A simple rule based Myanmar POS Tagger

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

Abstract

This paper presents a simple rule-based POS tagger to tag the correct syntactic categories of the Myanmar words by applying lexicon based word segmentation and heuristic rule based tagging method. Firstly, input sentence is tokenized into words by using syllable breaking and syllable merging with longest matching approach. Secondly, this system defines the detailed tag sets for POS tagging process by using nine different POS in Myanmar grammar. Finally, the proposed system solves the POS ambiguities of one word by applying word disambiguation rules which are generated from morphological features of Myanmar grammar. So, the proposed system can provide many benefits to Myanmar-English translation system and other NLP tasks.

Keywords: word disambiguation, morphological features

1. Introduction

Part of speech (POS) tagging is one of the most well studied problems in the field of Natural Language Processing (NLP). Part of speech (POS) tagging means assigning grammatical classes i.e. appropriate part of speech tags to each word in a natural language sentence [9]. POS tagging can be used in text to speech, information retrieval, shallow parsing , information extraction, word sense disambiguation in linguistics research for corpora [1] and also as an intermediate step for higher level NLP tasks such as parsing, semantics, translation and many more[1]. 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 of all possible tags and then uses context rules to disambiguate POS of the words in the text [2].

The rest of this paper is organized as follows. Section 2 describes about the related work of this system. Section 3 introduces Myanmar Language Section 4 presents about rule-base Myanmar POS tagger. Section 5 explains rule-based POS tagging. Section 6 describes the procedure of Myanmar tagger and Section 7 concludes this paper.

2. Related Work

In this section, previous works on word segmentation and part of speech tagging are reviewed. The authors in [4] showed that Myanmar word segmentation using syllable level longest matching. The authors showed Recall of 98.81%, a Precision of 99.11% and F-Measure is 98.95%.

In [8], word segmentation for the Myanmar language was presented. Their proposed strategy can be divided into two parts; one is rule based syllable segmentation and the other is dictionary based statistical syllable merging.

A simple rule based tagger for English which performs as well as existing stochastic taggers was described in [3].The perspicuity of a small set of meaningful rules as opposed to the large tables of statistics needed for stochastic taggers proved.

Hidden Markov model with rule based approach for part of speech tagging of Myanmar language showed in [5]. According to the experimental result accuracy of HMM model with rule based approach is high.

3. 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.

3.1. Myanmar Grammar

Myanmar Language Commission standardized that it is composed of nine part of-speech in Myanmar grammar such as noun, pronoun, particle, conjunction, adjective, verb, adverb, post-positional and interjection. The structure of sentence in Myanmar language may be simple and compound or complex [6].

Generally, sentence is subdivided into phrases. Phrase is subdivided into words. Word is subdivided into syllables. Syllable is the smallest unit of the language [6].

These are examples of simple and compound sentences.Simple: ဆရာက တပည့်များကို ပြောသည်။ Compound:ပါမောက္ခဦးဘသည်သားမောင်မောင်နှင့်အတူ အထက်မန္တလေးမှ အမြန်ရထားဖြင့် မနေ့မနက်က ရောရောမောမော ပြန်ရောက်လာသည်။

Types of Myanmar sentence can be distinguished into descriptive, negative, interrogative, command and opinion according to meaning [6]. The followings are the examples of all these five types.

Descriptive:တပ်မတော်သားတို့သည် နိုင်ငံတော်အတွက် အသက်ကိုပေးလှူကြသည်။

Negative: မောင်မောင်ကျောင်းမှ ပြန်မလာပါ။

Interrogative: ဘုရားပွဲကို သွားမလား။

Command: ခုံပေါ်မှာ မတ်တပ်ရပ်ပါ။

Opinion: လိုအင်ဆန္ဒတစ်လုံးတစ်ဝ ပြည့်ဝပါစေ။

3.2. POS Ambiguities on Myanmar Word

Myanmar language is mainly characterized as a SOV (subject, object and verb) language; it is also regarded as a free order of word in the sentence which means that the part of speech of the word in the text can vary according to its position in the sentence. Example: သူမ၏ပါးသည်နီရဲနေသည်။ဤစာအုပ်သည်အလွန်ပါးသည်။ကျန်းမာခြင်းသည်လာဘ်ကြီးတစ်ပါးဖြစ်သည်။By analyzing these example sentences, the word ပါး may be multiple POS such as noun or verb or particle. So, POS ambiguity of Myanmar word is a challenge to classify proper POS tag of the word in the text.

