Joint Word Segmentation and Stemming for Myanmar Language Based on Conditional Random Fields

Yadanar Oo, Khin Mar Soe
University of Computer Studies, Yangon, Myanmar
yadanaroo@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm

Abstract

In this paper, we describe a joint work on word segmentation and stemming of Myanmar sentences with syllabled-based tagging under Conditional Random Fields(CRF) framework. A manually-segmented corpus was developed to train the segmenter, and we implement it as a 7-tag syllable-based tagging and stemming with conditional random fields(CRF). And then, the trained CRF segmenter was compared to a baseline approached based on longest matching that used a dictionary extracted from manually segmented corpus. In our approach, we can achieve comparative performances compared to 4-tag syllable tagging approach. The experimental results show that the CRF with 7-tag set and word feature improve the stemming performance.

Keywords*: segmentation, stemming, syllable tagging, conditional random fields*.

1. Introduction

With the enormous amount of data available online, it is very essential to retrieve accurate data from user query. Stemming has been extensively used to increase the performance of Information Retrieval System. In Linguistic morphology, stemming is the process of reducing inflected words to theirs root form. Stemming is one of the parts of NLP. Myanmar written language is one of the languages that does not have word boundaries. In order to discover the meaning of the document, all texts must be separate into syllables, words, sentences and paragraphs. Word segmentation is the process of determining word boundaries in a piece of text. In English language, word boundaries are easily determined because of the presence of white spaces or punctuation between words. In Myanmar Language, segmenting sentences into words is a challenging task because sentences are clearly delimited by a sentence end marker, but words are not always delimited by spaces. Spaces may sometimes be inserted between words and even between a root word and the associated post-position. It is because there are no indicators such as blank spaces to show the word boundaries in Myanmar text. The same phenomenon does not happen only to Myanmar language but also many other Asia languages such as Japanese, Chinese and Thai. Therefore, to find the root word in the Myanmar text, the first thing that we need to do is to cut the sentences into word segments. Although it sounds easy to cut a sentence into a word sequence, however, from the past experience, we know that it is not a trivial task.

In this paper, word segmentation for Myanmar language was proposed with syllabled based tagging problem under Conditional Random Fields(CRF) framework that also find the root word in the sentence at the same time. This approach mainly aims to detect the syllable type in a certain sentence, e.g., root syllable, simple syllable, prefix syllable or suffix syllable. It achieves slightly better performances than traditional longest matching approach in the closed test. In order to increase the accuracy of overall segmentation and stemming process, different tagsets (using 4-tags and 7-tags) are tested. This implementation achieves the better performance in 7-tags but need much more training time and memory space.

In the remaining part of the paper, describe the related works of words segmentation (Section 2). Then some important concept of conditional random fields (Section 3). Proposed system architecture is described (Section 4). And then, describe how to obtain stem word and word segmentation using CRF approach (Section 5). Finally, experimental results were discussed (Section 6) and give conclusions with future direction (Section 7).

2. Related Work

In this section, we briefly review related works on Myanmar word segmentation. Hla Hla Htay, Kavi Narayana Murthy, 2008, "Myanmar Word Segmentation using Syllable Level Longest Matching "[4] introduced Myanmar word segmentation. Firstly, they have collected 4550

syllables from available sources of 2,728 sentences and build the words lists from available sources including dictionaries and by generating syllable n-grams as possible words, a total of 800000 words. Secondly, word segmentation is carried out with the longest syllable word matching using their 800000 strong stored word list.

In [7] Tun Thura Thet; Jin-Cheon Na, Wunna Ko Ko, October 2007, "Word segmentation for the Myanmar language ". The proposed strategy was in two parts: rule-based syllable segmentation and dictionary-based statistical syllable merging. First, input texts in Unicode character codes were scanned and segmented as syllables using a rule-based algorithm. Six syllable segmentation rules were applied in the algorithm. The next step adopted a dictionary-based statistical approach for syllable merging, using dictionaries and the collocation strength of a sentence or phrase.

In [8] W.P.Pa, N.L.Thein, February 2008 , "Myanmar Word Segmentation Using Hybrid Approach" Word Segmentation system consists of four components, sentence splitting, tokenization, initial segmentation by Maximum Matching Algorithm and statistical combined model (bigram model and modified word juncture model) for final segmentation.

