Comparison of Six POS Tagging Methods on 10K Sentences Myanmar Language (Burmese) POS Tagged Corpus

Khin War War Htike‡, Ye Kyaw ThuλΛ, Zuping Zhang‡, Win Pa Pa†,
Yoshinori SagisakaΛ, and Naoto Iwahashiλ

‡ School of Information Science and Engineering,
Central South University, Changsha, China
λ Okayama Prefectural University (OpU), Okayama, Japan
† University of Computer Studies Yangon (UCSY), Yangon, Myanmar
Λ Language and Speech Science Research Lab., Waseda University, Tokyo, Japan
Email: khinwarwarhtike@csu.edu.cn, ye@c.oka-pu.ac.jp, zpzhang@csu.edu.cn,
winpapa@ucsy.edu.mm, iwahashi@c.oka-pu.ac.jp, ysagisaka@gmail.com

Abstract. A robust Myanmar Part-of-Speech (POS) tagger is necessary for Myanmar natural language processing (NLP) research and not available publicly yet. For this reason, we developed a manually annotated ten thousand sentence POS tagged corpus for the general domain. We also evaluated six POS tagging approaches: Conditional Random Fields (CRFs), Hidden Markov Model (HMM), Maximum Entropy (MaxEnt), Support Vector Machine (SVM), Ripple Down Rules-based (RDR) and Two Hours of Annotation Approach (i.e. combination of HMM and Maximum Entropy Markov Model) on our developed POS tagged corpus. The POS tagging experimental results were measured with accuracy and also manual checking in terms of confusion pairs. The result shows that RDR approach give the best performance for open test-sets. The HMM and MaxEnt approaches also give strong results. We plan to release our POS tagged corpus and trained models in early 2017.

Key words: POS Tagging, Corpus Building, the Myanmar Language, Under resourced Language

1 Introduction

Part-of-Speech tagging is an important issue in Natural Language Processing applications such as Machine Translation, Information Retrieval and Text to Speech. 10 Part of Speech for the Myanmar language are defined by Myanmar Language Commission [Lwin, 1993], [Myanmar Language Commission, 2005] and based on these tags some papers modified tags to meet the intended applications. The grammatical pattern of the Myanmar language is different from English and some POS tags are different from English POS tags. E.g. "သွား–ခဲ့–ကြ–ပါ–သည်" ('went/Verb' in English), can be tagged as ("သွား"/Verb "ခဲ့ "/Particle-for-past-tense "ကြ"/Particle-for-plural "ပါ"/Particle "သည်"/Post Positional Marker). The suffixes of the verb, "ခဲ့" showed it is a past tense while following suffix "ကြ" showed that the corresponding subject/object is plural, continuously following suffix "ပါ" showed the polite form and the final suffix "သည်" usually used as sentence final. The present tense of this verb is written as "သွား–သည်" ("go" in English). There is no auxiliary verb in Burmese but more than one suffixes can be used in a word, depending on the context and style of writing or speaking. The ambiguity of POS tags and word segmentation are also an important issue for the Myanmar language. Like other Asian languages such as Khmer, Lao, Thai, POS tagging highly depends on word segmentation. The most important issue is the availability of resources since there is no big enough POS tagged corpus in Burmese to train with Machine learning approaches. In this paper, we introduce and describe ongoing work in the creation of a POS tagged corpus for the Myanmar language. We examine six existing POS tagging methods for incremental training with a new POS tagged corpus (10K). Evaluation results in terms of accuracy on experiments show RDR approach is the best among six POS tagging methods and well suited for the Myanmar language.

2 Related Work

Some studies on Myanmar Part of Speech Tagging can be found in the literature. They used Machine learning techniques such as Hidden Markov Models, Neural Networks.

[Khine, 2009] compared HMM with post-editing rules and pure rule-based approach for the POS tagging performance. They defined 36 POS tags and showed the results on increasing the number of training corpus size from 5K to 1M and the best result is 97.67%. The results showed that HMM with rules outperformed pure rules.