4. Rule based Myanmar Tagger

Part-of-speech tagging is an important part of Natural Language Processing (NLP) and is useful for most NLP applications. Therefore, it is necessary to have a tagger for the Myanmar language before developing other NLP activities.

4.1. Myanmar Word Segmentation

Word segmentation is the process of parsing concatenated text (i.e. text that contains no spaces or other word separators) to infer where word breaks exist.

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right. Burmese language users normally use space as they see fit, some write with no space at all. There is no fixed rule for word segmentation.

Thus, the computer has to determine syllable and word boundaries for Myanmar text. In this proposed system, word segmentation is divided into syllable segmentation and syllable merging.

Firstly, syllable segmentation is done by using the rules on syllable structure of Myanmar script. Myanmar script is composed of 33 consonants, 11 basic vowels, 11 consonant combination symbols and extension vowels, vowel symbols, devowelizing consonants, diacritic marks, specified symbols and punctuation marks.

And then, the next step was to merge the segmented syllables to determine word boundaries. Syllable merging is done by using the longest matching approach and mapped with lexicon. The algorithm starts from the beginning of a sentence, finding the longest matching word compared with lexicon and then repeating the process until it reaches the end of sentence.

In this case, longest matching cannot give the correct segmentation for all sentences. It can find segment conflicts for the word in some sentence. Example, သူမလိုင်အလွန်ကြိုက်သည်။With the longest matching approach, this sentence is segmented to wrong word into သူမ/ လိုင်/ အလွန်/ ကြိုက်/ သည် /။To avoid word segment conflict and then can improve the POS tagging quality, the assumption # symbol is used for the input sentence သူ#/ မလိုင် /အလွန်/ ကြိုက်/ သည်/ ။.

4.2. Myanmar POS Tag Sets

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

Table 1. Myanmar POS tag sets

POS	Kind	POS Tag set
Noun	11	NCCS, NCCP, NCU, NV, NCPS, NCPP, NCPU, NCLS, NCLP, NABQ, NABS
Verb	3	VAC, VST and VCP
Adjective	6	ADJQ, ADJDEM, ADJQ, ADJIS, ADJA, and ADJI
Adverb	8	ADVT, ADVM, ADVC, ADVQ, ADVI, ADVP, ADVTW and ADVR
Pronoun	9	PRSP, PRL, PRO, PRPOS, PRDEM, PRI, PRNQ, PRNA and PRR
Interjection	1	INT
Conjunction	14	COW, COS, COM, COC, COT, COCS, COCP, COCA, COCB, COCH, COCON, COCOR, COSP and COSQ
Particle	27	PANK, PANG, PAIDNUM, PADNUM, PAQ, PAS, PASI, PANSU, PANSA, PANGR, PANA, PANPC, PASDJC, PAADVC, PAVPC, PAVPP, PAVP2, PAP2C, PAVP2P, PAVF, PAVFC, PAVFP, PAVN, PAVI, PAVNU, PAVTF and PAVS
Postposition	19	PONOM, POOBJ, PODEP, PODIR, POARR, POACC, POREA, POACCP, POPLA, POT, POPOS, POAGR, POCOM, POINT, POSEP, POTCOM, POPCOM, POVP and POVP1
Other	9	UNK, SYM, SM, CN, ON, MM, DD, TT and FRA

This system is designed totally 107 tag sets for Myanmar text. These tag sets are designed according to Myanmar grammar book published by MLC. But, the proposed system cannot classify for all kind of Myanmar text both in

literary or written style used in formal and colloquial or spoken style.There are three types of verb namely into VAC=action, VST=state and VCP= compound. Action verb examples are ပြေး၊သွား၊စား၊ကန်. Some examples of state verb are ပျော်၊သနား၊ပူ and ကြင်နာ. Examples of compound verb contain ရေးသားစပ်ဆို၊ကာကွယ် and ရောင်းဝယ်.

The proposed system is designed to classify the correct POS tag of the word for structure and meaning in Myanmar text. Among them, the proposed system would like to describe some examples for POS in conjunction of Myanmar text. Normally, POS of conjunction can be often found in the simple and compound sentences.

In addition, conjunction in the sentence can act to join the two sentences or two phrases or two words.

