Comparison of two approaches for Part of Speech Tagging of Myanmar Language

Khine Khine Zin, Prof. Dr. Ni Lar Thein
University of Computer Studies, Yangon
khine2zin@gmail.com, nilarthein@gmail.com

Abstract

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. There are different approaches to the problem of POS Tagging. In this paper, we use two approaches (supervised and unsupervised approaches), and compare the performance of these approaches for Tagging using Myanmar language. These approaches use rule based POS Tagger and the former requires a large amount of annotated training corpus to tag properly and the later does not need annotated training corpus. We tried to see which technique maximizes the performance with limited resources. By experiments, the best configuration is investigated using supervised approach with rule based part of speech tagger and the accuracy is 97.56%. Therefore, this approach has better performance than unsupervised approach with rule based part of speech tagger.

1. Introduction

Part of Speech (POS) Tagging is the essential basis of Natural Language Processing (NLP). It is the process in which each word is assigned to a corresponding POS tag that describes how this word be used in a sentence. Typically, the tags can be syntactic categories, such as noun, verb and so on [5, 6, 7, 14]. For Myanmar language, word segmentation must be done before POS tagging, because, different from English sentences, there is no distinct boundary such as white space to separate different words. Also, word segmentation and POS tagging can be done at the same time. While a number of successful POS tagging systems have been available for English and many other languages, it is still a challenge to develop a POS tagging for Myanmar due to its language-specific issues.

Part of Speech Tagging is another important evaluation task for Myanmar language. A Part of Speech Tagger typically assigns a tag to each word in a sentence sequentially from left to right. When the words are tagged from left to right, the Part of Speech Tags assigned to the previous words are available to the tagging of the current word, but not the tags of the following words. We expect the use of the tags of the words on left side of the current word should improve the tagging of the current word. The words in both the training and testing data sets are already segmented into words [5, 9].

There are a variety of approaches for POS tagging. Among them, two approaches to POS tagging are

1. Supervised approach using Rule based POS Tagger, and
2. Unsupervised approach using Rule based POS Tagger

The former Taggers had large sets of hand-constructed Rules for assigning tags on the basis of words' character patterns and on the basis of the tags assigned to preceding or following words, but they had only small lexica, primarily for exceptions to the rules. But, the later Tagger does not require hand-constructed Rules and so it needs automatic rules. A POS tagger based HMM assigns the best sequence of tags to an entire sentence. The task of Part of Speech (POS) Tagging is to find the sequence of POS tags T = {t1, t2, t3, . . ., tn} that is optimal for a word sequence W = {w1, w2, w3, . . ., wn} [8].

This paper is structured as follows: a brief description of supervised approach using rule based POS tagging is described in Section 2, section 3 describes about Unsupervised approach using rule based POS Tagging and also describe which approach should be used to get better performance, and the main propose of our system is presented in section 4, experimental results in section 5 and we end with conclusion in section 6.

2. Supervised Approach using Rule Based POS Tagging

Supervised POS tagging is a machine learning technique using a pre-tagged corpora in which it requires training data. The training data consist of pairs of input objects and desired outputs. The output of the function can be a continuous value, or can

predict a class label of the input object. The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples i.e. pairs of input and target output. To achieve this, the learner has to generalize from the presented data to unseen situations in a "reasonable" way. This method also requires large amount of tagged data to achieve high level of accuracy.

Supervised learning can generate models of two types. Most commonly, supervised learning generates a global model that maps input objects to desired outputs. In some cases, however, the map is implemented as a set of local models. In order to solve a given problem of supervised learning, the type of training examples are determined. Then a training set (a set of input objects and outputs) is gathered either from human experts or from measurements. The learning algorithm is then run on the gathered training set. Parameters of the learning algorithm may be adjusted by optimizing performance on a subset of the training set. After parameter adjustment and learning, the performance of the algorithm may be measured on a test set that is separated from the training set. These rules are applied in construction the tagger in our native language, Myanmar by using Supervised POS Tagging as follows:

In this language, the tagger initially tags by assigning each word its most likely tag, estimated by examining a large tagged corpus, without regard to context. Since the goal of Supervised POS Tagging is to build a concise model of the distribution of class labels in terms of predictor features, supervised method learns tags and transformation rules from manually tagged corpus. So, the words in the following sentence in Myanmar language are tagged to the appropriate tags from the pre-tagged set like this:

အမေ_NN သည်_SPRP သား_NN ကို_SPRP ယပ်ခပ်ပေးသည်_V

3. Unsupervised Approach using Rule Based POS Tagging

When any training data does not require a annotated corpora, the learner is given only rule based examples which are closely related to the problem of density estimation in statistics. However, rule based learning also encompasses many other techniques that seek to summarize and explain key features of the data. In order to derive a rule of the learner, an objective function must be found for training that does not need a manually tagged corpus.

