Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language

Dim Lam Cing , Khin Mar Soe
University of Computer Studies, Yangon, Myanmar
dimlamcing@ucsy.edu.mm , khinmarsoe@ucsy.edu.mm

Abstract

In natural language processing (NLP), Word segmentation and Part-of-Speech (POS) tagging are fundamental tasks. The POS information is also necessary in NLP- based applications such as machine translation (MT), information retrieval (IR), etc. Currently, there are many research efforts in word segmentation and POS tagging developed separately with various approaches to reach high performance and accuracy. For Myanmar Language, there are also separate word segmentors and POS taggers based on statistical approaches such as Neural Network (NN) and Hidden Markov Models (HMMs). However, the Myanmar language has the complex morphological structure and the Out-of-Vocabulary (OOV) problem still exists. Thus, this paper proposed morphological analysis based joint Myanmar word segmentation and POS tagging using Hidden Markov Models (HMM) and morphological rules. This paper has also presented the comparison of accuracy result using HMM only, and HMM with morphological analysis.

1. Introduction

Word segmentation and Part-of-speech (POS) tagging are a fundamental process of natural language processing application such as machine translation, information extraction, speech recognition, grammar checking and word sense disambiguation, etc. There are many methods for development of POS taggers. Rule based, statistical based and neural network based are the most using techniques. In the rule-based approach, rule is developed by linguistic to define precisely how and where to assign the various POS tags. This approach has already been used to develop the POS tagger for Myanmar Language. In the statistical approach, statistical language models are built, refined and used to POS tag the input text automatically. Most commonly used statistical approaches are Hidden Markov Models based approach, Support vector machine based, Conditional Random Field based and Maximum Entropy based approach [1].

This paper describes Hidden Markov Models (HMM) and the proposed system for word segmentation and part-of-speech tagging for Myanmar language. Myanmar Language is morphologically rich, complex, and agglutinative in nature, words of which are inflected with many grammatical features. POS tagging is an important problem in the field of NLP and one of the basic processing steps for any language in NLP i.e., the capability of a computer to automatically POS tag a given sentence. Morphological analysis is an essential component in language engineering applications especially for morphologically rich and complex language like Myanmar. Performing a full morphological analysis of a wordform is usually regarded as a segmentation of the word into morphemes and gives basic insight to the natural language by studying how to distinguish and generate grammatical forms of words [4].

Normally, an early step of processing is to divide the input text into units called tokens where each is either a word or something else like a number. The main clue used in space-delimited language like English is the white space. In major East-Asian languages such as Chinese, Japanese, Thai and Myanmar, there is no spaces between words. Myanmar language, its writing style does not use any delimiter between words.

There has been very few researchs conducted on various language processing tasks including morphological analysis for Myanmar language compare to English, France, Chinese, India, and Thai., etc. Since high level language processing tasks such as POS tagging, machine translation, semantic analysis, syntactic analysis, sentiment analysis, information retrieval, classification, clustering system, etc. all process on smallest language unit; words. The morphology of the language through a systematic linguistic study is important in order to

reveal words that are significant to users such as historians, linguists, etc.

This paper is organized as follows: Section 2 discussed related work. Section 3 described about Myanmar Language and in Section 4, theory of Hidden Markov Models (HMMs). The proposed system has been presented in Section 5. Section 6 provides the evaluation and in Section 7 has concluded with future work.

2. Related Work

Bigram Part-of-Speech Tagger for Myanmar Language in [5] is used supervised learning approach for Myanmar Language. For disambiguating POS tags, HMM model with Baum-Welch algorithm is used for training and Viterbi algorithm is used for decoding. Myanmar lexicon is used for tagging a word with its all possible tags. Experimental results show that the approach achieves high accuracy (over 90%) for different testing input.