"Examples of detailed tag sets for conjunction"

ဝါကျဆက်သွယ်မှုနှုန်း ⟶ COS နှင့်တစ်ပြိုင်နက်,လျှင်,စေရန်

မိုးသံ/NCPU/ကို/PONOM/ကြား/VAT/သည်/POVP /နှင့်တစ်ပြိုင်နက်/COS/လယ်သမား/NCCS/တို့/PAIDNUM/ သည်/PONOM/ထွန်တုံး/NCCS/ကို/POOBJ/ပြင်/VAC/ကြ/ PANK/သည်/POVP/။/SM/

ပုဒ်ဆက်သွယ်မှုနှုန်း ⟶ COW နှင့်၊သို့မဟုတ်၊မှတစ်ပါး၊

နှင်းဆီပန်း/NCCS/နှင့်/COW/စံပယ်ပန်း/NCCS/ကို /POOBJ/ဝယ်/VAC/ခဲ့/PAVP2/ပါ/POVP/။/SM/

အဓိပ္ပါယ်ဆက် ⟶ COM ထို့ကြောင့်၊ထို့ပြင်၊။သို့ရာတွင်၊

သတင်းစာ/NCCS/ထဲ၌/POPLA/ပြည်တွင်းပြည်ပသတင်း/ NCPU/များ/PAIDNUM/ပါ/VAC/သည်/POVP/။/SM/ ထို့ပြင်/COM/ဆောင်းပါး/NCCS/များ/PAIDNUM/လည်း/C OCB/ပါ/VAC/သည်/POVP/။/SM/

ဆန့်ကျင်ပြသမှုနှုန်း ⟶ COC သို့ရာတွင်၊သို့သော်လည်း

မိုး/NCU/ချုန်း/VST/သည်/POVP/။/SM/သို့ရာတွင်/COC/မိုး /NCU/မ/PAVN/ရွာ/VAC/ပေ/POVP/။/ SM

အကျိုးမျှော်ပြသမှုနှုန်း ⟶ COCS အောင်၊ရန်၊ဖို့၊ရန်အလို့ငှာ

လူ/NCCS/တော်/ADJQ/ဖြစ်/VAC/ရန်/COCS/ကြိုးစား/VA C/ရ/PAVS/မည်/POVP/။/SM/

နှိုင်းယှဉ်ပြသမှုနှုန်း ⟶ COCP သကဲ့သို့၊သလို၊ထက်

သူ/PRO/စာကြိုးစား/VCP/သကဲ့သို့/COCP/ကျွန်တော်/PRSP /ကြိုးစား/VCP/သည်/POVP/။/SM/။

ကန့်သတ်ချက်ပြသမှုနှုန်း ⟶ COCON လျှင်၊မှ၊က

သားသမီး/NCPS/လိမ္မာ/ADJQ/လျှင်/COCON/မိဘ/NCPS/ စိတ်ချမ်းသာ/VCP/မည်/POVP1/။/SM/

Tag set abbreviation and their corresponding meanings are shown in Table: 2.

Table 2. Example tag sets and abbreviation

POS	Tag Abbreviation and their meaning
Conjunction	COW=word joint, COS=sentence joint, COM= meaning, COC=concession , COCS=consequential, COT=time, COCP=comparison , COCA=cause, COCB=combination, COCH=choice, COCON=condition, COCOR= correlative, COSP= Spontaneous and COSQ= Sequence
Pronoun	PRO=objective , PRDEM=demonstrative, PRSP=speaker,
Noun	NCCS=countable singular, NABQ=abstract quality, NCU=uncountable, NCPU=compound uncountable, NCPS=compound countable singular, NCLS=collective singular,
Postposition	PONOM= nominative, POPOS= objective, POVP=simple verb,POVP1=future verb, POOBJ= objective, POPLA=place , POT=time, POARR=arrival,
Verb	VST=state verb, VAC=action verb, VCP=compound verb,
Other	SM=end marker, CN=cardinal number,
Particle	PANK= noun kind, PAVN=negative, PANNU=verb tense numeric, PAVPC= present continuous change , PAIDNUM=indefinite number, PAVP2=past simple, PAVS=verb special,
Adjective	ADJQ=quality, ADJDEM=demonstrative,
Adverb	ADVC=conditional

4.3. Word Disambiguation Rules

The goal of POS tagging is to indentify grammatical classes by inducing the relationship with adjacent and related words in a phrase of the text. There are various ambiguities of POS in tagging for a word according to the position in the sentence. To solve the POS ambiguities of the words in the text, the proposed system uses some word disambiguation rules. But, these rules can't be solved to every word of POS in Myanmar text.