In order to tag each word with the most appropriate tag, the starting word is assumed with noun tag (NN) if this word is not found in the annotated corpus, and then the other tag sequence can be searched by using rules by untagged corpus. Unlike other languages, the starting word may be noun or other tags in Myanmar Language. Therefore, after assigning this word with noun (NN), firstly, it need to test whether this tag is correct or not by using rules. For example, if the word is ended with "သော", this must be changed with adjective Tag (JJ).

For example, the verb (V) in Myanmar language is terminated (eg., 'သည်', 'ပါ', 'မည်', etc) at the end of sentence, the adjective by **'သော'** and etc. This information is derived automatically from the training corpus. But, consider the following two sentences,

(၁)သူငယ်ချင်း_NN လာမယ့်_JJ တနင်္ဂနေနေ့_NN မှာ_SPRP ငါ_NN ရုပ်ရှင်_NN သွားမလို့_V

(၂)ငါ_NN ပြောသလို_V နင်_NN လုပ်ပါ_V

The words **'လာမယ့်'** in former sentence and **'ပြောသလို'** in later sentence are not found in the tagged corpus and so unsupervised approach with rules must be applied. eg., Adjective (ADJ) must be followed by noun (N) and simple sentences are constructed that subject (NN) is followed by verb (V) and also the sentences are ended with the verb (V). The word **'ပြောသလို'** doesn't found at the end of the sentences and but this word is tagged with verb (V) by using rules and also the word **'လာမယ့်'** also follow this rule and thus that word is tagged with adjective (ADJ).

Therefore, it needs to construct more and more rules to get better performance. But, in Myanmar Language, rules are difficult to construct manually and so the better performance cannot be achieved using only unsupervised approach using Rule Based Tagger. Therefore, by using supervised approach using rule based Tagger in this paper, the best accuracy will be achieved in POS Tagging of Myanmar Language. The most accurate taggers use estimates of lexical and contextual probabilities extracted from large manually annotated corpora. Therefore, if tagged text is needed in training, this would require manually tagging text each time the tagger is to be applied to a new language, and even when being applied to a new type of text. A supervised approach using rule-based POS tagger is described to achieve highly performance, and capture the learned knowledge in a set of rules instead of a large table of statistics. In addition, the learned rules can be converted into properties by HMM. We describe the proposal for supervised approach using rule based POS Tagger, and compare the performance with using unsupervised approach.

4. The Proposed Model

Our processing procedure of the system is shown in Figure 1.

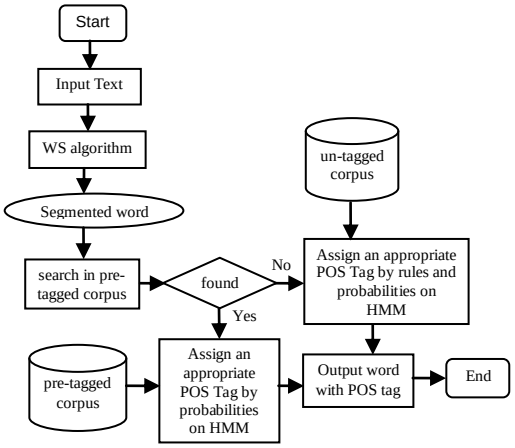

Figure 1. Processing steps of POS Tagging

In our system, Myanmar language has some fixed vocabulary, $\{w^1, w^2, ..., w^w\}$. This is a set of words, e.g., {မောင်နိုင်နိုင်, ကျောင်းသား, နှင်းဆီပန်း , ……, လှပသော} . And a fixed set of parts of speech, or tags, {t^1, t^2,….., t^T}, e.g., {adjective, adverb, noun,……., verb} is created. A text of n words is considered to be a sequence of random variables W ^{1,n}=W^1 W^2 …..W^n .Each of these random variables can take as its value any of the possible words in our vocabulary. More formally, let the function V(X) denotes the possible values (outcomes) for the random variable X. Then V(Wi)={w1, w2, …ww}. It denotes the value of W1 by w1, and a particular sequence of n values for W1,n by w1,n. In a similar way, the tags for these words are considered to be a sequence of n random variables T1,n = T1, T2,…., Tn. A particular sequence of values for these is denoted as t1,n and the ith one of these is ti. The tagging problem can then be formally defined as finding the sequence of tags t1,n which is the result of the following function:

$$\tau(w_{1,n}) \overset{def}{=} \arg_{t_{1,n} \in V(T_{1,n})} P(T_{1,n} = t_{1,n} | W_{1,n} = w_{1,n}) \quad (1)$$

$$\tau(w_{1,n}) = \arg\max_{t_{1,n}} P(t_{1,n} | w_{1,n}) \quad (2)$$

$$\tau(w_{1,n}) = \arg\max_{t_{1,n}} \frac{P(t_{1,n}, w_{1,n})}{P(w_{1,n})} = \arg\max_{t_{1,n}} P(t_{1,n}, w_{1,n}) \quad (3)$$

Next, this equation is broken down into "bit-size" pieces about which statistics can be collected.

