A PART OF SPEECH TAGGER FOR MYANMAR TEXT

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

ABSTRACT

POS tagging is the process of marking up the words in a text as corresponding to a particular part of speech based on both its definition, as well as its context. Tagsets and word disambiguation rules are fundamental parts of any POS tagger. A part of speech tagger for Myanmar text is proposed that can provide many benefits to Myanmar to English language translation system and other NLP tasks. It exploits heuristic rule based approach to tag the correct syntactic categories of the words in Myanmar text. The tagger of this system demonstrates on the following ideas: (i) Design detail tagsets for Myanmar text by using different part-of-speech in Myanmar grammar. (ii) Generate word disambiguation rules to assign the correct syntactic categories of each word in the text by using the morphological structure in Myanmar grammar. It also creates a small lexicon which can be applied to serve the training data for word segmentation and POS tagging for the words in the text. The proposed system can give the promising tagging accuracy by applying the morphological structure of Myanmar grammar and word disambiguation rules.

Keyword: Tagsets, Syntactic Categories, Heuristic Rule Based Approach.

1. INTRODUCTION

Part of speech (POS) tagging is one of the most well studied problems in the field of Natural Language Processing (NLP). Parts of speech (POS) tagging means assigning grammatical classes i.e. appropriate parts of speech tags to each word in a natural language sentence. POS tagging can be used in Text to Speech, information retrieval, shallow parsing , information extraction, word sense disambiguation linguistics research for corpora [10] and also as an intermediate step for higher level NLP tasks such as parsing, semantics, translation and many more[10]. D. Jurafsky and et.al generally divided the tagger into three classes, namely into rule based taggers, stochastic taggers and transformation based taggers. All these three approaches can be used with supervised as well as unsupervised taggers. Rule-based tagging assigns a word all possible tags and then uses context rules to disambiguate [2].

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right. Myanmar writing does not use white spaces between words or between syllables. Thus, the computer has to determine syllable and word boundaries for Myanmar text. The proposed method has two phases: word segmentation and part of speech tagging. Word segmentation is divided into syllable segmentation and syllable merging. In this proposed system, syllable segmentation is done by using the rules on syllable structure of Myanmar script. The next step was to merge the segmented syllables to determine word boundaries. Syllable merging is done by using forward and backward longest matching method and syllable merging rules. After the correct segmented word is achieved, the proposed method tries to tag correct syntactic class for the words in the text. In this case, searching and comparing with lexicon can achieve for multiple POS tag with for known word. And, it may also be unknown word which is not match with lexicon. To tag the correct single POS for both types of known and unknown words in the text, the proposed POS tagger applies the word disambiguation rules. These rules are created

based on the morphological structure of Myanmar grammar.

The rest of this paper is organized as follows. Section 2 describes about the related work of this system. Section 3 introduces Myanmar grammar. Section 4 presents about the Rule based POS tagging Section 5 explains the proposed POS tagger and section 6 describes conclusion and section 7 presents the future work.

2. RELATED WORK

In this section, previous works on word segmentation and part of speech tagging are reviewed. E.Brill presented a simple rule based tagger for English which performs as well as existing stochastic taggers, but has significant advantages over these taggers. A vast reduction in stored information are required, the perspicuity of a small set of meaningful rules as opposed to the large tables of statistics needed for stochastic taggers, ease of finding and implementing improvements to the tagger and better portability form one tag set or corpus genre to another [3]. H.H. Htay and K.N. Murthy described that Myanmar Word Segmentation using Syllable level Longest Matching. The authors have tested their algorithm on a 5000 sentence test data set containing a total of (35049 words) and manually checked for evaluating the performance. The program recognized 34943 words of which 34633 words were correct, thus giving them a Recall of 98.81%, a Precision of 99.11% and a F-Measure is 98.95% [4]. T.T. Thet and et.al described that Word segmentation for the Myanmar language. Their proposed strategy can be divided into two parts; one is rule based syllable segmentation and the other is dictionary based statistical syllable merging. They reported that their word segmentation achieved precision: 98.94%, recall: 99.5% and F-Measure:98.99% [8]. W.P. Pa and N.L. Thein experimented Disambiguation in Myanmar Word Segmentation. The authors described that word disambiguation accuracy is high when using forward and backward maximum matching with combined model. Precision of word segmentation for this approach is 92% and recall is 94% [9]. K.K. Zin and N.L. Thein described the Hidden Markov Model with Rule Based Approach for Part of Speech Tagging of Myanmar Language. The author compared two approach of tagging for Myanmar language, namely rule based approach and HMM model with rule based approach. According to the experimental result accuracy of HMM model with rule based approach is high [5].

