PU, UT, TE, ER, R

and the trigrams

C, *CO, COM, OMP, MPU, PUT, UTE, TER, ER*, R

Stemming

Stemming is the development of reducing distinct grammatical forms or word forms of a word such as its noun, verb, adjective, adverb etc. to its original or minimum root form. For example, in English language, the stem of "works" is "work" since -s is an inflectional suffix. The algorithm or model that used for stemming is called stemmer.

There are several types of stemming algorithmssuch as brute force algorithms which is finding a matching inflection through lexicon, suffix stripping algorithms whichtypically based on set of rules to strip the suffix of the word, lemmatization algorithms which is processing with determining the part of speech category, stochastic algorithms which trained the inflected word with root word and construct the probabilistic internal rule set to stem, affix stemmers which is similar to suffix but work on both prefix, suffix and even infix, and lastly the hybrid approach which used two or more above methods to build a proper stemmer[4].

POS Tagging

Part-of-Speech (POS) Tagging is the process of assigning the words with their categories that best suits the definition of the word as well as the context of the sentence in which it is used. Tagged keyword definition and numbers of tag-set may different from one another. All of the POS tagging algorithms fall into two distinctive groups: rule-based POS Taggers and stochastic POS Taggers.Typical rule-based approaches use contextual information to assign tags to ambiguous words. Disambiguation is done by analyzing the linguistic features of the word, its preceding word, its following word, and other aspects. The term stochastic tagger can refer to any number of different approaches to the problem of POS tagging. Any model which somehow incorporates frequency or probability may be properly labelled stochastic.

4. Proposed System

In this proposed system, there has five steps in total. Starting from normalization, syllabication and segmentation as preprocessing steps and followed by stemming and POS tagging respectively. In segmentation, N-gram matching algorithm is used and stemming is executed by rule based stemmer. The following figure is the overview of the proposed system.

For describing the proposed system, three types of corpusare predefined into consideration: prefix, infix and suffix.All of the corpusis collected from Myanmar Grammar Book published by Myanmar Language Commission[11].For N-gram segmentation,176 prefixesand 266 suffixesare predefined.For stemming and POS tagging, 196 prefixes, 30 infixesand 259 suffixes, totally 485 wordsare used in the proposed algorithm. The words

contain in this corpus have various number of ranging from one syllable to five syllables.

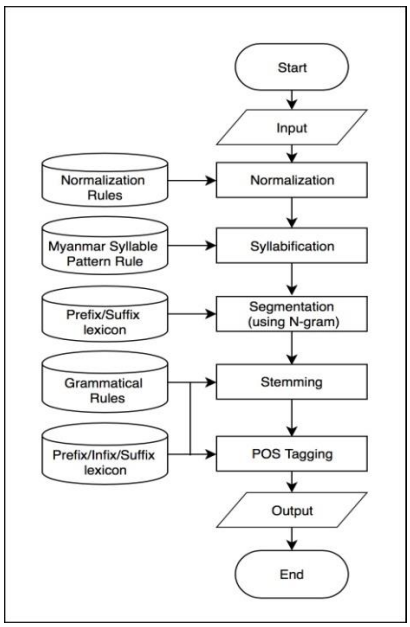

Figure 1. Proposed System Design

4.1 Normalization

The very first step to the system is to polish the input sentence. By normalizing the words to a standard format, the matching process and searching process will simplify. Normalization in Myanmar word is applied because the input may contain the converted text from different encoding and typing error.

In this study, 40 rules are using for normalizing character reordering. One of the rules is explained as an example.

E.g. In **"ဖြင့်"** word, correct from

"ဖ+ြ+c+ွ+ံ" to "ဖ+ြ+c+ံ+ွ"

4.2 Syllabification

After the normalization, the next step is syllabication. Syllabication is the task of breaking a sequence of words into syllables. Syllabication has no completely agreed definition of syllable boundaries. In this study, the following syllable structure of Burmese is used, C(G)V((V)C).

CV / သူC=သ, V=ူ

CVC / မိန်းC=မ, V=ိ, C=န်း

CGV / မြေ C=မ, G=ြ, V=ေ

CGVC / မြိန်C=မ, G=ြ, V=ိ, C=န်

CVVC / မောင် C=မ, V=ေ, V=ာ, C=င်

CGVVC /မြောင်းC=မ, G=ြ, V=ေ, V=ာ, C=င်း

The output gives syllables with separated space character.

Input = သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။ (no space contained)

Output = သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။ (space is added within each syllable)

4.3 Segmentation

In this step, five-grams is used for matching the prefix/suffix and output the segmented sentence. By matching with the predefined prefix and suffix lexicons, the input sentence is segmented with the correct lexicon and result as a space separated sentence.