[Phyu et al., 2011] used bigram in POS tagging on a very small training corpus that contains 1,000 sentences with 14 basic tags (54 tags in detail). The test corpora were both close and open data with ambiguous words and gained 95.77% F-score on open data. Backpropagation Neural Network based Myanmar POS tagged was applied by [Hnin et al., 2016]. They use manually tagged 5000 sentences for training (82,892 words) defining 5 more tags over traditional POS tag sets for Myanmar language. The training was done with 3-grams, 4-grams and 5-grams BPNNs on 3 different test sets and gained 0.80 F-score on open test data with 4-grams. [Ye et al., 2014] applied POS tags information in Machine translation with an unsupervised novel bilingual infinite HMM approach. They proved this approach gained two points in BLEU for Myanmar to English translation.

3 Word Segmentation

In Myanmar texts, words composed of single or multiple syllables are usually not separated by white spaces. Spaces are used for easier reading and generally put between phrases, but there are no clear rules for using spaces in the Myanmar language. Therefore, word segmentation is a necessary prerequisite for POS tagging. A Burmese word can usually be identified by the combination of root word, prefix and suffix. For example, a Myanmar word စားသည် (eat) can be segmented into two units: one is root verb, "စား" and the other unit is a postpositional marker "သည်", and forms a complete verb. Conditional Random Fields Myanmar word segmentation [Pa et al., 2016] is used for this POS tagging. The segmented example Burmese sentence, (If you do nothing you get nothing.) is shown as follow:

Unsegmented sentence: ဘာမှမလုပ်ရင်ဘာမှမရဘူး။
Word segmented sentece: ဘာ␣မှ␣မ␣လုပ်␣ရင်␣ဘာ␣မှ␣မ␣ရ␣ဘူး␣။

Most of the Myanmar words are formed by one to three syllables and in the above example sentence, all words are formed by one syllable. Myanmar negative statements are constructed with the prefix မ (ma) as in negative imperatives and prohibitions and generally followed by suffix as we have shown in the above sentence မ + လုပ် + ရင် (don't do) and မ + ရ + ဘူး (not get). It is similar to ne + conjugated verb + pas in the French language.

4 POS Tag-set

There are 10 POS tags in Myanmar language that is generally defined by Myanmar Language Commission [Lwin, 1993], [Myanmar Language Commission, 2005]. They are Noun, Pronoun, Adjective, Adverb, Verb, Post-positional-market, Particles, Conjunction, Interjection and Punctuation.

16 Myanmar POS are used in our tag set to meet the necessity of further NLP processing such as information extraction, semantic processing and machine translation. The definitions and descriptions of POS tags are presented in detail in Table 1. Comparing with a POS tag-set of Asian Language Treebanking (ALT) Tool [Thu et al., 2016], the definition of "fw (foreign word) POS tag" is different and we added "tn (text number) POS tag" for Myanmar text numbers such as တစ်ရာ (one hundred), တစ်ထောင် (one thousand) (see Table 1).

5 POS Tagging Methodologies

In this section, we describe the POS tagging methodologies used in the experiments in this paper.

5.1 Conditional Random Fields (CRFs)

Linear-chain conditional random Fields (CRFs) [Lafferty et al., 2001] are models that consider dependencies among the predicted segmentation labels that are

Table 1. Part-of-Speech Tag-set for Myanmar.