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(w_i | t_{1,i-1}, w_{1,i-1}) P(t_i | t_{1,i-1}, w_{1,i}) \quad (4)$$

"Markov assumptions" are possible to view the tagging as a Markov process. So it is needed to start with the simplest of these models (i.e., the one based upon the strongest Markov assumption). Equation (4) is started to make the following Markov assumption

$$P(w_i | t_{1,i-1}, w_{1,i-1}) = P(w_i | w_{1,i-1}) \quad (5)$$

$$P(t_i | t_{1,i-1}, w_{1,i}) = P(t_i | w_i) \quad (6)$$

Substituting these equations into equation (4) and substituting that into equation (3), the following equation is obtained.

$$\tau(w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(w_i | w_{1,i-1}) P(t_i | w_i) \quad (7)$$

$$= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i | w_i) \quad (8)$$

But, since the current tag is independent of the previous words and only dependent on the previous tag. Similarly we assume that the correct word is independent of everything except knowledge of its tag. With these assumptions we get the following equation:

$$\tau(w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i | t_{i-1}) P(w_i | t_i) \quad (9)$$

Our algorithm runs on a number of iterations. First, we process the tagged data by Supervised Learning then in each iteration it processes the untagged data using Rules and updates the transition probabilities i.e. p (tag | previous tag) and emission probabilities i.e. p (word | tag) for the Hidden Markov Model. Using tagged data each word maps to one state as the correct Part of Speech is known. But using untagged data (Rules), each word will map to all states because Part of Speech Tags are not known i.e. all states we considered possible. To realize the implementation of these approaches using Rule Based POS Tagger, firstly we have to design tag set (our approach has 36 tags). Some of these are shown in Table I:

TABLE I. SOME OF TAGS USED IN OUR SYSTEM

Tag	Description	Example
NN	Noun, singular or mass	နှင်းဆီပန်း၊ မောင်မောင်
NNS	Noun, plural	နှင်းဆီပန်းများ၊ လူများ
PRP	Personal pronoun	ကျွန်မ၊ သူ၊ ကျွန်တော်
POS	Possessive pronoun	၏
VB	Verb, base form	ရှိသည်၊ ဖြစ်သည်
V	Verb	သွားသည်၊ စားသည်
RB	Adverb	ကောင်းမွန်စွာ၊ လျင်လျင်မြန်မြန်
JJ	Adjective	လှပသော၊ ကောင်းသော
SYM	Symbol	၁၊ ၂၊ %
CD	Cardinal number	တစ်ခု၊ နှစ်ကောင်၊ သုံးယောက်

The following examples are shown that how to tag the segmented words in several sentences using Myanmar Language and their respective meaning are shown.

(1)ဦးမောင်မောင်_NN သည်_SPRP ကျောင်းဆရာ_NN တစ်ယောက်_CD ဖြစ်သည်_VB

(2) ကျောင်းသားများ_NN သည်_SPRP ကျောင်း_NN သို့_PRP အချိန်မှန်မှန်_RB သွားကြသည်_V

(3) ကြိုးစားခြင်း_NN သည်_SPRP အောင်မြင်ခြင်း _NN ၏_POS လမ်းစ_NN တစ်ရပ် _ CD ဖြစ်သည်_VB

(4) သူမ_NN သည်_SPRP ရိုးသားဖြောင့်မတ်သော_JJ မိန်းကလေး _NN တစ်ယောက်_CD ဖြစ်သည်_VB

(5) ကျမ်းမာချမ်းသာကြပါစေ_V

5. Experimental Results

In order to measure the performance of the system, we use the manually annotated test data using the tag set consisting of 36 grammatical tags and create corpus with different number of words collecting from Myanmar Newspaper. The corpus is defined as possible tags for each word. We also create un-annotated test data by grammatical rules in order to tag the unfound words in the annotated data. From our training data, we were able to extract data from unique unambiguous tag sequences which were then be used for better initializing the state transition probabilities. Then, we have tested many experiments using supervised and unsupervised approach using rule based POS Tagger on different corpus till we get the best accuracy. Then, we have seen that supervised approach using POS Tagging get the better accuracy than unsupervised approach. Table II shows the performances of POS tagging according to the different approaches on different number of words in corpus. Figure 2 also shows the comparison of these improvements in accuracy along with the increase in the size of annotated training data on different methods.