In [9], Ye Kyaw Thu, Win Pa Pa, Andrew Finch, "Word Boundary Identification for Myanmar Text Using Conditional Random Fields". Conditional random field is used to identify Myanmar word boundaries within a supervised framework. CRF approach is compared against a baseline based on maximum matching using dictionary from Myanmar Language Commission Dictionary (word only) and manually segmented subset of the BTEC1 corpus.

In [10], Upendra Mishra, Chandra Parkash, “MAULIK: An Effective Stemmer for Hindi Language”. The proposed strategy uses the Hybrid approach (Combination of brute force and suffix removal approach). It achieves the accuracy of 91.59%. This stemmer reduces the problem of under stemming and over stemming. In this paper, we deeply analyze word segmentation of syllable tagging on conditional random field approach and propose a new implementation of stemming approach at the same time.

3. Conditional Random Fields

The main title Conditional Random Fields are undirected graphical models trained to maximize a conditional probability of the whole graph structure. A common case of a graph structure is a linear chain, which corresponds to a finite state machine, and is suitable for sequence labeling. A linear-chain CRF with parameter Λ = {λ1, ..., λK} defines a conditional probability for a label sequence y = y1...yT given an input words sequence x = x1...xT to be:

$$P_\lambda(y|x) = \frac{1}{Z_x} \exp(\sum_{t=1}^{T} \sum_{k} \lambda_k \, f_k(y_{t-1}, y_t, x))$$

where Zx is the normalization factor that makes the probability of all state sequences sum to one; fk(yt-1, yt, x) is a feature function, and λk is a learned weight associated with feature fk. The feature function measures any aspect of a state transition, yt-1, yt, and the entire observation sequence, x. Large positive values for λk indicate a preference for an event, and large negative values make the event unlikely.

The most probable label sequence for an input x,

$$y^* = argmax_y \; P_\lambda(y|x)$$

can be efficiently determined using the Viterbi algorithm. CRFs are trained using maximum likelihood estimation, i.e., maximizing the log-likelihood LΛ of a given training set

$$T = \langle x_i \, , y_i \rangle_{i=1}^{N}$$

$$L_\lambda = \sum_i \log P_\lambda \, (y_i|x_i)$$

$$= \sum_i (\sum_{t=1}^{T} \sum_k \lambda_k \;\; f_k(y_{t-1} \, , y_t \, , x \,) - Log \; Z_{x_i}$$

CRFs are discriminative models and can capture many correlated features of the inputs. Therefore, it is suitable in many tasks in NLP for sequence labeling. Since they are discriminatively-trained, they are often more accurate than the generative models, even with the same features. CRF++ is a customizable implementation of linear-chain CRFs for labeling sequential data.

Another attractive aspect of CRFs is that one can implement efficient feature selection and feature induction algorithm for them. That is, rather than specifying in advance which features of (X, Y) to use, we could start from feature-generating rules and evaluate the benefit of generated features automatically on data.

Conditional random fields offer a unique combination of properties: discriminatively trained models for sequence segmentation and labeling; combination of arbitrary, overlapping and agglomerative observation features from both the

past and future; efficient training and decoding based on dynamic programming; and parameter estimation guaranteed to find the global optimum.

4. Proposed System Architecture

The enormous amount of data available online, it is very essential to retrieve accurate data for some user query. Stemming has been extensively used in various Information Retrieval System to increase the retrieval accuracy. Stemming is a method that reduce morphology similar variant of word into a single term called stems or roots without doing complete morphological analysis. In English, relating a word like "children" to its root "child" is an obvious necessity. The importance of morphology, however, is even greater in a language like Myanmar, Chinese, Japanese and Korean. In fact, Asian text is written with limited or no space separations. The task of segmenting the initial text into a sequence of words is strongly related to the stemming process. In this paper, word segmentation for Myanmar language was proposed with syllabled based tagging under Conditional Random Fields(CRF) framework that also find the root word in the sentence at the same time. The proposed system architecture for CRF-based Myanmar Word Segmentation and Stemming is shown in Figure 1.

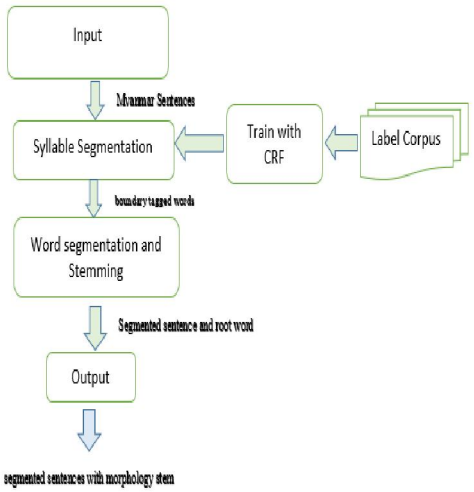

Figure 1. Overview of the Proposed System

5. Myanmar Word Segmentation Framework

Segmentation is a process to divide a sentence into meaningful word and is also the process of taking a sequence of syllables and producing meaningful word units. In Myanmar language, a "word" is difficult to define, as it does not explicit word boundaries. In order to produce the meaningful words, word segmentation task has to be done as a preprocessing stage of stemming.