These rules are based on contextual information known as context frame rules. Here are some examples word disambiguation rules.R1 is the general rule for all nine POS in text. R2 is generated for noun. But the others (R3, R4 and R5) are to solve the ambiguity of POS in one word.

R1: General Rule

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 previousword =verb

d.If currentword==adv degree two then nextword=verb

e. If currentword==noun then nextword= positional nominative

f. If nextword==POVP && previousword==verb then currentword = particle verb

g. If currentword==postpositional verb then previousword=verb

h. If currentword ==sentence end marker then previous word= the verb postpositional

i. If nextword==verb then currentword=adverb If nextword ==postpositional verb then currentword =verb present

j. If nextword==particle verb past then currentword=verb past

k. If nextword==particle verb present continuous" then currentword =verb present continuous

l. If nextword==particle verb future then currentword=verb future

m. If nextword==particle verb present perfect then currentword=verb present perfect

R2: Noun

a. If currrentword==possessive postpositional then nextword=noun

b. If currentword==particle noun phrase change then previousword=noun

c. If nextword==postpositional objective then currentword=noun

R3: Conjunction OR Postpositional

a. If previousword==noun&&nextword==noun then currentword=word joint conjunction

b. If nextword==verb&&previousword==noun then currentword= accusative postpostional

There may be POS ambiguities in one word based on the position in the sentence. Example: ရေနှင့်ကြာသည်လွန်စွာပနာရသည်။သူမသည်လက်ဖက်ခြောက်ကိုချိန်ခွင်နှင့်ချိန်သည်။The word နှင့် may be word joint conjunction or accusative postpositional word according to position.

R4: Time OR Place postpositional marker

a. If currentwordtype==place then currentword=postpositional place

b. If currentwordttype==time then currentword=postpositional time

R5: Adjective OR Verb

a. If currentword==verb postpositional then previousword=verb

b. If current word==noun then nextword=adjective

If nextword==noun then current word=adjective

4.4. Components of 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 required for word segmentation and part of speech tagging.

The lexical entries are stored as the alphabetically order used in Myanmar script and classified into nine different part of speech and morphological features in Myanmar grammar.

There are some assumptions of the Noun word. It contains the following fields such as ID, MyanmarWord, Tag, Type, Group and Remark, Prefix, Postfix, Affix. Attributes values for Group of noun contain People, Thing, Animal, Instrument, Timeday, TimeMonth, TimeSession, Place and None. Classification for Group attribute can provide many benefits to noun phrase identification, etc. Verb can be classified into action verb, state verb and compound verb which can be applied in the translation system for Myanmar-English. Example of lexical entries is described in table: 3.

Table 3. Example Entries in Lexicon

Mword	POS tag	Type	Group	Postfix
စာအုပ်	NCCS	Noun singular	Thing	
ပြေး	VAC	Verb Action	Intransitive	သည်
ပြောဆို	VCP	Verb compound	Transitive	သည်

5. 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].

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 [1].

6. Procedure of Myanmar Tagger

This section presents the processing steps of POS tagging. Firstly, the proposed system accepts the Myanmar sentence and then this sentence is tokenized into words by using word segmentation rules. The process of breaking a text up into its constituent tokens is known as tokenization. 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 based on Combining consonant and vowel, Devowelising and Kinzi, Contractions, Syllable Chaining, Distinct Letter, Single Character and Loan Words.

After the segmented syllables are received, the proposed system merges these syllables into meaningful word by using longest matching approach.

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 mapping with lexicon can achieve multiple POS tags for known word in the text. And, it may also be unknown words which is not match with lexicon. To tag the correct single POS tag for both types of known and unknown words in the text, the proposed POS tagger applies the word disambiguation rules. The process of POS tagging is shown in

Figure: 1.The following steps are the detail process of this system.