TABLE II. PERFORMANCE OF POS TAGGING FOR DIFFERENT NUMBER OF WORDS

Number of words	Accuracy using Unsupervised Approach	Accuracy using Supervised Approach
5,000	56.43%	60.23%
10,000	61%	70.6%
50,000	67.0%	80.19%
100,000	71.23%	83.67%
150,000	75.6%	85.46%
200,000	77.89%	87.2%
300,000	80.9%	90.70%
500,000	84.32%	93.67%
1,000,000	89.56%	97.56%

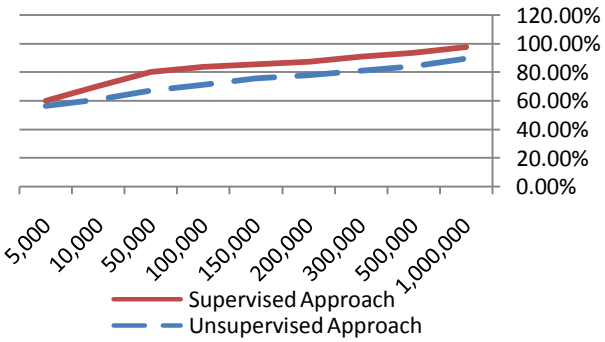

Figure 2. Tagging accuracies on different words

6. Conclusion

In this paper, we have presented an approach for Part of Speech Tagger of Myanmar Language using HMM with Rule Based Approach. And also, we describe this approach has the better accuracy than using Rule Based Approach only. This tagger significantly outperforms Natural Processing System which assigns to a word the most likely tag assigned to that word in the training corpus. So, the present system is still under the development, especially in morphological knowledge acquisition. For the future work, we hope to improve our system with information retrieval and Machine Translation in Myanmar Language. Thus, the use of features involving the POS tags of the following words further improves the performance of the tagger.

7. References

[1] A.Steven, Part of Speech Tagging and Partial Parsing, Conference of Corpus-Based Methods in Language and Speech, 1997, pp. 118-136.

[2] B.Eric, A Simple Rule-Based Part of Speech Tagger, Department of Computer Science, University of Pennsylvania, U.S.A. Human Language Technology

Conference. 1992. ISBN: 1-55860-272-0. pp. 112-116

[3] B.Eric, Unsupervised Learning of disambigua- tion Rules for Part of Speech Tagging, NSF grant IRI-9502312. 1995. pp.1-13.

[4] B.Michele and C.Robert, Part of Speech Tag- ging in Context, Microsoft Research, WA 98052 USA. International Conference on Computational Linguis-tics. August. 2004.

[5] C.Aitao, Z.Ya, S.Gordon, A Two-Stage App- roach to Chinese Part-of-Speech Tagging, Sixth SIGHAN Workshop on Chinese Language Processing Yahoo! Inc. 701 First Avenue Sunnyvale, CA 94089. 11 January 2008. pp. 82-85

[6] C.Eugene and H.Curtis and J.Neil and P.Mike, Equations for Part-of-Speech Tagging, Providence RI 02912. 1993.

[7] D.Sandipan, S.Sudeshna and B.Anupam, A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali, International Conference on Computatio-nal Intelligence, December 2004. pp.169-172.

[8] E.Asif, M.Samiran and B.Sivaji, POS Tagging using HMM and Rule-based Chunking, 2007.

[9] F.Guohong, J.Webster, A Morpheme-based Part-of-Speech Tagger for Chinese, Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing. pp. 124-127

[10] H.Fahim Muhammad, U.Naushad and K.Mumit, Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla, Tuesday August 28 2007, ISBN 978-1-4020-6263-6 (Print) 978-1-4020-6264-3 (Online), pp. 121-126

[11] H.M.Fahim, U.Naushad, K.Mumit, Comparison of Unigram, Bigram, HMM and Brill's POS Tagging approaches for some South Asian Languages, Center for Research on Bangla Language Processing (CRBLP), BRAC University, 2007.

[12] S.B.Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31 October 2007. pp. 249-268

[13] S.Rama, P.K.Kusuma, Combining POS Taggers for improved accuracy to create Telugu annotated texts for Information Retrieval, 2007 ICUDL, Carnegie Mellon University, Pittsburgh,USA-ULIB.

[14] X.Hongzhi and L.Chunping, Combining Con- text Features by Canonical Belief Network for Chinese Part-Of-Speech Tagging, 30 December 2007. pp. 907-912