In this paper, we firstly have focused on syllable based boundary tagging and proposed approach for stemming at the same time.

5.1. Sentence Segmentation

In Myanmar language, there is no white space between words, but sentences are delimited by sentence end marker called ***pote-ma*** "။". Firstly, we separate the sentences by using sentences end marker.

5.2. Syllable Segmentation

Syllable is a basic sound unit. A word can be consisted of one or more syllables. Every syllable boundary can also be word boundary. In some a word can include other words, it is called a compound word. For syllable segmentation, this system uses the algorithm of [8]. Example of syllable segmentation is shown in Figure 2.

လေ့လာတွေ့ရှိရချက်များ အရ ယခု နှစ်သည် အပူဆုံးနှစ် ဖြစ်ပြီး ၎င်းသည် သဘာဝပတ်ဝန်းကျင်ဆိုင်ရာ စိန်ခေါ်မှုကြီး ဖြစ်သည်။

လေ့ လာ တွေ့ ရှိ ချက် များ အ ရ ယ ခု နှစ် သည် အ ပူ ဆုံး နှစ် ဖြစ် ပြီး ၎င်း သည် သ ဘာ ဝ ပတ် ဝန်း ကျင် ဆိုင် ရာ စိန် ခေါ် မှု ကြီး ဖြစ် သည်။

Figure 2. An example of syllable segmentation for Myanmar sentence

5.3. Syllable Boundary Tagging in CRF

The word segmentation process starts by tagging a word boundary on initial syllable segmentation. The segmentation task is to classify each syllable with a tag of 'B' or not, which represent a word boundary appears after the syllable or not. There are several classification algorithms which can be applied to do the segmentation, such as maximum entropy, perceptron algorithm and conditional random field (CRF). In this system, we applied CRF approach for syllable boundary tagging.

The sentence is first segmented into syllable. Then, from the output, syllable boundary tagging is used to classify the word type and detect the boundary of words. This process uses manually tagged corpus to train the boundary of the syllables with conditional random field (CRF) learning approach. In this paper, syllable is classified with four classes of word types. There are Root word,

Simple word, Prefix and Suffix. All the type of words also has their boundary marker (“B”) except from prefix. Figure 3 shows the example of boundary tagging in word segmentation.

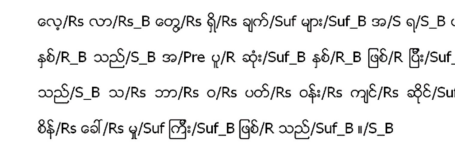

Figure 3. An Example of Boundary Tagging for Word Segmentation and Stemming

In Figure, each syllable is tagged with ‘Rs’ (for example လေ့/Rs) represents the sub syllable of root word. Syllable tagged with ‘Rs_B’ (for example လာ/Rs_B) represents the last syllable of root word that is also boundary of word. Therefore, one-word type contains one or more syllables. For example: root word သဘာဝပတ်ဝန်းကျင် contains six segmented sub syllables which are သ/Rs ဘာ/Rs ဝ/Rs ပတ်/Rs ဝန်း/Rs ကျင်/Rs .

In the boundary tagging process, the local context features, w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2} are used.

5.4. Morphological Stemming

The basic order of the Myanmar languages is subject-object-verb. There are nine Part-of-Speech classes for all Myanmar words. These are Noun, Pronoun, Verb, Adjective, Adverb, Conjunction, Postpositional Marker, particles and Interjection.

Noun is the content word that can be used to refer a person, place, thing. Noun is the root word in a sentence. But Noun in Myanmar language can be combined with particles to form plural by suffixing the particle “ တွေ” [-twei] or “များ “[-myar] e.g., ကျောင်းသားများ is the plural form of Noun Student.

