Disambiguation using joint entropy in part of speech of written Myanmar text

Sin Thi Yar Myint & G. R. Sinha

International Journal of Information Technology
An Official Journal of Bharati Vidyapeeth's Institute of Computer Applications and Management

ISSN 2511-2104
Volume 11
Number 4

Int. j. inf. tecnol. (2019) 11:667-675
DOI 10.1007/s41870-019-00336-4

Springer

Int. j. inf. tecnol. (2019) 11:667–675
https://doi.org/10.1007/s41870-019-00336-4

ORIGINAL RESEARCH

Disambiguation using joint entropy in part of speech of written Myanmar text

Sin Thi Yar Myint[1] · **G. R. Sinha**[2]

Received: 17 November 2018 / Accepted: 29 July 2019 / Published online: 19 August 2019

Abstract Myanmar language also known as Burmese language is a free order of word language and syntactic patterns of one word can vary based on the position and the structure in the sentence. There are many ambiguous part of speech (POS) tags on one word in the sentence of Myanmar text. This research work presents disambiguation for the POS of the words in written Myanmar text. We aim at removing this ambiguity on Myanmar word and assigning single POS to each word of sentence. This is demonstrated on the following ideas: (i) input the sentence and segmented into words using syllable segmentation rules and forward maximum matching approach with monolingual Myanmar dictionary and (ii) apply the Joint Entropy (JE) for POS ambiguous for each word in the sentence with monolingual Myanmar tagged corpus. Joint probability value could be given the useful and accurate disambiguation of POS for free order and structure of words in Myanmar text. The monolingual Myanmar tagged corpus and tagged dictionary are created including 620 sentences and 15,000 words, respectively. This study attempts practical word segmentation and POS tagging system which can really overcome bottleneck of the machine translation system for Myanmar to other languages and research activities related to natural language processing (NLP).

✉ Sin Thi Yar Myint
cynthiamyint@gmail.com; sin_thi_yar_myint@miit.edu.mm

G. R. Sinha
gr_sinha@miit.edu.mm

[1] Faculty of Computer Science, Myanmar Institute of Information Technology, Mandalay, Myanmar

[2] Myanmar Institute of Information Technology, Mandalay, Myanmar

Keywords Part of speech · Joint entropy · Natural language processing · Myanmar language

1 Introduction

The nature of Myanmar language is completely different with the nature of English and other western languages. There are many word combinatory and overlapping ambiguity problems when tokenizing the syllables or words from the sentence because Myanmar script has no fixed delimiters between words or syllables. The structure of Myanmar sentence is subject, object and verb. There are many ambiguous POS tags on one word in the sentence of Myanmar text. POS tagging for the Myanmar text is the most important and main NLP tasks of Myanmar Language. There is different word segmentation and POS tagging methods which are proposed by many researchers but still needed the best and complete method for this task. Myanmar POS disambiguation is also the special issue for machine translation system and other preprocessing tasks of NLP. POS ambiguity is a kind of syntactic ambiguity that arises not from the range of meanings of single words, but from the relationship between the words and clauses of a sentence, and the sentence structure underlying the word order therein [20]. In addition, machine translation is a subfield of computational linguistics that investigates the use of software to translate text or speech from one language to another [5]. Part of speech tagger using hidden Markov model was applied by researcher [2] as the initial study for POS tagging. Brent studied about models for word and speech segmentation using these three techniques the utterance-boundary strategy, the predictability strategy, or the word-recognition strategy. Selected predictions of the models are explained and compared the performance for

computation [1]. Therefore, we studied about POS ambiguity solving problem for Myanmar text.

In this research, a small monolingual dictionary and a small monolingual corpus which contains detail POS tag of the word are created because there have no linguistic resources appropriate with this study. This system applied on the POS tags of the word in the sentence on 94 tag sets for Myanmar text which was already published by Myint et al. [14].These tag sets are designed according to Myanmar grammar book published by Myanmar Language Commission to the detail structure and meaning of Myanmar Grammar. But, the proposed work cannot classify for all kind of Myanmar text both in literary or written style used in formal and colloquial or spoken style. It mainly focuses on written style of text. This study was implemented the POS tagging model by using statistical approach of JE method with corpus. This study applied the power of joint probability value in solving POS disambiguation of words in the sentence.