3. MYANMAR GRAMMAR

Myanmar Language Commission standardized that it is composed of nine part of-speech in Myanmar grammar such as noun, pronoun, adjective, verb, adverb, post-positional marker, particle, conjunction and interjection. Two types of sentence in Myanmar language is simple and compound or complex [6]. Grammar rules are studied behind languages. The aspect of grammar, which does not concern directly with meaning, is called *syntax*. Myanmar language is mainly characterized as a SOV (subject, object and verb) language; it is also regarded as a free word order language which means that the part of speech of the word in the text can vary according to its position in the sentence.

One can usefully distinguish two types of SOV language in terms of their type of marking. The first, referred to in linguistic typology as dependent-marking, has case markers to distinguish the subject and the object. This type usually places adjectives and numerals before the nouns they modify and is exclusively suffixing without prefixes. SOV languages of this first type include Japanese and Tamil.

The second is head-marking and distinguishes subject and object by affixes on the verb rather than markers on the nouns. It also differs from the dependent-marking SOV language in using prefixes as well as suffixes, usually for tense and possession. In most SOV languages with a significant level of head-marking or verb-like adjectives, numerals and related quantifiers (like "all", "every") also follow the nouns they modify.

3.1. Myanmar POS Tagset

The language tagset represents part of speech and tagset consists on 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. There are totally 109 tagsets for according to different part of speech in Myanmar grammar.

Table 1. Example POS tagsets for Myanmar text

Myanmar POS	English Meaning	POS Tag	Example
ကတ္တားဝိဘတ်	Nominative	PONOM	သည်၊က၊မှာ
ကံဝိဘတ်	Objective	POOBJ	ကို
အကြောင်းပြဝိဘတ်	Reason	POREA	ကြောင့်၊ဖြင့်၊နဲ့
အများဆိုင်နာမ်	Common Countable Singular	NCCS	စာအုပ်၊ခဲတံ၊မြစ်
အများဆိုင်နာမ်	Common uncountable	NCU	ရေ၊ထမင်း၊ဆံပင်
ပေါင်းစပ်နာမ်	Compound Countable Singular	NCPS	ရေအိုး၊လက်စွပ်
လိင်ညွှန်းပစ္စည်း	N_Gender	PANG	မ၊ဒေါ်၊ဦး
ကိန်းညွှန်းပစ္စည်း	Indefinite Number	PAIDM	များ၊တို့
ပြောသူနာမ်စား	Speaker	PRSP	ကျွန်တော်၊ကျွန်မ
အညွှန်းနာမ်စား	Demonstrative	PRDEM	ဤ၊သည်၊ထို
အညွှန်းနာမဝိသေသန	Demonstrative	ADJDEM	ဤ၊သည်၊ထို
ပုဒ်ဆက်သမ္ဗန္ဓ	Word conjuntion	COW	နှင့်၊သို့မဟုတ်၊
ဝါကျဆက်သမ္ဗန္ဓ	Sentence	COS	၍၊လျက်၊လျှင်