Moreover, the roots of the Verbs are always suffixed with at least on particle to form a tense, politeness, mood, etc. The root of the Verbs remains unchanged when they have the particle suffix to them. For instance, စားသည် [-sar the]; စား၏ [-sar ei]; စားခဲ့သည် [-sar kae the] have different verb particles. But they have the same root verb is စား [-sar]. And, Verbs are negated by the particle “ မ “[-ma], which is prefix to the verbs to form the negative verb and which also unchanged the root of verb.

And then, Reduplication occur in Myanmar sentences and most of the reduplicated words are Adverbs and their root form are Adjective. Many Myanmar words, especially adjectives or verbs with two syllables, such as “ လှပ ” (pretty)[adjective] or “ ကျန်းမာ ” (healthy)[verb], can be reduplicated as “ လှလှပပ ,,

(pretty) or “ ကျန်းကျန်းမာမာ ” (healthily). Some of the Adjective has suffix “ သော “[-thw] and Adverb also has suffix “ စွာ “[-swar]. Their root form remains unchanged when suffix removal.

The other type of words, such as Conjunction, Postpositional Marker, Interjection and Particles are not root word. Our goal in this paper is to detect root word and their boundary correctly. Example of morphological stemming and segmentation output result is shown in figure 4 and figure 5.

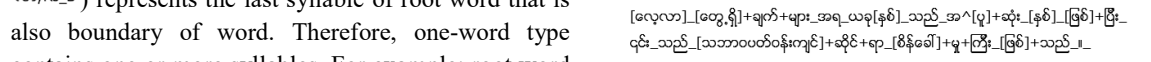

Figure 4. An Example of Morphological Stemming Result

In figure, the root word is placed within the boundary marker [], and suffix words are marked by + and then prefix are delimited by ^ marker. Moreover, spaces between words are separated by _ marker.

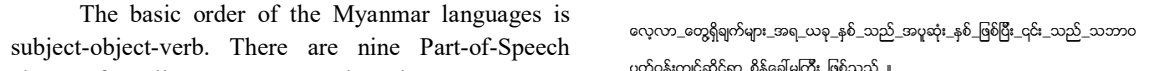

Figure 5. An Example of Word Segmentation Result

6. Experiments

In this section, we evaluate the result of CRF segmentation model by comparing against a baseline model based on longest matching. A detail evaluation is presented in the following.

6.1 Data Setup

The CRF model was trained using a training set selected from manually annotated 5000-sentences. In order to train the CRF segmentation models, we have to build a corpus of newspaper articles from many sites and in various domains. Although the corpus is not enough to cover a broad range of Myanmar words, it contains many sentences from multiple domains.

The syllable is annotated by one kind of labels, such as "Rs" (Sub syllable of Root word), "Rs_B" (Boundary of sub syllable Root word), “Pre” (Prefix), “Suf” (Suffix), “Suf_B” (Boundary of suffix word), “S” (Simple), “S_B” (Boundary of simple word).

6.2 Training with CRFs

We used the CRF++ toolkit to train the CRF models. Before CRF model training, we must transfer the document into the tagging sequences. And then, we train a CRF model that can label the syllable type.

The tagged data are used to train the CRF model in advance. In the CRF++, the output is a CRF model file. The feature set used in the model was as follow.

Table 1. Features Template of CRF

Feature	Description
w_0	current word
w_{-1}, w_0, w_{+1}	previous word, current word, next word
w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}	two previous word, previous word, current word, next word, two next word

6.3 Evaluation

The performance of CRF models on different data set was measured the commonly used precision(P), recall(R), and F1-Measure.

$$\textbf{Precision} = \frac{Number\ of\ correct\ words}{Number\ of\ words\ in\ the\ test\ corpus}$$

$$\textbf{Recall} = \frac{Number\ of\ correct\ words}{Number\ of\ words\ in\ system\ output}$$

$$\textbf{F-score} = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

6.4 Result and Discussion

In the experiment, we compare agains t a baseline model based on longest matching (LM) [4] and our approach, CRF model. Longest matching approach have tested on 5000 sentences including a total of (35049 words) and then all the sentences are from English-Myanmar parallel corpus. In our approach, we have tested on 1000 sentences including a total of (18601 words) including (49683 syllables) and all the sentences are from Myanmar New website (Thit Htoo Lwin, 7Days News, Eleven New Journal). Table 2 shows the results of the experiments.

Table 2. Result based on LM and CRF Model

	Precision	Recall	F1-Measure
LM	99.11	98.81	98.95
CRF	99.20	98.70	98.96

The result shows the accuracy of the Longest matching and CRF is not very different. In the result, CRF is slightly better than longest matching approach. But CRF model has competitive advantages because while longest matching relies heavily on words listed in the dictionary and it always prefers compound words over simple words, CRF only relies on syllables.