2 Related work

We have studied a number of research papers, thesis, articles, databases, etc., in the field of Myanmar language, scripts and others. This section reports findings and weakness of major research contributions in this research area.

Win et al. [22] proposed that it was easier to segment words after decomposing the phrases of sentence. The system had been tested by developing a phrase segmentation system using CRF++. Although scores were highly efficient, authors faced some difficulties could not solve because of limited lexical resources. They had been tested phrase segmentation of various types of corpus with 5000 and 50,000 phrase-model of general corpus, respectively. Average scores of phrase segmentation are above 70% according to the F-Measure. The corresponding scores were tested upon four different corpus types. In the corpus type of sport was shown F-Measure 83%. The corpus type of newspaper was achieved F-Measure 72%. The corpus type of general F-Measure was tested 70% and in the corpus type of novel was achieved F-Measure 62%.

Nam et al. [17] studied for the solution of Khmer language based on the syllable division into component clusters applied by two syllable models. Authors used a lexical database which was collected from the online Khmer dictionaries and supported dictionaries serving role of training data and complementary linguistic characteristics. Each component cluster is labelled and located by the first and last letter to identify entirety a syllable. This approach proves that it can workable and the test results achieve high accuracy, eliminate the ambiguity, contribute to solving the problem of word segmentation and applying efficiency in Khmer language processing and it achieved high precision of test results for simple word, compound word, phrase are 95%, 92.5% and 89.5%, respectively. Nguyen et al. [18] proposed the word segmentation and part of speech tagging using two methods namely pipeline and joint. Pipeline strategy displays segmented the words first and then applied as into part of speech tagging step. In joint method, authors predicted a combined segmentation and POS tag for each syllable and then compared with state-of-the-art (SOTA) feature based and neural network-based models. On the benchmark Vietnamese Treebank (Nguyen et al. 2009), experimental results show that the pipeline strategy produces better scores of POS tagging (93.39%) from unsegmented text than the joint strategy (87.53%), and the highest accuracy is obtained by using a feature-based model.

Tedla et al. [21] proposed first part-of-speech (POS) tagging research for Tigrinya (Semitic language) from the newly constructed Nagaoka Tigrinya Corpus. Data are collected from a newspaper published in Eritrea in the Tigrinya language with formatted in plaintext and the Text Encoding Initiative (TEI) XML format. A tag set of 73 tags was designed, and the corpus for POS was manually annotated. This tag set defines as three levels of grammatical information, which contain the main POS categories, subcategories, and POS clitics. The POS tagged corpus contains 72,080 tokens. The authors used the conditional random fields (CRFs) and support vector machines (SVMs) methods to get the correct POS tags. For a reduced tagset of 20 tags, an overall accuracy of 90.89% was obtained on a stratified tenfold cross validation. Enriching contextual features with morphological and affix features improved performance up to 41.01% point. Ding et al. [3] studied about the experiments on various word segmentation approaches for the Burmese language. The authors described that there are dictionary-based, statistical, and machine learning approaches for Myanmar word segmentation are tested. According to the experiment results describe that that statistical and machine learning approaches perform significantly better than dictionary-based approaches on an annotated corpus of relatively considerable size (containing approximately a half million words and discovered the properties and proper approaches to Burmese textual processing and to promote further researches on this understudied language. Phyu et al. [19] studied about the Burmese (Myanmar) word segmentation using character clustering and CRF methods. Characters are clustered into groups of some inseparable characters due to language characteristics. Authors proposed rules for a set of 29 types of Burmese Character, and Conditional Random Fields is applied as a sequential labelling machine learning method. According to the comparison results, the

proposed method with CRF without BCC and Syllable-based CRFs and proves that it hits the highest. Myint [15] proposed a new approach for POS tagging of using rule based and probabilistic approach try to solve ambiguities tag sets for Burmese texts. The tag sets for Myanmar POS, segmentation rule, tagging algorithm and CRF method are designed. The proposed approach is used and tested on Bilingual lexicon. The accuracy result is above 89.53% for known words, 79.89% for unknown words. Cutting et al. [3] presented an implementation of a part-of-speech tagger based on a hidden Markov model. The methodology enables robust and accurate tagging with few resource requirements. Only a lexicon and some unlabeled training text are required. Accuracy exceeds 96%. Based on study of a number of prominent research papers, a critical review reported above highlights findings and suitable applications. Word segmentation and POS tagging are the necessary tasks for Myanmar NLP activities. Lack of appropriate lexical resources such as corpus, lexicon and dictionary and lack of robust approach for POS disambiguation are observed as major challenges in existing work. Table 1 highlights summary of recent research in the area of word segmentation and POS tagging.

