A HYBRID APPROACH FOR PART-OF-SPEECH TAGGING OF BURMESE TEXTS

Cynthia Myint
University of Computer Studies, Mandalay
Mandalay, Myanmar
cynthiamyint@gmail.com

Abstract*—In Myanmar to English language translation system, in order to provide meaningful sentence from one language to another is non-trivial task. POS tagging is used as an early stage of linguistic text analysis in many applications. POS tagging is a process of assigning correct syntactic categories to each word. Tagsets and word disambiguation rules are fundamental parts of any POS tagger. This paper presents a new approach for POS tagging of Myanmar Language. Firstly, Users input a simple Myanmar sentence and then this sentence is segmented into words by using segmentation rules. These words are assigned to appropriate syntactic categories of Myanmar language by using rule based and probabilistic approach. This system applied CRF method for tagging POS ambiguities on words. CRF is a framework for building discriminative probabilistic models for segmenting and labeling sequential data. The tagsets for Myanmar POS, segmentation rule, tagging algorithm and CRF method are designed. The proposed approach is used UCSM Lexicon. So, this hybrid approach for POS tagging can give the optimal accuracy and robustness of machine translation system.

Keywords-POS tagging; CRFmethod; Syntatic Categories;

I. INTRODUCTION

The need for computers to understand natural language is growing by the day as human-computer interfaces become more intuitive. Part of Speech tagging is a process of assigning syntactic categories like noun, pronoun, verb and adjective to each word in the text document [1]. POS Tagging is an essential part of Natural Language Processing (NLP) applications such as speech recognition, text to speech, word sense disambiguation, information retrieval, semantic processing, parsing, information extraction and machine translation. Taggers can be divided as supervised or unsupervised. (i) Rule Base Taggers, (ii)Stochastic Taggers (iii) Transformation Based Taggers All these three approaches can be used with supervised as well as unsupervised taggers. Among these approaches this system is rule based and stochastic tagger. This system applied CRF (Conditional Random Field) method for tagging Burmese Texts. CRF is a framework for building discriminative probabilistic models for segmenting and labeling sequential data.

The rest of this paper is organized as follows. Section 2 describes the related work of this system. Section 3 introduces Myanmar Language. Section 4 discusses the part_of_speech tagging of this system. Section 5 presents the proposed model and Section 6 concludes this paper.

II. RELATED WORK

K.Tuntonova, C. D. Manning "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger"[2] examined K.tutonova and et.al present results for a maximum entropy-based part of speech tagger, which achieves superior performance principally by enriching the information sources used for tagging. They also showed the improvement of results by incorporating three features such as (i) capitalization for unknown words.(ii) disambiguation of the tense forms of verbs and(iii) disambiguating particles from prepositions and adverbs .The best resulting accuracy for the tagger on the Penn Treebank is 96.86% overall, and 86.91% on previously unseen words.

A combination of both statistical and rule-based methods has also been used to develop hybrid taggers. D. Shamsfard and et. al [3] presents tagging algorithm in which combines the features of probabilistic and rule-based taggers to tag Persian unknown words. Approximately 97% accuracy is achieved. M .Q. Murata and et.al [4] describes hybrid system for Thai part of speech tagging that consists of a neuro tagger and a rule bases corrector. The accuracy of 99.1% is achieved which is higher than the HMM and rule-based approaches. L. Altunyur and et. Al [5] presents a composite part of speech tagger for Turkish in which they combine the rule-based and statistical approaches with two additional features and achieved approximately 85% accuracy.

III. MYANMAR LANGUAGE

The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which is descended from the Brahmi script of ancient South India [6]. Other Southeast Asian descendants, known as Brahmic or Indic scripts, include Thai, Khmer and Lao.

A. Myanmar Grammar

Myanmar Language commission standardized that it is composed of nine part- of- speech in Myanmar Grammar such

978-1-4244-9283-1/11/$26.00 ©2011 IEEE

as noun, pronoun, adjective, verb, adverb, post-positional marker, particle, conjunction and interjection. Two types of sentences in Myanmar are simple and complex or compound. Words are traditionally grouped into equivalence classes called POS, word classes, morphological classes or lexical tags. In English sentence, POS for a word gives a significant amount of information about the word and its neighbors.