POS Tag	Brief Definition	Examples
abb	Abbreviation	အထက(Basic Education High School), လ.ဝ(Confidentiality)
adj	Adjective	ရဲရင့်(brave), လှပ(beautiful), မွန်မြတ်(noble)
adv	Adverb	ဖြေးဖြေး(slow), နည်းနည်း(less)
conj	Conjunction	နှင့် (and), ထို့ကြောင့်(therefore) သို့မဟုတ်(or)
fw	Foreign word	1, 2, 3, Myanmar, ミャンマー (Myanmar in Japanese), BBC, Google.
int	Interjection	အမလေး(Oh My God!)
n	Noun	ကျောင်း(school), စာအုပ်(book), ဒေါ်အောင်ဆန်းစုကြည်(Daw Aung San Suu Kyi), လွတ်လပ်ရေး (freedom)
num	Number	၁ (1), ၂ (2), ၃ (3), ၁၀ (10), ၁၀၀(100), ၁၀၀၀ (1000)
part	Particle	များ (used to form the plural nouns as "-s , -es"), ခဲ့ (the past tense "-ed"), သင့် (modal verb "shall"), လိမ့် (modal verb "will"), နိုင်(modal verb "can")
part_neg	Negative Particle: Particle that is used to form negative meaning of adjective and verb	မဆိုးပါဘူး (not bad), မသွားနိုင်ဘူး(can not go)
ppm	Post-positional Marker	သည်, က, ကို, အား, သို့, မှာ, တွင် (at, on ,in, to)
pron	Pronoun	ကျွန်တော် (I), ကျွန်မ (I), သင် (you) , သူ (he), သူမ (she)
punc	Punctuation	။, ၊, (,), \, _ ,“ ”
sb	Symbol	?, #, &, %, $, €, £, π, λ, ÷, +, ×,
tn	Text Number	တစ် (one), နှစ် (two), သုံး (three), တစ်ရာ (one hundred), တစ်ထောင် (one thousand)
v	Verb	ကူညီ (help), လိုက်နာ (observe), အားပေး (encourage)

inherent in the state transitions of finite state sequence models and can incorporate domain knowledge effectively into segmentation. Unlike heuristic methods, they are principled probabilistic finite state models on which exact inference over sequences can be efficiently performed. The model computes the following probability of a label sequence $\mathbf{Y} = \{y_1, ..., y_T\}$ of a particular character string $\mathbf{W} = \{w_1, ..., w_T\}$.

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T} \sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t)) \tag{1}$$

where $Z(\mathbf{W})$ is a normalization term, f_k is a feature function, and $\boldsymbol{\lambda}$ is a feature weight vector.

5.2 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a probabilistic sequence model: given a sequence of units (words, letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the best label sequence [Rabiner, 1989], [Jurafsky & Martin, 2000]. In an HMM for POS tagging, the observation is a sequence of words $\mathbf{o} = x_1, \ldots, x_n$ and is associated with a state sequence of POS tags that we cannot observe $\mathbf{s} = y_1, \ldots, y_n$. The model describes the joint state and observation sequence:

$$p(y_1, \ldots, y_n, x_1, \ldots, x_n) = p(y_1)p(x_1|y_1) \prod_{i=2}^{n} p(y_i|y_{i-1})p(x_i|y_i) \tag{2}$$

and the probability of the observation sequence can be obtained by marginalizing:

$$p(x_1, \ldots, x_n) = \sum_{\mathbf{y}} p(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{3}$$

Here, the key assumptions of HMM are:

- The state sequence $p(y_i|y_1, \ldots, y_{i-1}) = p(y_i|y_{i-1})$ is Markovian
- The observations are conditionally independent of next and previous states and observations given the current state:

$$p(x_i|x_1, \ldots, x_n, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = p(x_i|y_i) \tag{4}$$

A graphical representation of a HMM can be seen in Fig 1.

5.3 Maximum Entropy (MaxEnt)

The original concept of Maximum Entropy (MaxEnt) comes from physics [Jaynes, 1957]. MaxEnt models have been used in various tasks of natural language

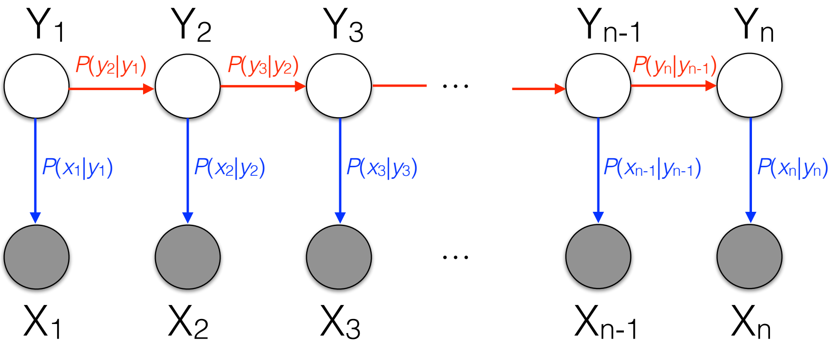


Fig. 1. Graphical representation of a Hidden Markov Model. Here, state variables $Y_1, Y_2, \ldots, Y_n$ form a Markov chain, but this sequence of variables is not observed (i.e. hidden). The $X_1, X_2, \ldots, X_n$ are observable variables (i.e. output) of the Markov chain. Horizontal and vertical arrows indicate conditional dependence relations of variables.