Many researchers, scholars and professionals have been attempting to achieve the complete NLP system for Myanmar language from different point of views and perspectives. These research works concerned with Myanmar word segmentation and POS tagging have been made but better techniques and algorithms are still required for Myanmar word segmentation and POS tagging. A practical Myanmar word segmenter and POS tagger which can be solved the bottlenecks of Myanmar language that have not been done yet. The challenges of statistical POS disambiguation for Myanmar language have become the fact that Myanmar has free order of word in sentence structure and a complex morphological system. NLP tools and applications for Myanmar language also need to keep up the NLP technologies in information and communication technology world.

3 Myanmar language

Myanmar language is the official language of the Union of Republic of Myanmar (formerly known as Burma) and is more than 1000 years old. Texts in the Myanmar (Burmese) language use the Myanmar script, which is descended from the Brahmi script of ancient South India. Myanmar is a country having a population of about 54,219,768 [11] people comprising of 135 ethnic groups speaking their own vernacular languages or dialects. Myanmar language still remains as one of the less privileged Asian languages in cyberspace. Many people have put considerable effort into the computerization of the Myanmar script [7]. Nowadays, Myanmar still lacks support on computers and not many NLP tools and applications are available for this language [13, 17]. And then, Myanmar Language Commission (MLC) standardized that it is composed of nine parts of-speech in Myanmar grammar such as noun, pronoun, adjective, verb, adverb, post-positional marker, particle, conjunction and interjection [10]. Moreover, there are many detail POS on the above general nine grammar according to their structure and meaning. By studying and specializing upon these detail grammar structures we can achieve the correct translation of Myanmar from other languages. This may take many advantages to the machine translation system for Myanmar to other Languages. In addition, types of Myanmar sentence can be distinguished into descriptive, negative, interrogative, command and opinion according to meaning. The language tag sets represents part of speech. According to contextual and morphological structure, natural

Table 1 Summary of different Word Segmentation and POS tagging methods

Categories	Year	Method	Training data
Word Segmentation	Ding et al. (2016)	Dictionary-based, statistical, and machine learning approaches	Approximately a half million words
Phrase Segmentation	Win et al. (2011)	CRF++	5000 and 50,000 Myanmar phrases general corpus
Word Segmentation	Nam et al. (2017)	Syllable division into component clusters applied by two syllable models	Khamer Dictionaries
POS Tagging	Cutting et al. (1999)	Phrase recognition; word sense disambiguation; and grammatical function assignment	Hidden Markov Model with Lexicon
POS Tagging	Nguyen et al. (2017)	Pipe line and Joint (neural network-based)	Vietnamese Treebank
POS Tagging	Tedla et al. (2016)	the conditional random fields (CRFs) and support vector machines (SVMs)	Nagaoka Tigrinya Corpus

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. Two types of sentence in Myanmar language is simple and compound or complex [6].Generally, sentence is classified into phrases. And then, phrase can be subdivided into words. Finally, word can be partitioned into syllables. Syllable is the smallest unit of the language [6].

3.1 Example categories of Myanmar script

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right. Myanmar characters can be classified into three groups: consonants, medials and vowels. Medials are subscript characters which can modify the basic consonants to form hundreds of consonants. Figure 1 shows different categories of Myanmar script which contains consonant, medials, vowels (compound and independent, punctuations, digits and symbols). Myanmar script has consonants, vowels (attached and free standing), diacritics, medials, a vowel killer or asat, digits and punctuation marks. Myanmar is a tonal language and is syllable-based, also an abugida. Spaces are used to separate phrases, rather than words. Words are composed of syllables. These start with a consonant or initial vowel. An initial consonant may be followed by a medial consonant. After the vowel, a syllable may end with a nasalisation of the vowel or an unreleased glottal stop. At the end of a syllable a final consonant usually has an 'asat' sign above it, to show that there is no inherent vowel. In multisyllabic words derived from an Indian language such as Pali, where two consonants occur internally with no intervening vowel, the consonants tend to be stacked vertically, and the asat sign is not used. There are a set of Myanmar numerals, which are used just like Latin digits [7–9, 13, 14].