B. Myanmar POS tagsets

The language tagsets represents part of speech and tagsets consists of syntactic classes. According to contextual and morphological structure, natural languages are different from each other. Therefore, it is necessary to have a tagset for the Myanmar language before developing part of speech tagger.

According to Myanmar Grammar, the general parts of speech are noun, pronoun, adjective, verb, adverb, post-positional marker, particle, conjunction and interjection. There are 88 tagsets for Myanmar POS according from different grammar categories. Noun can be divided into Noun (NN), Collective singular(NNCLS), and so on. Numerals are divided into cardinal (CA),ordinal(OD), fractional(FR),multiplication (MUL).Two type of marker are used in the text ,sentence end marker (SEM),phrase marker(PM). English style date and time is used in the Burmese text. Therefore, (DATE) and (TIME) tags are used.

TABLE I. SOME POS TAGSETS FOR MYANMAR

Tagset	English	Myanmar
SPSM	Nominative	သည်ကမှာ
VPSM	Verb postposition	သည်၊ ၏
OPSM	Objective	ကို
DPSM	Departure	မှ
DIPSM	Direction	သို့
NNCMS	Common countable singular	အလံတော်
JJD	Adjective Descriptive	ကြီး
PRPNS	Personal nominative singular pronoun	ကျွန်ုပ် ကျွန်တော်

IV. PART OF SPEECH TAGGING

There are a variety of techniques for POS tagging. Two approaches to POS tagging are

1. Supervised POS Tagging
2. Unsupervised POS Tagging

Supervised tagging technique requires a pre tagged corpora where unsupervised tagging technique do not require a pre tagged corpora. Supervised POS tagging is a machine learning technique using pretagged corpora in which it required training data. In Myanmar, there have no Myanmar corpora. So, this system uses UCSMlexicon. Data are manually trained.

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.

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 from the UCSM Lexicon.

V. PROPOSED MODEL

In this section, this paper presents the proposed model for processing steps for POS tagging. In First step, users can input Myanmar simple sentence to tag appropriate syntactic categories, this sentence used pretagged UCSMlexicon and a rule based POS tagger. Firstly, system accepts input sentence and segments the sentence into words. The sentence is segmented into ကျောင်းသား များ သည် အလံတော် ကို တိုင်ထိပ် ၌ နံနက်ဖိုင်း တွင် လွှင့်ထူ သည်။This system looks up all postposition and then mark these words .And then search sentence end marker ။.The previous position of endmarkerသည် assign into VPSM .The previous word of VPSM assign လွှင့်ထူ CPVT and then search into(သည်) , (ကို) , (၌) , (တွင်) and marked as(SPSM) , (OPSM), (PPSM) , (TPSM). ကျောင်းသားများis the previous word of (သည်) as assign into NNCMP. အလံတော် is the previous word of (ကို) as assign into NNCMS. တိုင်ထိပ် is the previous word of (၌) as assign into NNCPS. နံနက်ဖိုင်း is the previous word of (တွင်) as assign into NNCPS according to rule. And then the system assigns to appropriate POS tagset of Burmese texts such as

ကျောင်းသား များ _NNCMP သည် _ SPSM အလံတော် _ NNCMS ကို _OPSM တိုင်ထိပ် _NNCPS၌_PPSM နံနက်ဖိုင်း_NNCPS တွင် _TPSM လွှင့်ထူ _ CPVT သည် _VPSM

ညီမလေး NNCMS သည် SPSM ကြ NNCMS ကို OPSM ဆင် NNCMS ကြီး JJD အား ACPPPSM ကျွေး VT သည် VPSM

A. Segmentation and Word Disambiguation Rules

Rule1: Tokenize input text using segmentation rule of longest matching.
//Expection case(two repetitive postposition)
Rule2:If wordi==“arr” && wordi-1==“ko” then segment the sentence “ko” as SPSM
For verb
Rule3:find sentence end marker “the,ei,pyi,par,lar,tae,myee,eant,latent”
If (wordi==“the”) then assign the wordi-1 to verb(present).
Else if (no endmarker)&& (lastword==VPSM) then assign the word to verb.(present).
Else if ((wordi== “the”) && (wordi-1) ==“kya”))then omit the “kya” && assign the word to verb(present).
Else if ((wordi== “the”) && (wordi-1) ==“khae”))then assign the word to verb(past).
Else if ((wordi== “the”) && (wordi-1) ==“nay”))then assign the word to verb(ing).
Rule4:If((wordi==”Verb”)&& (wordi-1)==(“swar”)) then (wordi-1) into Adverb.
For Noun
Rule5:If(wordi= =”SPSM”) then wordi-1 into Noun.
Rule6:If(wordi==”Noun”) then wordi-1 into Adjective.
Rule7:If{(wordi==”Noun”) && ((wordi-1)==”thaw”)} then wordi-1 into Adjective.