Analysis of Myanmar Word Boundary and Segmentation by using Statistical Approach proposed a unified approach for Myanmar Word analysis using Finite State Automata (FSA), Rule Based Heuristic Approach and Statistical Approach. The Rule Based Heuristic Approach and Statistical Approach are used with corpus based dictionary. Evaluation results showed that the method is very effective for the Myanmar language [1].

Myanmar Word Segmentation using Hybrid Approach proposed a hybrid approach that works by longest matching on syllable segmented sentences. By using Longest matching method, the known words from a dictionary are first segmented and an n-gram model predicts the segmentation of the unknown words. The principal problem of this approach stems from the ambiguity in the longest matching process, since words can be formed in more than one way [6].

Y. Zhang and S. Clark proposed a joint Chinese word segmentation and POS tagging model, which achieved a considerable reduction in error rate compared to a baseline two stage system. A challenge for this joint approach is the large combined search space, which makes efficient decoding very hard. They used a single linear model for combined word segmentation and POS tagging, and chose the generalized perceptron algorithm for joint training and beam search for efficient decoding. The joint model gives an error reduction in segmentation accuracy of 14.6% and an error reduction in tagging accuracy of 12.2%, compared to the traditional pipeline approach [7].

H. Fadaei and M. Shamsfard presented a POS tagger for Persian. They exploited a hybrid approach which is a combination of statistical and rule-based methods to tag Persian sentences. The proposed tagger uses a novel probabilistic morphological analysis to tag unknown words. As a secondary result of this research a knowledge base of Persian morphological rules with their probabilities is built according to a corpus. Experimental results show that their method improves the tagging performance and accuracy [8].

A two-stage discriminative approach based on CRFs for a Korean morphological analysis is presented in [9]. Similar to methods used for Chinese, they perform two disambiguation procedures based on CRFs: 1) morpheme segmentation and 2) POS tagging. In morpheme segmentation, an input sentence is segmented into sequences of morphemes. In the POS tagging procedure, each is assigned a POS tag. Once the POS tagging is complete, they carry out a post-processing of the compound morphemes, where each compound morpheme is further decomposed into atomic morphemes, which is based on pre-analyzed patterns and generalized HMMs obtained from the given tagged corpus.

3. Myanmar Language

Myanmar language is highly agglutinative and is morphologically rich and complex. Moreover, Myanmar scripts do not use white-spaces to separate the one word from another, there is no way of knowing whether a group of syllables form a word, or is just a group of separate monosyllabic words. Every syllable has a meaning of its own. A word in Myanmar may consist of one or more syllables which are combined in different ways. Based on the way of constructing words from syllables, we can classify them into three categories: single simple words, complex words and reduplicative words.

eg. ငေပါင်း (steam) + အိုး(pot) =>ငေပါင်းအိုး (rice cooker), မီး(fire) + ပူ (hot) =>မီးပူ (iron), ပန်း(flower) + ချီ (carry) =>ပန်းချီ (painting), all have their referential meaning and each monosyllable within words also has their own meaning.

4. Hidden Markov Models (HMMs)

An HMM is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), they compute a probability distribution over possible sequences of labels and choose the best label sequence.

A Hidden Markov models(HMMs) are appropriate for situations where somethings are observed and some things are hidden:

-observed events (the words in a sentence)

-hidden events (part-of-speech tags)[6].

In an HMM hidden states model the hidden events which are thought of as generating the observed words.

An HMM is specified by the following components:

$Q= q_1q_2...q_N$	a set of N **states**
$A= a_{11}a_{12}....a_{n1}....a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$.
$O = o_1o_2...o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V= v_1,v_2,....,v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
q_0 ,q_F	a special **start state and end (final) state** that are not associated with observations, together with transition probabilities $a_{01}a_{02}...a_{0n}$ out of the start state and $a_{1F}a_{2F}..a_{nF}$ into the end state