3.2 POS ambiguities in Myanmar word

Myanmar language is mainly characterized as a SOV (subject, object and verb) language; but it is also regarded as a free order of word language which means that the part of speech of the word in the text can vary according to its position in the sentence. There are many POS ambiguities of words in sentence [10]. Part of speech of the Myanmar word can have many POS ambiguity based on the position in the sentence.

In Myanmar, words are mainly classified into two categories: independent word and helping word. An independent word represents a word which can convey its meaning itself without depending on another word (MLC 2005). Nouns, pronouns, adjectives, verbs, adverbs and interjections are used as independent words. A word used for helping in conveying the meaning of another word is called a helping word or an auxiliary. Postpositions, conjunctions and particles can be used as helping words or auxiliaries. According to their formation, another classification of words can be made: simple word, compound word, particle-affixed word and postposition-joined word. A word that can stand alone and convey its meaning without the help of affixing any postposition or particle is called a simple word. A word formed by the combination of two three independent words without using any postposition or particle or conjunction is called a compound word. A word formed by the combination of one independent word and a helping word or an auxiliary is called a

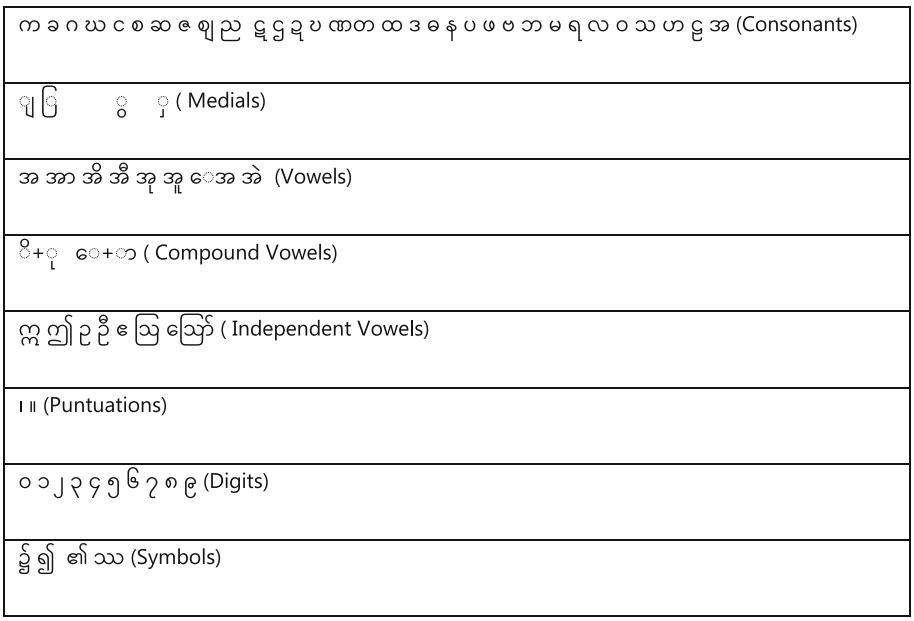

Fig. 1 Catagories of Myanmar script

သူမသည် ပျောက်**သွား**သောလက်စွပ်ကို မတွေ့မချင်းရှာ၏။[**သွား(have been—Particle)**—PAVS, auxiliaries] She searches the ring that **has been** lost until she finds.
သူမသည် သူမ၏ **သွား**ကိုတိုက်သည်။ [**သွား(tooth--noun)**---NCCS, singular countable noun] She brushes her **tooth**.
ကျွန်တော်သည် မန္တလေးသို့ မနက်ဖြန်**သွား**ဖြစ်မည်။ [**သွား(go--verb)**---VAC, action verb] I will **go** to Mandalay tomorrow.

Fig. 2 Examples of sentences for POS ambiguous words

particle-affixed word. A word joining with a postposition is called a postposition-joined word [6]. The following sentences are example ambiguous POS words for the sentences based on the position and structure. Figure 2 shows few examples of sentences for POS ambiguous words.