processing including POS tagging [Berger et al., 1996a], [Ratnaparkhi, 1996], [Toutanova & Manning, 2000], [Denis & Sagot, 2012]. The principle of ME is to estimate the probability distribution based on the minimal bias (i.e. maximal entropy) while verifying the statistical properties measured on the observation set. Those properties are referred to as the constraints that derived from the training data. The exponential form of MaxEnt model for POS tagging can be stated as Equation (5):

$$P(t|h) = \frac{1}{Z(h)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(h,t)\right) \tag{5}$$

where, t is the POS tag, h is the history/context, $f_i(h,t)$ is a feature/class with associated feature-weight parameter λ_i and normalization function $Z(h)$. The POS tagging problem can be formally stated as Equation (6):

$$P(t_1, \ldots, t_n, |w_1, \ldots, w_n) = \prod_{i=1}^{n} P(t_i|h_i) \tag{6}$$

where, given a sequence of words $w_1, \ldots, w_n$ and finding the conditional probability of a tag sequence $t_1, \ldots, t_n$. In MaxEnt modeling, the features are binary or multiple valued functions, which associate a POS tag with various elements of the context. For example:

$$f_i(h,t) = \begin{cases} 0, & \text{if } word(h) = \text{Yangon } \& \ t = \text{N} \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

5.4 Ripple Down Rules-based (RDR)

Ripple-Down Rules (RDR) is an approach to building knowledge-based systems (KBS) incrementally, while the KBS is in routine use [Compton & Jansen, 1990]. [Nguyen et al., 2016], [Nguyen et al., 2014] present a new error-driven approach to automatically restructure transformation rules in the form of a Single Classification Ripple Down Rules (SCRDR) tree [Compton & Jansen, 1990] [Richards, 2009]. A SCRDR can be notated as a triple $< rule, X, N >$, where X and N are the exception RDR and the succeeding RDR (i.e. if-not rules) respectively (see Figure 2) [Scheffer, 1996]. Cases in SCRDR are evaluated by passing a case to the root (Rule 0 in Figure 2). At any node in SCRDR tree (i.e. Rule 1 to Rule 6), if the condition of a node n met, the case is passed on to the exception child of n using the except link if it exists. Otherwise, the case is passed on to the if-not child of n. In the SCRDR approach, a conclusion is always given by the last node in the process. To ensure that a conclusion is always given, the root node (also known as default node) is usually set up with the condition which is always satisfied.

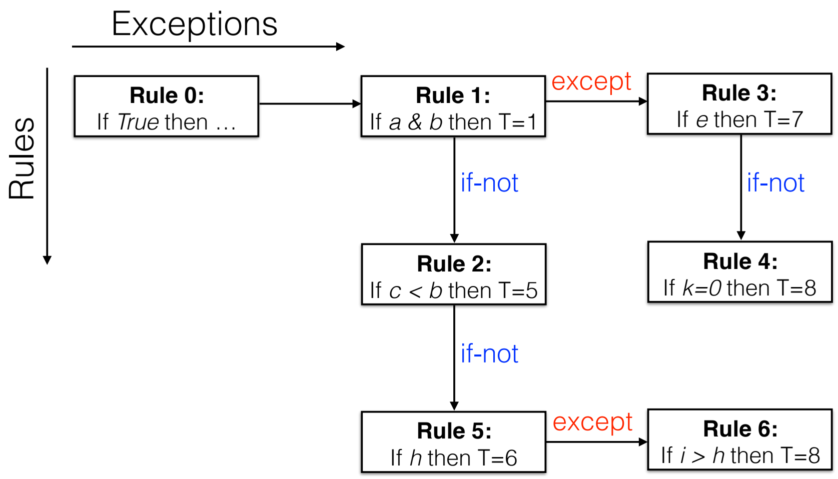


Fig. 2. A binary tree of Single Classification Ripple Down Rules.