Step1: Accept the input sentence:

သူမ၏ပါးသည်နီရဲနေသည်။

Step2: Break the sentence into syllable:

သူ/မ/၏/ပါး/သည်/နီ/ရဲ/နေ/သည်။

Step3: Merge syllables into the word: သူမ/၏/ပါး/သည်/နီရဲ/နေ/သည်။

Step4: Receive multiple POS tag sets for each word in the sentence mapped with a tagged lexicon:

သူမ- PRO

၏- POPOS

ပါး- NCCS/ADJQ/VST/PANK

သည်- PONOM/POVP/PRDEM/ADJDEM

နီရဲ- ADJQ/VST

နေ- PAVPC/NCCS

သည်- PONOM/POVP/PRDEM/ADJDEM

။- SM

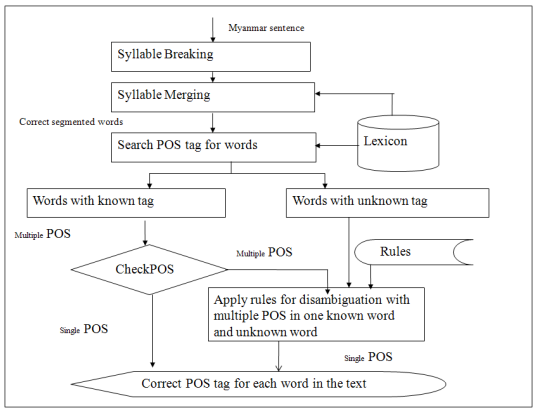

Figure 1. Process of POS tagging

Step5: Output the correct POS tag for each word after applying word disambiguation rules.

In this case, the word ပါး may be multiple POS tags. According to word disambiguation rules which are generated in section 4.3. From R2 for Noun, the proposed system looks up current word and its type is possessive postpositional. So, the tagger finally outputs the word ပါး into NCCS. And then, the word သည် can also be PONOM/POVP/PRDEM/ADJDEM. By using R1: General Rule, the first word သည် in the sentence is PONOM because of the current word is noun. And then the second သည် can classify into POVP because of the current word is sentence end marker. According to R1, the word နေ can be assigned into verb particle PAVPC by looking at the previous and next word in the sentence. The word နီရဲ may be ADJQ or VST from lexicon result. But the proposed tagger may tag as VST mapping with R5.

သူမ/PRO/၏/POPOS/ပါး/NCCS/သည်/PONOM/နီရဲ/VST/နေ/PAVPC/သည်/POVP/။/SM

ဖွဲ့/NABQ/ရှိ/VAC/က/COCON/အောင်မြင်/VST/သည်/POVP/။/SM/

6.1. Expected results of the system

The accuracy of the POS tagger is evaluated in Figure 2. Totally 200 (both type of simple and compound) sentences are collected from Myanmar grammar book published by MLC and are manually tested. Among them, the proposed system can correctly tag to the 50 sentences for disambiguation of multiple POS in one word and 20 sentences for unknown word.

The accuracy results of the proposed system can improve by adding more word disambiguation rules to classify every word of the POS in the text. Figure shows that known word accuracy is the 82.67%, average accuracy is 81.01% and unknown word accuracy is 79.35%.

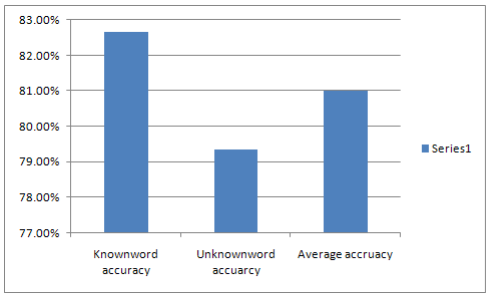

Figure 2. Accuracy analysis on POS tagging

7. Conclusion

An approach for Myanmar POS tagger is presented by using rule based method. The proposed tagger accepts the Myanmar sentence in the text. This sentence is tokenized into words by using syllable breaking rules and syllable merging with longest matching method. It also mapped with lexicon to achieve POS tags of the words. In addition, the proposed system tries to solve the POS ambiguities in one word and also for unknown word in the text by applying word disambiguation rules and morphological features of Myanmar grammar. The POS tagger of this system can be properly tagged on both known and unknown words in the text.

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 Segmentation 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, Yangon, 2005.

[7] Myanmar Orthography, Second Edition, Myanmar Language Commission, Yangon, 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] 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.

[10] Lexique Pro- Myanmar lexicon (Version-2).