By analyzing the above example sentences, the Mynamar word "သွား" can have multiple parts of speech 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 sentence according to their morphological structure and position in the sentence. The process of correct and meaningful translation from Myanmar Language to Others is still a difficult task. Most of the NLP researchers and scholars for Myanmar Language are trying to find the solutions for this problem.

4 Architecture of the system

The system architecture of the proposed work that includes segmentation of Myanmar words and POS tagging process are shown in Fig. 3. Firstly, input sentence is accepted by the system and broken into mono-syllables by using the breaking rules based on the Myanmar syllable structure highlighted in Fig. 1. These syllables consist of one initial consonant, zero or non-zero medials and vowels; auxiliary number of depended signs. Number of independent vowels, symbols, various signs and digits can be considered as standalone syllables.

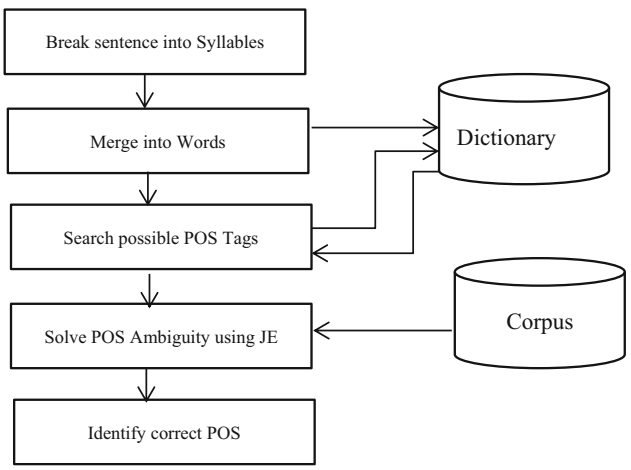

Fig. 3 Architecture of the POS tagging

Secondly, the corrected segmented syllables were merged into meaningful word by using forward longest matching approach and monolingual dictionary. There may be multiple POS ambiguities in one word in the sentence. This program had been searched and retrieved all possible POS tags for the words in the sentence compared with monolingual dictionary. So, we need to define correct single POS tag for the words in the input sentence. Thirdly, the POS ambiguities of words in the input sentence had been defined by the program. In this case, this system was used by JE statistical approach with monolingual tagged corpus to solve the ambiguities of words for the input sentences. Finally, this study had successfully identified the single correct POS tag for the specific words in the input sentence.

5 Lexicon, corpus structure and tag sets

The example of monolingual lexicon structure is shown as below in Table 2. It contains Myanmar words and possible POS tags for one word. Currently it has 15,000 words for Myanmar Language. These words are collected from the words of various Myanmar dictionaries and the words from Myanmar grammar book published in 2005 [10]. These words are stored as word with POS tag sets.

The following Fig. 4 shows the structure of the monolingual corpus in this study. Currently, this corpus is trained with different POS patterns of Myanmar sentences. It has 620 sentences of tagged sentences are trained from Myanmar Grammar book published by MLC and Myanmar newspapers.

Table 2 Monolingual Lexicon Structure

Mwords	POS tag
"သွား"	NCCS
"သွား"	VAC
"သွား"	PAVS

သူမ_/pronoun_PRO#မီးရထား_/noun_NCCP#ဖြင့်_/preposition_POACC#အိမ်_/noun_NCCS#သို့_/preposition_PODIR#သွား_/verb_VAC#သည်_/postposition_POVP#~,သူ_/pronoun_PRO#သည်_/none_PONOM#ကျောင်း_/noun_NCCS#သွား_/verb_VAC#သည်_/postposition_POVP#~,ဤ_/adjective_ADJDEM#စာအုပ်_/noun_NCCS#အလွန်_/adverb_ADVQ#ပါး_/verb VST#သည် /postposition POVP#~,

Fig. 4 Monolingual corpus structure

Table 3 Example POS tag sets

Sr.no	English Meaning	Tagsets	Example Myanmar Words
1	Adj-Descriptive Quality	ADJQ	လိမ္မာသော၊အေးသော၊ယူသော
2	Sentence Conjunction	COS	လျက်၊လျှင်၊သကဲ့သို့၊တစ်ဖြိုင်နက်
3	Verb Action	VAC	ခြေး၊ကန်၊ခွဲ၊ရိုက်
4	Adverb Time	ADVT	ယနေ့ ၊မနက်ဖြန်၊ ယခု