To solve the ambiguities between words that can not be solved by the rule. This system used probabilistic approach. So, tagger of this system is designed using hybrid approach of rule based and probabilistic. Probabilistic approach determines POS tags based on conditional probabilities given surrounding context features, where these probability values are obtained from a manually tagged corpus. This system applied CRF method for calculating the probability of tagsets for Myanmar Language.

Let ၀၀ သည် အနိုရောင် ကုတ်အကျီ ' ကို ဝတ်ဆင် သည်. In this sentence, the system can be assigned in အနိုရောင် to noun or adjective from Myanmar Grammar Rule. To solve the disambiguation between words which can not be solved by the rule, this system used probabilistic approach. This system applied CRF method for calculating the probability of tagsets for Myanmar Language.

CRF is a framework for building discriminative probabilistic models for segmenting and labeling sequential data .CRF is an undirected graphical model in which each vertex represents a random variable Yi whose distribution is conditioned on some observation sequence X, and each edge represents a dependency between two random variables. CRF model estimates the probabilities based on the imposed constraints. Such relationships are derived from the training data, maintaining some relationships between features and outcomes. Training data from this system is used from UCSM Lexicon.

The conditional dependency of each Yi on X is defined through a set of feature functions of the form f(Yi-1,Yi,X,i),which can be thought of as measurements on the observation sequence X that partially determine the likelihood of each possible value for Yi. After calculating feature function, the system will obtain the conditional dependency of Yi on X. And the , this system calculates the probability of a label sequence Y given an observation sequence X and CRF model λ by using the equation. Thus, the probability of a label sequence Y given an observation sequence X and CRF model λ can be written as

$$p(Y \mid X,\lambda) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right)$$

Let ၀၀ သည် အနိုရောင် ကုတ်အကျီ ' ကို ဝတ်ဆင် သည်။be observation sequence X,(X1,X2,……,Xi).
X1 ၀၀ X2 သည် X3 အနိုရောင် X4 ကုတ်အကျီ X5 ကို X6ဝတ်ဆင် X7သည်.
NN SPSM JJDP NNCMS OPSM VB VPSM be label sequence Y (i.e,Y1,Y2,……,Yi).

CRF model for NOUN of အနိုရောင်

$$\sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)) \quad = 0.27$$

$$\cdot \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right) \quad = 0.7631$$

$$\frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right) \quad = 0.8478$$

CRF model for ADJECTIVE of အနိုရောင်

$$\sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)) \quad = 0.68$$

$$\sum_j \lambda_j \quad = 0.8$$

$$\cdot \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right) \quad = 0.7180$$

$$\frac{1}{Z(X)} \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(Y_{i-1}, Y_i, X, i)\right) \quad = 0.8975$$

Where Z(X) is a normalization factor, and the λj values are weights learned from manually training data. where f(Yi-1,Yi,X,i), Yi-1 be အနိုရောင်and Yi be ကုတ်အကျီ ' and i be total word of input sentence in X.

Comparing form probability of NOUN 0.8478 and probability of ADJECTIVE 0.8975, this system chooses for word အနိုရောင် into POS tags JJD.
Example Sentences for POS tagging using CRF is

မမ NNCMS သည်SPSM အနီရောင် JJD ကုတ်အင်္ကျီ ‘NNCMS ကို OPSM ဝတ်ဆင် VT သည်VPSM

တတိန်ကား:NNCMS ဖြင့် RPSM ကျောင်း:NNCMS သို့.ARPSM သွားနေသော်JJD ဒိန်းကလေး:NNCMS သည် SPSM ယဉ်ကျေး VBSBJ သည် VPSM

POS Tagging Algorithm

Take input text
1. Tokenize input text.
2. Store all words into array WORD
3. Select each word one by one from array WORD.
4. Search and compare selected word from UCSM Lexicon
5. If word is found one or more times, then store associated tag or tags of word into array TAGS
6. Else display “the word is not found”, add this new word with corresponding tag into lexicon. Add linguistic rules for new word.
7. If one tag is stored in array TAGS, then display word with associated tag as an output.
8. Else select one or more linguistic rules and search most appropriate tag for a word by applying rules
9. Display words with associated tag as an output.