4.1 The Basic equation of HMM Tagging

HMM decoding, that is to select the tag sequence that is most probable given the observation sequence of n words w_1^n:

$$\hat{t_1^n} = argmax_{t_1^n} P(t_1^n|w_1^n) \quad (1)$$

by using Bayes' rule to instead compute:

$$\hat{t_1^n} = argmax_{t_1^n} \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \quad (2)$$

by dropping the denominator w_1^n

$$\hat{t_1^n} = argmax_{t_1^n} P(w_1^n|t_1^n)P(t_1^n) \quad (3)$$

HMM taggers make two further simplifying assumptions. The first assumption, the probability of a word appearing is independent of neighboring words and depends only on its own tag:

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i) \quad (4)$$

The second assumption, the bigram assumption, is that the probability of a tag is dependent only on the previous tag, rather than the entire tag sequence:

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \quad (5)$$

For the best tag sequence from a bigram tagger, simplifying assumption corresponds to the emission probability and transition probability is described in equation (5). [2]

$$\hat{t_1^n} = argmax_{t_1^n} P(t_1^n|w_1^n) \approx argmax_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \quad (6)$$

5. Design of Proposed System

The framework of the proposed system is shown in figure 1. There are two modules: training and testing modules. In the training phase, the collection of segmented and tagged-sentences are used to develop the proposed HMM model. This model is used in the testing phase. In testing phase, the input Myanmar sentences are identified into each sentence using the sentence end marker called pote-ma '။ '. After that, morphological word segmentation and POS tagging is performed.

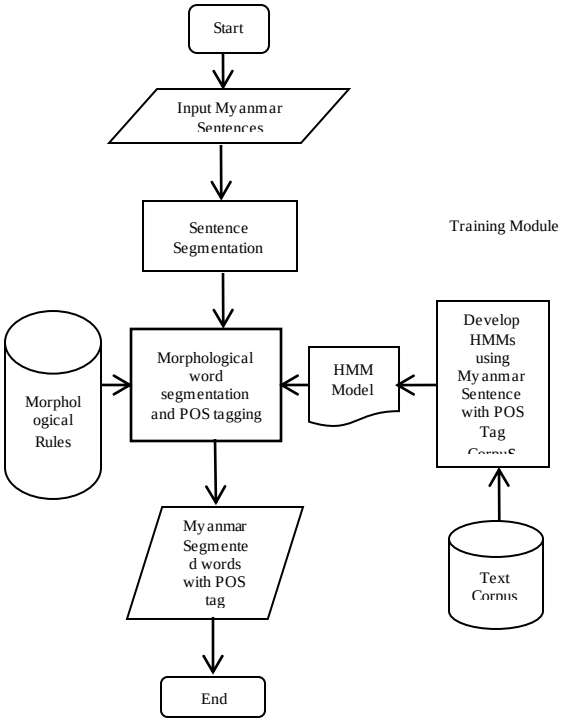

Figure 1. Framework of the Proposed System

5.1 Training Data

We collected the training data from Myanmar News, Myanmar grammar books, eBooks and journals (general domain). In training data included

two parts: Corpus Creation and estimating probabilities.

5.1.1 Corpus Creation

We created the Myanmar text corpus to be used as including sentences from Myanmar News, Myanmar grammar books, eBooks and journals (general domain). Since, documents used various Myanmar font styles; these are converted to standard Unicode font (Myanmar3) and save into our corpus. There are total 7500 sentences covering 112500 words and each sentence has an average of 15 words. The collected sentences are segmented and tagged with proposed POS in [3] manually. Figure 2 shows the sample sentences in our corpus. "/" is word break and "@" is put between word and its POS tag.