The following Table 3 shows that the examples tag sets among the 94 detailed POS tag sets. These tag sets are referenced on detailed 94 POS tag sets which was already proposed POS tag sets by Myint et al. [16]. Detail POS tag sets for Noun are define as proper noun, common countable singular, common countable singular, common uncountable, verbal noun, compound countable singular, compound uncountable singular, collective singular, abstract quality and abstract state and so on from the Myanmar Grammar book published in 2005. The system applied the power of detail POS tags sets for Myanmar Language to support the correct translation for other NLP tasks and activities.

6 POS tagging process

The POS tagging process consists of different modules which perform different functionalities to achieve better accuracy of POS tagger. The sentence is segmented into syllables by using syllable breaking rule in based on syllable structure of Myanmar script. After that, the resulted syllables are merged into words by using forward maximum matching approach. And then, the meaningful segmented output words can be assigned into multiple POS tags according to POS tagging algorithm which are retrieved from monolingual lexicon also it applies the disambiguation model with joint entropy method combined with monolingual corpus for known words. And finally, it displays the appropriate correct tag for each word of input text. The following steps are the example process of this system.

Breaking into syllables The system accepts the input sentence with pyidaungsu font [12] and breaks them into syllables.

Step 1: Accept the input sentence

သူမသည်ပျောက်သွားသောလက်စွပ်ကိုမတွေ့မချင်းရှာ၏။

English meaning for this Myanmar sentence is “She searches the ring that has been lost until she finds”.

Step 2: Break the sentence into syllable:

သူ/မ/သည်/ပျောက်/သွား/သော/လက်/စွပ်/ကို/မ/တွေ့/မ/ချင်း/ရှာ/၏/။

Segmentation into words

Step 3: These syllables are merged into the word using forward longest matching method retrieves from monolingual dictionary.

သူမ/သည်/ပျောက်/သွား/သော/လက်စွပ်/ကို/မတွေ့မချင်း/ရှာ/၏/။

Extraction of Multiple POS Tag sets for one word

Step 4: In this case, the system receives multiple POS tag sets for each word in the sentence by matching with a tagged monolingual dictionary.

သူမ -> PRO, သည် -> PONOM, သည် -> POVP, သည် -> PRDEM,သည် -> ADJDEM, ပျောက် -> VAC, သွား -> NCCS, သွား -> VAC, သွား -> PAVS, သော -> PAADJC, သော -> ADJQ, လက်စွပ် -> NCPS, မတွေ့မချင်း-> ADVM, ကို -> POOBJ, ကို -> PANG, ရှာ -> VAC, ရှာ -> PAVS, ၏ -> POPOS, ၏ -> POVP

Solving POS disambiguation using JE

This system solves POS ambiguity of the words in sentence by using disambiguation JE model described in equation: 1. JE model of this system was taken the joint probability of other words from the sentence to get the maximum probability of POS tag. This joint probability values can solve the POS ambiguous problems for the words in the sentence. The joint Shannon Entropy of two variables X and Y are defined as Where, x and y are particular values of X and Y, respectively, is the probability of these values occurring together, and $P(x, y) \log_2[P(x, y)]$ is defined to be 0 if $P(x, y) = 0$ [4].

$$H(X, Y) = -\sum_{x}\sum_{y} P(x, y) \log_2[P(x, y)] \quad (1)$$

POS ambiguity on one word in the sentence is solved and can produce the correct single POS tags for the words in the sentence.

Table 4 Precision of the experiment results of system

Type	WS	WA	NCA	NTA	Precision
Simple	15	5	480	500	96.00%
Compound	8	5	107	120	89.16
Both	23	10	587	620	94.67%

Table 5 Recall of the experiment results of system

Type	WS	WA	NCA	NTA	Recall (%)
Simple	15	5	480	500	95.83
Compound	8	5	107	120	87.85
Both	23	10	587	620	94.37

သည် -> PONOM, သွား-> PAVS, ကို ->POOBJ, ရှာ-> VAC, ၏->POVP

Display Single POS tag sets for one word

The system finally assigned into the correct and precise single POS tag of the every word in the sentence.