5.5 Support Vector Machine (SVM) based Point-wise Classification

Generally, sequence-based pronunciation prediction methods such as [Nagano et al., 2005] require a fully annotated training corpus. To reduce the cost of preparing a fully annotated corpus and also considering possible future work on domain adaptation from the general to the target domain, the techniques involving only partial annotation have been developed [Ringger et al., 2007], [Tsuboi et al., 2008]. [Neubig & Mori, 2010] proposed the combination of two separate techniques to achieve more efficient corpus annotation: point-wise estimation and word-based annotation. Point-wise estimation assumes that every decision about a segmentation point or word pronunciation is independent of the other decisions [Neubig & Mori, 2010]. From this concept, a single annotation model can be trained on single annotated words, even if the surrounding words are not annotated such as ငါ/{ngar} ကျေးဇူး/{kyei: zu:} တင်ပါတယ်/{tin ba de} (Thank you in English). In this paper, we applied this approach to phonemes of syllables within a word and thus the previous example will change to ငါ/{ngar} ကျေး/{kyei:} ဇူး/{zu:} တင်/{tin} ပါ/{ba} တယ်/{de}.

5.6 Two Hours of Annotation Approach

Semi-supervised learning of POS tagging from two hours manually annotated data was proposed for low-resource languages [Garrette & Baldridge, 2013]. Proposed system has four main steps, (1) Tag dictionary expansion, (2) Weighted model minimization, (3) Expectation maximization (EM) HMM training and (4) Maximum Entropy Markov Model (MEMM) training. The label propagation, the Modified Adsorption (MAD) algorithm [Talukdar & Crammer, 2009] is used for tag dictionary expansion, step 1. It is a graph-based technique for spreading labels between related items and their graphs are seeded with POS-tag labels from the two hours human-annotated data. After step 1, all words have information, but still noisy and thus induce a cleaner hard tagging [Ravi et al., 2010], [Garrette & Baldridge, 2012] from a noisy soft tagging in step 2, model minimization. Step 3 uses the EM algorithm initialized with the noisy labels and constrained with the expanded tag dictionary to produce an HMM. However, the HMM produced by stage 3 will contain zero probabilities for out-of-vocabulary (OOV) words of the test-corpus and thus it cannot be used directly for POS tagging. It is used to provide a Viterbi labeling of the raw corpus, following the auto-supervision [Garrette & Baldridge, 2012]. The output of step 3 can be concatenated with the token-supervised corpus if it is available, and used to train a MEMM POS tagger.

6 Experimental Setup

6.1 Corpus Developing

We collected 11,000 sentences (234,802 words) from Wikipedia (including various area such as economics, history, news, politics, philosophy) [Wikipedia, 2014]. Word segmentation, tagging with defined POS tags for each word and error checking were done manually by three Myanmar natives who already have experience in NLP areas such as POS tagging, parallel corpus building. We used the POS tag-sets that we mentioned in Section 4. The average number of words per sentence in the whole corpus is 21.35. Here, two Myanmar punctuation symbols little section "၊" (significance close to comma) and section "။" (significance close to full stop) also counted as words. The shortest sentence in the corpus contained 2 words (for example: တဆိတ်လောက် ။, “please” in English) and there are 17 shortest sentences in total. The longest sentence of the current corpus contained 419 words and we found that is a Myanmar language translation of the preamble of the Universal Declaration of Human Rights [The United Nations, 1948]. As we mentioned about “fw” (foreign word) POS tag in Table 1, we didn’t remove non Myanmar words from the original data. And thus, some sentences of our corpus contained non Myanmar words such as “ထို ကဲ့သို့ site များ ကို ဝင်ကြည့် တာ ပဲ ဖြစ် ဖြစ် download ဆွဲ တာ ဖြစ် ဖြစ် spyware ပါ လာ တတ် သည် ။” (here, three English words “site”, “download”, “spyware” contained), “ဂျပန် နိုင်ငံ သည် မြန်မာ နိုင်ငံ အား မြန်မာ ミャンマー ဟု ခေါ် ပေ မဲ့ မြန်မာ လူ များ အား ဘားမိစ် Burmese ビルマ人 ဟု ဆက်လက် သုံးစွဲ လျက် ရှိ သည် ။” (here, two Japanese words, “ミャンマー” and “ビルマ人” contained). This characteristic is a significant difference from existing proposed POS tag-sets [Khine, 2009], [Phyu et al., 2011], [Hnin et al., 2016], [Thu et al., 2016], [Thant et al., 2012]. Statistic of POS tag-set in 10K corpus is as shown in Table 2.