Tagsets for noun can be subdivided into 12 sub groups which contain NN, NCCS, NCCP, NCU, NV, NCPS, NCPP, NCPU, NCLS, NCLP, NABQ, NABS. Tagsets for pronoun are divided into 10 categories which are PR, PRSP, PRL, PRO, PRPOS, PRDEM, PRI, PRNQ, PRNA and PRR. Tagsets for adjective can be classified into 7 types namely ADJ, ADJQ, ADJDEM, ADJQU, ADJIS, ADJA, and ADJI. Tagsets for adverb can be indentified into 9 items that contains ADV, ADVT, ADVM, ADVC, ADVQ, ADVI, ADVP, ADVTW and ADVR. Tagset for interjection is INT. Tagsets for postposition are composed into 19 subgroups such as POP, PONOM, POOBJ, PODEP, PODIR, POARR, POACC, POREA, POACCP, POPLA, POT, POPOS, POAGR, POCOM, POINT, POSEP, POTCOM, POPCOM and POVP.

Tagsets for particles include 28 kinds. They are PANK, PANG, PAIDNUM, PAQ, PAS, PASI, PANSU, PANSA, PANG, PANA, PANPC, PASDJC, PAADVC, PAVPC, APVPP, PAVP2, PAP2C, PAVP2P, PAVF, PAVFC, PAVFP, PAVN, PAVI, PAVN1, PAVTF and PAVS. Tagsets for conjunction contain 15 kinds which are CO, COW, COS, COM, COC, COT, COCS, COCP, COCA, COCB, COCH, COCON, COCOR, COSP and COSQ. Tagsets for others are constituted with 8 items namely SYM, SM, CN, ON, MM, DD, TT and FRA.

4. RULE BASED POS TAGGING

Rule based Tagging contains large database of disambiguation rules. It usually uses dictionary to assign each word list of potential POS. It also uses lists of hand-written disambiguation rules to cut list down to single POS for each word. Rules based on contextual information known as context frame rules. If unknown word X is preceded by determiner, followed by noun, tag it as adjective ('the yellow bird'). Another way is using rules based on morphological information. If unknown word ending in –s are likely to be plural ("pencils") [1].

The second is head-marking and distinguishes subject and object by affixes on the verb rather In addition to contextual information, many taggers use morphological information to aid in the

disambiguation process. One such rule might be: if an ambiguous/unknown word ends in an -ing and is preceded by a verb, label it a verb (depending on your theory of grammar) [1]. Some systems go beyond using contextual and morphological information by including rules pertaining to such factors as capitalization and punctuation. Information of this type is of greater or lesser value depending on the language being tagged [10].

4.1. Lexicon

Lexicon as a kind of expanded dictionary that is formatted so that a computer can read it (that is, machine readable) [1]. In this proposed system, lexicon is used to serve the training data for word segmentation and part of speech tagging. The lexicon used in this system is composed into nine POS tables according to different part of speech in Myanmar grammar. The lexical entries in these tables are stored as the alphabetically order used in Myanmar script. Moreover, there is some relationship between these nine POS tables. For example, noun and verb tables are connected has a one to one relationship to particle and postposition tables in the lexicon.

Adjective table is a part of lexicon. The attributes are ID, Myanmar Word, Tag, Type, Group, Kind and Remark. Attribute values of Kind are Opinion, Shape, Colour, Age and Material. Noun table contains the following fields such as ID, Myanmar Word, Tag, Type, Group and Remark. Attributes values for Group of noun contain People, Thing, Animal, Instrument, Timeday, TimeMonth, TimeSession, Place and None.

4.2. Word Disambiguation Rules

There are various ambiguities of POS in tagging for a word. To solve the POS ambiguities of the words in the text, the proposed system used some word disambiguation rules. These rules are based on contextual information known as context frame rules. It also used head-marking and distinguishes subject and object by affixes on the various part of speech. The proposed tagger is applied on morphological information in Myanmar grammar such as prefix, postfix and affix of the words to aid in the disambiguation process and unknown word. Here are some examples word disambiguation rules.