စားပွဲ@NN/ပေါ်မှာ@PPM/စာအုပ်@NN/တစ်@Number/
အုပ်@Part/ရှိ@V/သည်@PPM/။@Symbol/
ပန်းသီး@NN/များ@Part/ချို@Adj/သည်@PPM/။@Symbol/
မဖြူ@NN/သည်@PPM/စေး@NN/မှု@PPM/ပေတံ@NN/များ@Part/ဝယ်@V/လာ@Part/သည်@PPM/။@Symbol/
မြန်မာ@NN/နိုင်ငံ@NN/တွင်@PPM/ကမ်းခြေ@NN/များ@Part/ရှိ@V/သည်@PPM/။@Symbol/
ဒေါ်အေး@NN/သည်@PPM/ဦးမြ@NN/၏@PPM/ဇနီး@NN/ဖြစ်@V/သည်@PPM/။@Symbol/

Figure 2. Format of Corpus

5.1.2 Estimating probabilities (HMMs)

POS tagging using HMM, the probabilities are estimated by counting on a tagged training corpus instead of using the full power of HMM learning.

The tag transition probabilities $P(t_i|t_{i-1})$represent the probability of a tag given the previous tag. Estimation of transition probability is computed by counting, out of the times we see the first tag in a labeled corpus, how often the first tag is followed by the second

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})} \quad (7)$$

The emission probabilities, $P(w_i|t_i)$ given a tag, it will be associated with a given word [2]. The MLE of the emission probability is

$$P(w_i|t_i) = \frac{C(t_i,w_i)}{C(t_i)} \quad (8)$$

5.2 Testing

The input sentences are firstly separated by pote-ma "။ ". The words in each sentence is segmented and assigned POS proposed in [3] by using HMMs probabilistic models. This system employs a sentence based approach rather than a word based approach. First all the possible tags for the words and the word sequences in the sentence are determined, and then the combination of the tags with the highest probability for the whole sentence is selected.

For example, the input is as follows:

မိုးရွာလျှင်ကလေးများလမ်းပေါ်တွင်ဘောလုံးကန် ကြသည်။

In Myanmar Language, since words are formed by combining more than one syllable that is one word can have one or more syllables and one syllable has more than one character, syllable identification must be done before word level segmentation.

After Syllable Identification, the right output is come out as follows:

မိုး|ရွာ|လျှင်|က|လေး|များ|လမ်း|ပေါ်|တွင်|ဘော|လုံး|ကန်|ကြ|သည်|။|

A common approach to word segmentation is to use the N-grams Language models which scans an input sentence from left to right, and selects the maximum probability.

In Unigram:

မိုး,ရွာ,လျှင်,က,လေး,များ,လမ်း,ပေါ်,တွင်,ဘော,လုံး,,ကန်,ကြ,သည်,။,

In Bigram :

မိုးရွာ,ရွာလျှင်, လျှင်က, ကလေး, လေးများ, များလမ်း, လမ်းပေါ်, ပေါ်တွင်, တွင်ဘော, ဘောလုံး, လုံးကန်, ကန်ကြ, ကြသည်, သည်။

Maximum probability for word segmentations are like this:

မိုးရွာ,လျှင်,ကလေး,များ,လမ်း,ပေါ်တွင်,ဘောလုံး,ကန်, ကြ,သည်

All possible tags for sequence:

မိုးရွာ@V/ လျှင်@Conj/ကလေး@NN/များ@Part,V/လမ်း@NN/ပေါ်တွင်@PPM/ဘောလုံး@NN/ကန်@V,NN/ကြ@Part/သည်@PPM/

The highest scoring tag sequence:

မိုးရွာ@V/လျှင်@Conj/ကလေး@NN/များ@Part/
လမ်း@NN/ပေါ်တွင်@PPM/ဘောလုံး@NN/ကန်@V/
ကြ@Part/သည်@PPM/

$$t_1^{\hat{}n} = argmax_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

Unknown words cause segmentation errors because OOV words, that are not seen in the training corpus, in an input text normally are incorrectly segmented into pieces of words.
For instance:
input: ချက်လက်မှတ် ပေးပါ (give cheque)

1. ချက်(latch)@NN/လက်(hand)@NN/မှတ်(note)@V/ ပေး(give)@V/ ပါ@Part/
2. ချက်(latch)@NN/လက်မှတ်(ticket)@NN/ ပေး(give)@V/ပါ@Part/
3. ချက်လက်မှတ်(cheque)@NN/ပေး(give)@V/ ပါ@Part/

For all tags t€T: $P(w|t) = 0$

Performance of taggers depends largely on treating unknown words.