Table 2. Statistic of Part-of-Speech Tag-set in 10K corpus.

No.	POS-tag	Frequency	Proportion
1	n	59957	28.04%
2	part	44074	20.61%
3	ppm	34958	16.35%
4	v	28702	13.42%
5	punc	14374	6.72%
6	conj	10578	4.95%
7	adj	6302	2.95%
8	num	3527	1.65%
9	adv	2671	1.25%
10	pron	2579	1.21%
11	tn	2121	0.99%
12	fw	2080	0.97%
13	part_neg	1409	0.66%
14	abb	264	0.12%
15	sb	159	0.07%
16	int	95	0.04%

6.2 Closed and Open Test Set

There are two types of test data: one closed data set (10% of the training data) and one open data set. The open test set contains 1,000 sentences (20,952 words) and it is also taken from the Myanmar language Wikipedia [Wikipedia, 2014].

6.3 Software

We used following open source POS Taggers for the experiments:

- CRFSuite: We used the CRFsuite tool (version 0.12) [Okazaki, 2007], (`https://github.com/chokkan/crfsuite`) for training and testing CRFs models. The main reason was its speed relative to other CRFs toolkits.
- Jitar (version 0.3.3): is a simple part-of-speech tagger, based on a trigram Hidden Markov Model (HMM). It (partly) implements the ideas set forth in [Brants, 2000]. Jitar is written in Java [de Kok, 2014] and thus easy to use in other Java programs, or languages that run on the JVM.
- Maximum Entropy Modeling Toolkit for Python and C++: provides a (Conditional) Maximum Entropy Modeling. We used a python extension module (maxent module) for building a Maximum Entropy POS tagger [Zhang, 2003], [Berger et al., 1996b], [Pietra et al., 1997].
- RDRPOSTagger (Version 1.2.3): is a rule-based Part-of-Speech and morphological tagging toolkit [Nguyen et al., 2014], [Nguyen et al., 2016]. It is a robust, easy-to-use and language-independent toolkit. It employs an error-driven approach to automatically construct tagging rules in the form of a binary tree. The main properties of RDRPOSTagger are it obtains fast performance in both learning and tagging process and achieves a very competitive accuracy compared to the state-of-the-art results.
- KyTea: is a general toolkit (version 0.47) [Neubig & Mori, 2010], (`https://github.com/neubig/kytea`) and it is able to handle word segmentation and tagging. It uses a point-wise classifier-based (SVM or logistic regression) approach and the classifiers are trained with LIBLINEAR (`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`). We used the KyTea toolkit for studying G2P bootstrapping with SVM based point-wise classification for Myanmar language.
- Low-Resource POS-Tagging toolkit (2014): contains Scala code for training and tagging using the approach described in the papers [Garrette & Baldridge, 2013], [Garrette et al., 2013]. We used it for experiments on the two hour annotation approach that we mentioned in Section 5.6.

We ran all above software with default parameters for building the POS tagging models. Although feature engineering is usually an important component of machine-learning approaches, the POS tagging models were built with features coming only from the corpus, to allow for a fair comparison between the six approaches.