R1: Particle and Postpositional marker

a. If currentword == particle adv phrase change then previousword =adverb
b. If currentword==particle adj phrase change then previousword=adjective
c. If currentword==postpositinal verb then
d. previousword =verb
e. If currentword==adv degree two then nextword==verb
f. If currentword==noun then nextword=verb
g. If nextword==postpositionalverb then currentword=verb
h. If nextword==verb then currentword=adverb

R2: Noun

a. If currrentword==possessive postpositional then previousword=noun and nextword=noun
b. If currentword==particle noun phrase change then previousword=noun
c. If currentword==Postpositional possessive then nextword=noun
d. If nextword==postpositional objective then currentword=noun

R3: Conjunction and postpositional

a. Ifpreviousword==noun&&nextword==noun then currentword=word joint conjunction
b. If nextword==verb&&previousword==noun then currentword=postpostional accusative

R4: Time and Place postpositional marker

a. If wordtype==place then currentword=postpositional place
b. If currentword wordttype==time then word=postpositional time

R5: Verb

a. If nextword ==postpositional verb then currentword =verb present

b. If nextword==particle verb past then currentword=verb past
c. If nextword==particle verb present continuous"then currentword =verb present continuous
d. If nextword==particle verb future then currentword=verb future
e. If nextword==particle verb present perfect then currentword=verb present perfect

5. PROPOSED SYSTEM

This section is presented the proposed system of the processing steps for POS tagging.

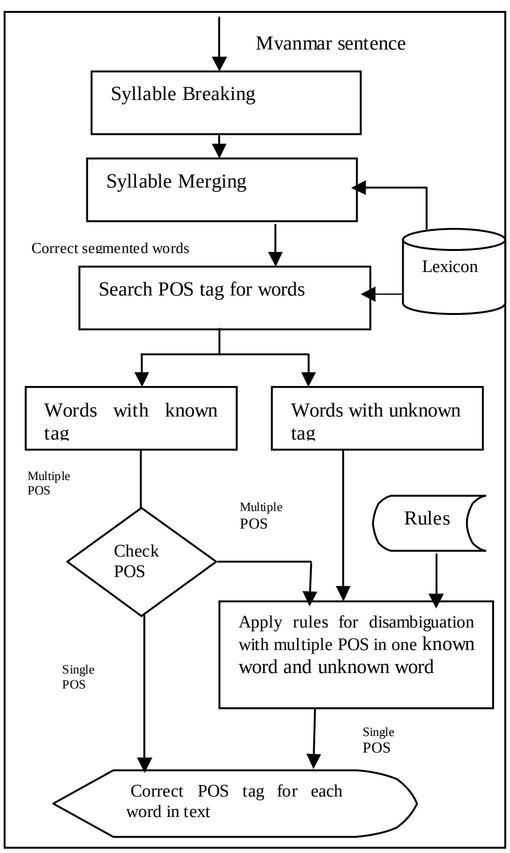

Figure 1. Procedure for POS tagging

Firstly, system accepts the Myanmar sentence and then this sentence is tokenized into words by using word segmentation rules. Word segmentation is divided into two steps which contain syllable breaking and syllable merging. Syllable breaking is the process of identifying syllable boundaries in a text. Syllable breaking rules are created based on the syllable structure of Myanmar script. A Myanmar syllable consists of one initial consonant, zero or more medials, zero or more vowels and optional dependent various signs. Independent vowels, independent various sign and digits can act as standalone syllables. After the segmented syllables are received, this proposed system merges these syllables into meaningful word by using forward and backward Maximum Matching approach. In maximum matching (MM), a character string is compared with the entries of a lexicon so that all the substrings constituting lexicon items are highlighted. The principal of Maximum Matching, or called longest match, is to find the best segmentation with the longest words among all possible substring chains. The algorithm starts from the beginning of a sentence, finding the longest matching word and then repeating the process until it reaches the end of sentence. When performing word segmentation stage, the next step is to assign the correct POS tag for the words in the text. There may be multiple POS ambiguities in one word and unknown words in the sentence. So, the tagger of this system is applied word disambiguation rules to assign the correct syntactic category of the word in the sentence. The above figure shows the procedure of the tagging process.