သူမ/PRO/သည်/PONOM/ပျောက်/VAC/သွား/PAVS/သော/PAADJC/လက်စွပ်/NCPS/ကို/POOBJ/မတွေ့မချင်း/ADVM/

ရှာ/VAC/၏/POVP/

7 Experimental results on part of speech tagging

The experiment results of Precision, Recall and F-Measure are described in Tables 4, 5 and 6. This study was tested upon the both types of simple and compound sentences of written Myanmar text. According to the experimental results, accuracy results of simple sentences had the highest in Precision, Recall and F-Measure which are calculated upon Eqs. 2, 3 and 4. The compound sentences had the lowest accuracy. But the accuracy results of both types of sentences proved the medium accuracy results. The accuracy of the POS tagger was evaluated on both types of simple and compound 620 sentences of Myanmar text which were collected from Myanmar grammar book published by Myanmar Language Commission. The accuracy results of the system could be improved by adding more words and example sentences into the corpus.

In above Tables 4, 5, and 6, WS is wrong sentences; WA indicates wrong ambiguity words; NCA infers number of correct ambiguous sentences by the program on the input and NTA is number of total sentences verified manually.

Table 6 F-Measure of the experiment results of system

Type	WS	WA	NCA	NTA	F-Measure (%)
Simple	15	5	480	500	95.91
Compound	8	5	107	120	88.50
Both	23	10	587	620	94.51

Now, we also can calculate

$$precision = \frac{no.of \text{ correct amb sentences}}{no.of \text{ total correct amb sentences}} \times 100 \quad (2)$$

[သွား NCCS, သွား VAC,သွား],[PAVS,ပါး NCCS,ပါး],[VST,သည် POVP,သည် PONOM,သည် ADJDEM],[သို့ PODIR,သို့ POARR],[က PONOM,က VAC,က PODEP,က POPLA,က POT,က COW],[တွင် NCCS,တွင် POSEP,တွင် POPLA,တွင် POT,တွင် VAC], [ဖြင့် POACC,ဖြင့် POREA],[နှင့် POACC,နှင့် COW,နှင့် POVP],[၏ POPOS,၏ POVP], [နေ PAVPC,နေ NCCS,နေ VAC],[တို့ PAIDNUM,တို့ PRSP,တို့ VAC],[ဖြစ် VST, ဖြစ် PAVS],[ရန် NCCS,ရန် COS], [အောင် VAC,အောင် COS, အောင် POINT],[များ PAIDNUM, များ VST],[မှ COS,မှ PODEP,မှ PAVS],[လေ NCCS,လေ PAVS, လေ VAC,လေ POVP], [တက် VAC,တက် NCCS,တတ် VST,တတ် PAVS],[တိုင်း PANA, တိုင်း COS],[ဖို့ COS,ဖို့ POINT,ဖို့ VST],[ကိုယ့် PRPOS, ကိုယ့် ADJDEM], [မှာ POPLA,မှာ POT,မှာ PONOM,မှာ VAC], [ထက် VAC,ထက်

Fig. 5 Example POS ambiguous words solved by the system

$$\text{Re}call = \frac{no.of \text{ corrected amb sentences by program}}{no.of \text{ total correct amb sentence}} \times 100 \quad (3)$$

$$F-Measure = \frac{2 \times \text{Pr}ecision \times \text{Recall}}{\text{Pr}ecision + \text{Recall}} \quad (4)$$

There are so many POS ambiguities in one Myanmar word. So, it is a difficult and challenge task to assign the correct POS tags for one word. The following words are the some findings and examples of POS ambiguity words studied by this system as mentioned in Fig. 5. This study tested upon 73 kinds for POS tags out of 94. The research tested and solved 195 POS ambiguities pattern with 620 Myanmar sentences.