To observe the morphological information can assist the performance of the stemming and segmentation, we compare the 4-tags set and 7-tags set. In 4-tags set, it contains only Root word and Simple word. Table 3 shows the definition of the 4-tags set.

Table 3. Definition of 4-tags Set

Tags	Definition
Rs	Sub syllable of Root word
Rs_B	Boundary of Sub syllable Root word
S	Simple word
S_B	Boundary of Simple word

Another corpus is 7-tags corpus. It defined the type of the word morphologically. Table 4 shows the definition of the 7-tags set.

Table 4. Definition of 7-tags Set

Tags	Definition
R	Sub syllable of Root word
R_B	Boundary of Sub syllable Root word
S	Simple word
S_B	Boundary of Simple word
Pre	Prefix
Suf	Suffix
Suf_B	Boundary of Suffix word

We analyzed the comparison of Precision and Recall between 4- and 7-tags on two types of testing, open and closed test. Experimental results are shown in Table 5.

Table 5. Result based on Different Tag sets with CRF Model

	No of tags	Precision	Recall	F1-Measure
Closed Test	4-tags	97.38	97.78	97.58
	7-tags	99.20	98.70	98.96
Open Test	4-tags	84	87	85.47
	7-tags	94	96	95

In the experimental results, the performance is not very different in Closed test. In the Open test, the performance is increasing from 4-tag to 7-tag set. It can be observed that morphological

information gives a performance improvement when it works with 7-tag set. On the contrary, 4-tag set will not give better performance because it can only give Root word and Simple word.

7. Conclusion and Future Work

This paper has described our initial effort to deal with Myanmar word segmentation based on the concept of conditional random fields. We present a framework for performing word segmentation and morphological Stemming. In the experiment we compare our model with Longest matching approach. And then, we analyze the differences between 4-tags and 7-tags corpus in open test and closed test. We have shown the 7-tag set can give a substantial performance improvement in both open test and closed test. Our approach produces not only root word but also detect the boundary of the word which is basic requirement for Myanmar Language.

In the future work, we intend to increase the size of manually segmented corpus in order to improve the performance of segmentation. And then, in the current work, we use only basic features for learning and predicting. However, CRFs allow using arbitrary, overlapping and non-independent features. So, we will use more features to increase the accuracy of word segmentation. Moreover, we will detect the more words especially names, cities name and proper nouns, which further improve the performance.

References

[1] C.Kruenghkrai, V.Sornlertlamvanich and H.Isahara, " A Conditional Random Field for Thai Morphological Analysis" Journal, Thailand.C.T Nguyen, T.K Nguyen and X.H Pham, " Vietnamese Word Segmentation with CRFs and SVMs: An Investigation ", The 4th International Conference on Computer Science , Vietnam National University, Hanoi, 2006.

[2] Goh.C.L, Y.Matsumoto, and K.Shikano, " Unknown Word Identification for Chinese Morphological Analysis", Nara Institute of Science and Technology, 29th September 2006.

[3] H.H.Htay and K.N.Murthy, " Myanmar Word Segmentation using syllable level longest matching", IJCNLP, 2008, pp.41-48.

[4] Lafferty,J. McCallum, A., Pereira,F.: "Conditional Random Fields: Probabilistic Models for Segmentation and labeling Sequence Data." In: Proceedings of the International Conference on Machine Learning, 2001, pp. 282-289.

[5] Myint.P.H and T.M.Htwe, " Bigram Part-of-Speech Tagger for Myanmar Language", University of Computer Studies, Yangon, 2011.

[6] T.T.Thet, J.C Na and W.K.Ko, " Word Segmentation for Myanmar Language", Jourrnal of Information Science, Nanyan Technological University, Singapore, 2008, pp.688-704.

[7] Thu, Y.K., Finch, A., Sagisaka, Y., Sumita, E.: A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation. In Proceedings of 12th International Conference on Computer Applications, Yangon, Myanmar, pp.167-179(2014).

[8] Pa. W.P., N.L.: "Myanmar Word Segmentation using Hybrid Approach.", ICCA, Yangon, 2008, pp.166-170.

[9] W.P.Pa, Y.K.Thu, A.Finch and E.Sumita, "Word Boundary Identification for Myanmar Text Using Conditional Random Field", Springer, Switzerland, 2016.