Accept the input sentence:

þ pmt kfo n f v Efygo n f

Break the sentence into syllable:

þ /pm/tkyf/onf/t /vGef/yg / onf/ /

Merge syllables into the word:

þ /pmtkyf/onf/t vGef/yg/ onf/ /

Multiple POS tagsets for each word in the sentence compared with a tagged lexicon:

þ	-	ADJDEM/PRDEM
pmtkyf	-	NCCS
onf	-	PONOM/POVP/PRDEM/ ADJDEM
tvGef	-	ADVQ
yg	-	VST/ADJQ/NCCS/PANK
onf	-	PONOM/POVP/PRDEM/ ADJDEM-
/	-	SM

Correct POS tag for each word by applying word disambiguation process:

þ - ADJDEM - pmtkyf- NCCS - onf - PONOM - tvGef- ADVQ-yg- VST- onf- POVP - / - SM

6. CONCLUSION

This paper presents an approach for a POS tagger by using rule based method to assign the correct syntactic category of each word for Myanmar text. The proposed tagger accepts the Myanmar sentence in the text. This sentence is tokenized into words by using word segmentation rules. Word segmentation is divided into two steps which contain syllable breaking and syllable merging. It also designs detail tagsets for Myanmar text. The proposed POS tagger uses word disambiguation rules to solve the multiple POS ambiguities in one word and unknown word. A tagged lexicon serves as the training data to obtain correct segmented word and POS tag for the words in Myanmar text. The POS tagger of this system can be properly tagged on both known and unknown words in the text. So, the proposed tagger of this system can give the promising accuracy for the part of speech of the words in Myanmar text. However, there are some limitations on this proposed tagger. It can't be solve to tag all the words in the Myanmar text. The limitation encountered in this proposed system can be categorized into five groups: missing words in the lexicon, proper nouns such as names of people and places, some adopted words in Pali (or other languages (Hindi, English), and some numerical words and also some negative words in the text.

7. FUTURE WORK

This research is an ongoing stage and also a part of developing Myanmar to English machine translation system. We hope to improve our system that can be assigned the correct syntactic categories of every word in the text within near future. We will try to implement our system by applying not only the rule based approach but also statistical approaches (Conditional Random Field, Naives Bayesian) to improve the accuracy of the tagger. Later, we will try to compare the accuracy of POS tagger by using different statistical approaches.

REFERENCES

[1] C. D. Manning, H. Schiitze," Foundations Of Statistical Natural Language Processing", The MIT Press, Cambridge, Massachusetts London, England, 2000.

[2] D. Jurafsky, J.H. Martin, "SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistic and Speech Recognition", Prentice-Hall, 2000.

[3] E. Brill, "A Simple Rule-Based Part of Speech Tagger", Department of Computer Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.PP152-155.

[4] H.H. Htay, K.N. Murthy, Myanmar Word Segmetation using Syllable Level Longest Matching, "Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages,pages 51-58,Hyderabad,India,Janauary 2008.

[5] K.K. Zin, N.L. Thein," Hidden Markov Model with Rule Based Approach for Part of Speech Tagging of Myanmar Language", 2009.

[6] Myanmar Grammar, Myanmar Language Commission, memorable for 30th anniversary, 2005.

[7] Myanmar Orthography, Second Edition, Myanmar language conmmisinon, 2003.

[8] T.T. Thet, J.C. Na, W.K. Ko, "Word segmentation for the Myanmar Languge", Journal of Information Science, 2007, PP. 1-17.

[9] W.P.Pa, N.L.Thein "Disambiguation in Myanmar Word Segmentation", "Proceedings Of the Seventh International Conference On Computer Applications", Yangon, Myanmar,2009,PP. 1-4.

[10] Y. Halevi, "Part of Speech Tagging ", Seminar in Natural Language Processing and Computational Linguistics, (Pro.Nachum Dershowitz), School of Computer Science, TeL Aviv University, Israel, April, 2006.

[11] Lexique Pro- Myanmar lexicon (Version-2), July, 2011.