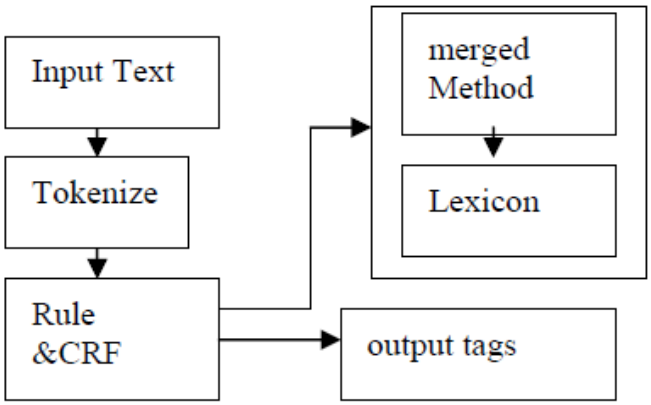

Figure 1. Architecture of POS tagging

The POS tagging architecture consists of different modules which perform different functionalities to achieve better accuracy of POS tagger. Input text is tokenized by using linguistic rules of Myanmar Language. And then untagged tokenized output words can be tagged according to POS tagging algorithm using UCSM Lexicon for known words. Unknown words for untagged POS must be tagged by using CRF model of this system. And finally, outputs the appropriate tag for input text.

The performance of the POS tagger in terms of accuracy is evaluated using the hybrid approach of this system This system is used the manually annotated test data using the tag set consisting of 88 grammatical tags and manually training data from UCSM Lexicon with different number of words .This system creates un-annotated test data by grammatical rules in order to tag the unfound words in the annotated data. Many example sentences are tested and trained by using this system. This system is tested on of 400 sentences and 5225 tokens.

TABLE I. ACCURACY ANALYSIS ON POS TAGSETS

average accuracy tagsets=88	known words accuracy	Unknown words accuracy
84.71%	89.53%	79.89%

VI. CONCLUSION

This paper presents a new approach for POS tagging of Myanmar Language. Firstly, Users input a simple Myanmar sentence and then this sentence is segmented into words by using segmentation rules. These words are assigned to appropriate syntactic categories of Myanmar language by using rule based and probabilistic approach try to solve ambiguities tag sets for Burmese texts. The tagsets for Myanmar POS, segmentation rule, tagging algorithm and CRF method are designed .The proposed approach is used and tested on UCSM Lexicon. So, this hybrid approach POS tagging can give the optimal accuracy and robustness of machine translation system.

REFERENCES

[1] J.A. Mahar and G. Q Memon,” Sindhi Part of Speech Tagging System Using Wordnet”, International Journal of Computer Theory and Engineering (IACSIT), Vol. 2, No. 4,pp538 to 545,1793-8201 August, 2010.

[2] K . Tuntonova, C. D. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger ”,In proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000),pp 63-70.

[3] D. Shamsfard, M . Fadaee, “A Hybrid Morphology-Based POS Tagger for Persian”, Proceeding of the 6th International Language Sources and Evaluation, pp. 3453-3460,2008.

[4] M .Q. Murata, M. Uchimoto, K.Isahar, “Hybride Neuro and Rule-Based Part of Speech Taggers”, International Conference on Computation Linguistics”, pp. 509-515,2000.

[5] L. Altunyur, Z. Orhan, T.Gungor, , “Towards Combining Rule-Based and Statistical Part of Speech Tagging in Agglutinative Languages”, Computer Engineering, Vol 1. pp. 66-69,2000.

[6] Unicode Consortium, The Unicode Standard 4.0: Southeast Asian Scripts (Addison Wesley, California, 2004).

[7] Myanmar Sar A Phwae, Myanmar Spelling Bible, Myanmar Thut Pon Kyan, (Myanmar Sar A Phwae, Yangon, 2003).

[8] Myanmar Grammar, Memorable for 50th anniversary, Myanmar Sar A Phwae, Yangon, 2005, June.