Input =သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။ (unsegmented sentence)

Five-grams set = ["သူသည်ပင်ပန်းသော", "သည်ပင်ပန်းသောကြောင့်", "ပင်ပန်းသောကြောင့်ကျောင်း", "ပန်းသောကြောင့်ကျောင်းကို", "သောကြောင့်ကျောင်းကိုကား" "ကြောင့်ကျောင်းကိုကားဖြင့်", "ကျောင်းကိုကားဖြင့်သွား", "ကိုကားဖြင့်သွားသည်", "ကားဖြင့်သွားသည်။"]

Output = သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။

The following figure shows how the segmentation process executed using N-gram.

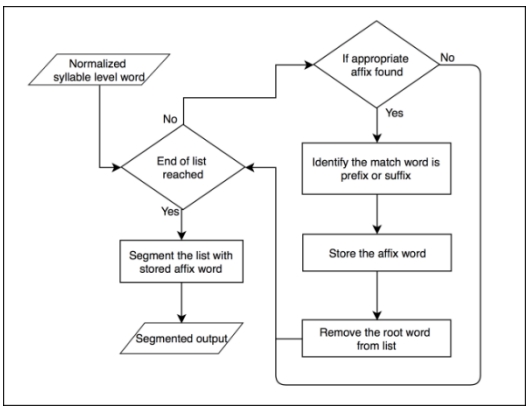

Figure 2. Architecture of Segmentation using N-gram

4.4 Stemming

After processing all the above steps, the segmented words are ready to stem. The proposed stemmer was developed using three corpus prefix, infix and suffix. In this step, 8 kinds of grammar rules and 8 kinds of additional rules are applying for stemming. The eight kinds of grammar rule contain rule for conjunction, postpositional marker, particle, interjection, adjective, adverb, verb and pronoun. The list of additional rules is covered for symbols, gender, counter, number, number units, months, day and negative word.

Grammar rules contain 37 rules for of Prefix/Infix/Suffix stemming. This stripping is done by matching the word with the input phrase. Additional rules are constructed by 12 rules. All the rules are using 485 predefined prefix/infix/suffix corpus for respectively. These segmented words are applied with the above rules sequentially. If the input word contains the word from corpus list, then it is considered as a root word. This process is continuing until it finds the root word, its prefix/suffix or till the word is stripped into minimum word length that is one. The algorithm does affix striping the word according to rule and tag with the rule indicator.

Input =
သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။

Output = ['သူ', 'ပင်ပန်း', 'ကျောင်း', 'ကား', 'သွား']

The word
'သည်","သောကြောင့်","ကို","ဖြင့်","သည်","။"are

removed as they are prefix and suffix to the stemmed word

The figure below indicates how the stemming and POS tagging work.

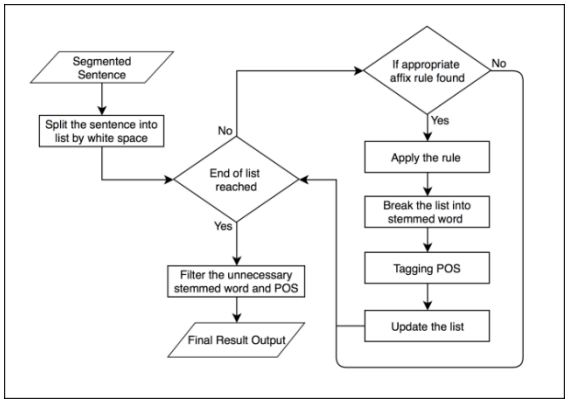

Figure 3. Architecture of Stemming and POS Tagging

4.5 POS Tagging

In this stage, 14 kinds of tag-sets are tagged according to the rule of stemming for POS tagging. They are Nouns (NN), Pronouns (PRON), Verbs (VB), Adverbs (ADV), Adjectives (ADJ), Conjunctions (CONJ), Particles (PART), Postpositional Markers (PPM), Interjections (INTJ), Symbols (SB), Numbers (NUM), Text Number (TN), Punctuation (PUNC) and Foreign Word (FW). Addition to the above tag-sets, there is also an ambiguous POS corpus. The POS tagging is working under the same framework with stemming algorithm. In this system, not only the stemmed word but also the stripped word is tagged and stored in a separated file. The stemmed word only is defined as output (1) and the whole sentences which include stripped word is described as output (2).