8 Conclusion

This system is the initial study of the Myanmar word segmentation and part of speech tagging using maximum forward word segmentation using monolingual dictionary and POS disambiguation with JE model using monolingual tagged corpus. Although the scope of evaluation experiment was limited, it proved that this research worked properly to solve the ambiguities of POS in Myanmar words in sentence. The Precisions, Recall and F-measure results of the system are above 80% on the simple and compound sentences for Myanmar text and it can prove the promising tagging accuracy. So, many problems solving issues still exist to complete this study. So, there is a need to try to build the linguistic resources of tagged dictionary and corpus suitable for this study by adding more and more ambiguity patterns of words and sentences from different domains of Myanmar text. The accuracy results of the system could be improved by adding more words and example sentences into the corpus. Although the scope of evaluation experiment was limited, it proved that this research was worked properly to solve the ambiguities of POS in Myanmar words in sentence. Although the experiments have been performed on a very small lexicon and corpus, the results had shown that the use of a composite approach and heuristics improves the accuracy of the tagger.

References

1. Brent MR (1999) Speech segmentation and word discovery: a computational perspective. Trends Cogn Sci 3(8):294–301
2. Cutting D, Kupiec J, Pederson J, Sibun P (1992) A practical part-of-speech tagger. In: Proceedings of the 3rd conference on applied NLP, pp 133–140
3. Ding C, Thu YK, Utiyama M, Sumita E (2016) Word segmentation for Burmese (Myanmar). ACM Trans Asian Low-resour Lang Inf Process 15(4):Article 22
4. Joint Entropy (2019) https://en.wikipedia.org/wiki/Joint_entropy. Accessed 10 Mar 2019
5. Machine Translation (2019) https://en.wikipedia.org/wiki/Machine_translation. Accessed 4 Apr 2019
6. Minn AZ (2009) A comparative study of the two grammatical systems of written english and MYANMAR and its significance to learning english as a foreign language. Ph.D Dissertation, Department of English, University of Mandalay, Myanmar
7. Myanmar (2019) https://en.wikipedia.org/wiki/Myanmar. Accessed 3 Apr 2019
8. Myanmar Language Commission (1993) Myanmar—English dictionary. Republic of the Union of Myanmar, Myanmar Language Commission, Naypyitaw
9. Myanmar Language Commission (2003) Myanmar orthography, 2nd edn. Republic of the Union of Myanmar, Myanmar Language Commission, Naypyitaw
10. Myanmar Language Commission (2005) Myanmar Grammar, 1st edn. Republic of the Union of Myanmar, Myanmar Language Commission, Naypyitaw
11. Myanmar Population (2019) http://worldpopulationreview.com/countries/myanmar-population/. Accessed 4 Apr 2019
12. Myanmar (pyidaungsu) Font (2019) https://www.mmunicode.org/wiki/pyidaungsu-font/. Accessed June 2019
13. Myanmar Script Summary (2015) https://r12a.github.io/scripts/myanmar/. Accessed 9 Mar 2019
14. Myanmar Unicode and NLP Research Center (1998) http://mcf.org.mm/myanmar-unicode.html. Accessed 3 Apr 2019
15. Myint C (2011) A hybrid approach for part-of-speech tagging of Burmese texts. In: Proceedings of 2011 international conference on computer and management, Wuhan, China, pp 648–651
16. Myint STY, Khin MM (2013) Lexicon based word segmentation and part of speech tagging for written Myanmar text. Int J Comput Ling Nat Lang Proc 2(6):396–403
17. Nam TV, Hue NT, Khanh PH (2017) "Building a syllable database to solve the problem of Khmer word segmentation. Int J Nat Lang Comput 6(1):2278–2307

18. Nguyen DQ, Thanh V, Nugyen DQ, Dras M, Johnson M (2017) From word segmentation to POS tagging for Vietnamese. In: Proceedings of Australasian language technology association workshop, pp 108–113
19. Phyu ML, Hashimoto K (2017) Burmese word segmentation with Character Clustering and CRFs. In: 14th international joint conference on computer science and software engineering (JCSSE), Nakhon Si Thammarat, Thailand
20. Syntatic ambiguity (2019) https://en.wikipedia.org/wiki/Syntactic_ambiguity. Accessed 3 Apr 2019
21. Tedla YK, Yamamoto K, Marasinghe A (2016) Tigrinya part-of-speech tagging with morphological patterns and the new Nagaoka Tigrinya Corpus. Int J Comput Appl (0975–8887) 146(14):33–41
22. Win MT, Win MM, Than MM, Than M, Aye K (2011) Burmese Phrase Segmentation. In: Proceedings of conference on human language technology for development, Egypt, pp 27–33