To solve unknown words or OOV words, morphological rules approach that has been described in [3] is used.

6. Evaluation

In order to evaluate the experiment result for POS tagging, the system used the parameters of Recall, Precision and F-score. These parameters are defined as follows:

$$Recall, R = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ \ words\ in\ the\ test\ set} \quad (9)$$

$$Precision, P = \frac{Number\ of\ correct\ POS\ tag\ assinged\ by\ the\ system}{Number\ of\ POS\ tag\ assigned\ by\ the\ system} \quad (10)$$

$$F\_score, F = \frac{2PR}{P+R} \quad (11)$$

6.1. Result and Discussion

In the experiments, two test sets are used for evaluation in order to calculate the accuracy of the word segmentation and tagging. The comparison of accuracy result has been described using HMM only and HMM with morphological analysis. Each testing contains 200 sentences. First test set (A) has 15% unknown words and second test set (B) has 30% unknown words. All tested 400 sentences are randomly chosen from news websites and Myanmar grammar books.

Table 1 depicts the experimental results of the two models.

Table 1: Experiment results

Models	Test set: A			Test set: B		
	R	P	F	R	P	F
HMM	86.57 %	88.24 %	87.4 %	74. 8%	76. 25 %	75.52 %
HMM + MA	92.3 %	94.49 %	93.38 %	83. 21 %	85. 33 %	84.26 %

7. Conclusion

This paper presents a joint word segmentation and POS tagging in Myanmar using HMM and morphological rules. The accuracy of using HMM only has no attraction. Therefore, the combination of HMM and morphological rules have been proposed to get the better Myanmar word segmentation and POS tagging. We also intend to make a larger corpus in order to reduce the OOV words and also incorrect tag. For Syllable Identification using [10] and for word segmentation and POS tagging N-grams and HMMs is used. The system is implemented using Python 3.7.

References

[1] Aye Myat Mon , Myint Myint Thein, Su Su Htay, Soe Lai Phyue, Thinn Thinn Win, *"Analysis of Myanmar Word Boundary and Segmentation By using Statistical Approach"*, University of Computer Studies, Mandalay

[2] Daniel Jurafsky,James H. Martin, "*Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition",* Copyright 2006, Draft of June 25, 2007

[3] Dim Lam Cing, Khin Mar Soe, "*Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language*", 16th International Conference on Computer Application, Yangon, 22-23, February,2018

[4] Hopple, "The structure of nominalization in Burmese. Ph.D Dissertation", University of Texax, Arlington, 2003

[5] Phyu Hninn Myint, Tin Myat Htwe and Ni Lar Thein, *"Bigram Part-of-Speech Tagger for Myanmar Language",*2011 International Conference on Information Communication and Management, IPCSIT vol.16 (2011), Singapore

[6] Win Pa Pa, Ni Lar Thein ,"*Myanmar Word Segmentation using Hybrid Approach",* Proceedings of 6th International Conference on Computer Applications, Yangon, Myanmar

[7] Yue Zhang, Stephen Clark, *"Joint Word Segmentation and POS Tagging using a Single Perceptron",* Proceedings of ACL-08: HLT, pages 888–896, Columbus, Ohio, USA, June

[8] Hakimeh Fadaei , Mehrnoush Shamsfard, *" Persian POS tagging using probabilistic morphological analysis", Int. J. Computer Applications in Technology, Vol. 38, No. 4, 2010*

[9] Seung-Hoon Na, "Conditional Random Fields for Korean Morpheme Segmentation and POS Tagging", *ACM Transactions on Asian Language Information Processing*, Vol. 14, No. 3, Article 10, Publication date: June 2015

[10] https://github.com/ye-kyaw-thu/sylbreak