Input =
သူသည်ပင်ပန်းသောကြောင့်ကျောင်းကိုကားဖြင့်သွားသည်။

Output 1 = ['သူ-PRON', 'ပင်ပန်း-VB', 'ကျောင်း-NN', 'ကား-NN', 'သွား-VB']

Output 2 = ['သူ-PRON', 'သည်-PRON/PPM', 'ပင်ပန်း-VB', 'သောကြောင့်-CONJ', 'ကျောင်း-NN', 'ကို-PPM', 'ကား-NN', 'ဖြင့်-PPM', 'သွား-VB', 'သည်-PPM', '။-PUNC']

5. Experimental Evaluation

In this section, data used for experiment and achieved results with precision, recall and F-measure have been presented.

5.1 Data used for Experiment

The experiment is conducted using data collected 100 sentences each from Myanmar Wikipedia website, Myanmar Computer Federation (MCF) dataset and Myanmar Grammar Book published by Myanmar Language Commission. Therefore, testing 300 sentences and an average number of sentences per document are 10 so that the total words for input is over 3,000. These testing data was evaluated based on different domains.

5.2 Performance Measures and Result

In this paper, the evaluation metrics for the data set is precision, recall and F-measure. These are defined as following: -

Precision (P) = Number of words correctly stemmed by the system / Total number of words

Recall (R) = Number of words stemmed by the system / Total number of words

F-Measure (F1) = (β^2 +1) PR/ (β^2 R+P)

Where β is the weighting between precision and recall andtypically β = 1.

Table 1. Evaluation Results

Category	P(%)	R(%)	F1(%)
Segmentation	83.67	88.80	86.16
Stemming	80.59	82.35	81.46
POS Tagging	73.78	83.95	78.54
Average	79.34	85.03	82.05

6. Conclusion

In this study, n-gram is used to segment the input sentence and rule-based stemmer is applying to get the root word of Myanmar phrases and generate the POS tagged.The output of this study can be used as further study research on text classification, categorization, information retrieval, machine learning, text mining and etc.However, no language in the world strictly follows a deterministic set of rules, so it is difficult to achieve this purpose systematically. That is why a perfect stemmer, able to accurately obtain the stems of any term independently of its features, does not exist. This paper algorithms are limited to work for words containing conjugated letters, only verb and noun inflections; the inflections for other parts of speeches are not considered. Advantages of this algorithms are following: algorithmic stemmer is fast, require only low storage and processor power, performance can improve by adding rules or lexicons and can add lexicons and rules at any time. However there still have some limitations to the current research. They are compound word cannot be stemmed, additional rules can be conflict with each other and require complex grammar rules. The overall accuracy of the current proposed system is over 80% but, the future work can be done by the following areas.

i. The accuracy can improve by adding lexicons in all three types.
ii. It can be compared with other rule based algorithms and also with other types of stemmer.
iii. It can use as the rough data for specific POS tagging research.

Acknowledgements

Thank to my supervisor of this thesis, Dr. Khin Mar Soe from University of Computer Studies, Yangon (UCSY), for her support and inspiration throughout the duration of the work. Finally, thanks go to all the teachers, staff and friends who support and generous help to me all along the time.

References

[1] A. Paul, A. Dey, An Affix Removal Stemmer for Natural Language Text in Nepali, International Journal of Computer Applications, Volume 91, 2014

[2] C.D. Paice, "An evaluation method for stemming algorithms", In the Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1990, pp. 42 – 50.

[3] Chen Din, Word Segmentation for Burmese (Myanmar), 2016

[4] EimanTamah, Towards An Error- Free Stemming, 2008

[5] H.H. Htay, K. N. Murthy, Myanmar Word Segmentation using Syllable Level Longest Matching, Proceeding of the 6th Workshop on Asian Language Resources, 2008.

[6] J. Dawson, "Suffix removal and word conflation", LLCbulletin, 2(3), 1974, pp. 33– 46.

[7] J.B.Lovins, "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics 11, 1968, pp. 22-31.

[8] M.F. Porter, "An algorithm for suffix stripping", Program, 14(3) 1980, pp. 130−137.

[9] Michael F. Lynch, and Peter Willett, Stemming and N-gram matching for term conflation in Turkish texts, 1996

[10] M. Ali, S. Khalid, M. H. Saleemi, A Rule based Stemming Method for Multilingual Urdu Text (IJCA, Volume 134 – No.8, January 2016)

[11] Myanmar Grammar, Myanmar Sar Myanmar Sakar, Department of Myanmar Language commission, Ministry of education, Union of Myanmar June 2016.

[12] Paice Chris D. "An evaluation method for stemming algorithms". Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. 1994, 42- 50.

[13] . Krovetz, "Viewing morphology as an inference process", In Proceedings of the 16 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, pp. 191-202.

[14] Rohit Kansal, Rule Based Urdu Stemmer, 2012

[15] Savoy, Stemming of French words based on grammatical categories, 1993

[16] Wahiba Ben, A new stemmer to improve information retrieval, 2013.