7 Evaluation Criteria

The POS tagging performance was measured using the accuracy defined as follows:

$$Accuracy = \frac{\#of\ correct\ POS - tags}{\#of\ tokens\ in\ test\ corpus} \tag{8}$$

We also used the SCLITE (score speech recognition system output) program from the NIST scoring toolkit SCTK version 2.4.10 [The National Institute of Standards and Technology, 2015] for making dynamic programming based alignments between reference and hypothesis POS-tag strings and calculation of Word Error Rate (WER). In our case, WER will be equal to POS tagging error rate (POS-ERR). The SCLITE scoring method for calculating the erroneous words in WER: first make an alignment of the hypothesis (the output from the trained model) and the reference POS strings (POS-tagged manually) and then perform a global minimization of the Levenshtein distance function which weights the cost of correct words, insertions (I), selections (D), substitutions (S) and the number of POS (N). The formula for WER can be stated as Equation (9):

$$WER = \frac{(I + D + S) \times 100}{N} \tag{9}$$

For example, scoring I, D and S for the POS-tagged Buremese sentence ဒီ/adj အထုတ်/n သေး/adj လေး/part ကို/ppm ချိန်/v ပေး/part ပါ/part လား/part /punc ခင်ဗျာ/part /punc (”Can you measure this small package?” in English) is as follow:

```
Scores: (#C #S #D #I) 10 1 1 1
REF: adj n  ADJ part ppm * v PART part part punc part punc
HYP: adj n PART part ppm N v **** part part punc part punc
Eval:        S             I   D
```

In this case, one substitution (ADJ => PART), one insertion (* => N) and one delection (PART => ****) happened, that is $S = 1$, $D = 1$, $I = 1$, $C = 10$, $N = 12$ and thus WER or POS-ERR is equal to 25%.

8 Results

The accuracies of six POS-tagging methods that we trained with 10K POS-tagged sentences are shown in Table 3. Test data sizes are 1K sentences for both closed (10% of the training data) and open test data. Underlined numbers indicate the highest scores of the six different approaches. The experimental results show that RDR achieved the highest acccuracy 97.05% with open-test data. On the other hand, SVM gives highest accuracy 99.83% with closed-test data.

In order to study how six models behave with varying amounts of training data, we run a sequence of experiments that trained CRFs, HMM, MaxEnt, RDR, SVM and Two-Hours models from 2K, 4K, 6K, 8K to 10K sentences respectively. From the results with 1K open test data in Figure 3, it is clear that the RDR and HMM models are more able to learn and performs identically from 2K to 10K training data. The MaxEnt, CRFs and SVM learning curve are

Table 3. Accuracies of six POS tagging approaches with 10K model on 1K open test data set.

Methods	**Closed Test-set**	**Open Test-set**
CRFs	97.77%	95.05%
HMM	97.31%	96.43%
MaxEnt	96.55%	96.31%
RDR	98.42%	97.05%
SVM	99.83%	93.55%
Two-Hours	95.83%	92.87%

also similar and accuracies increase gradually as the training data size grow. Although the learning curve of the Two-Hours approach drops from 4K, final results obtained after trained with 10K are comparable with SVM.

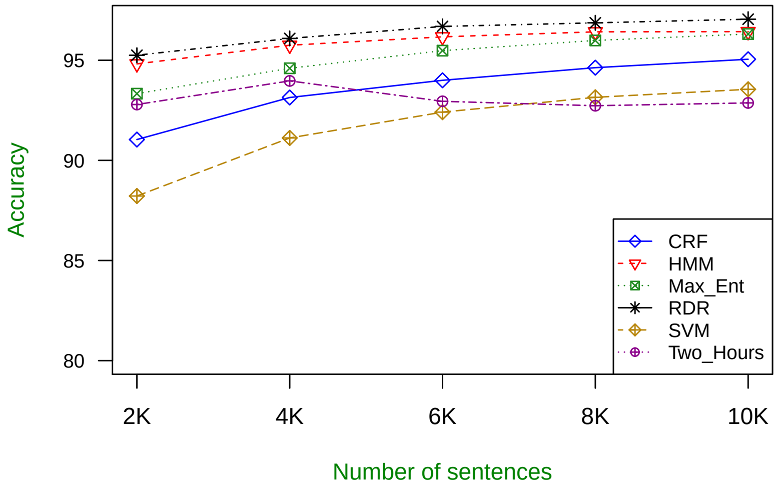
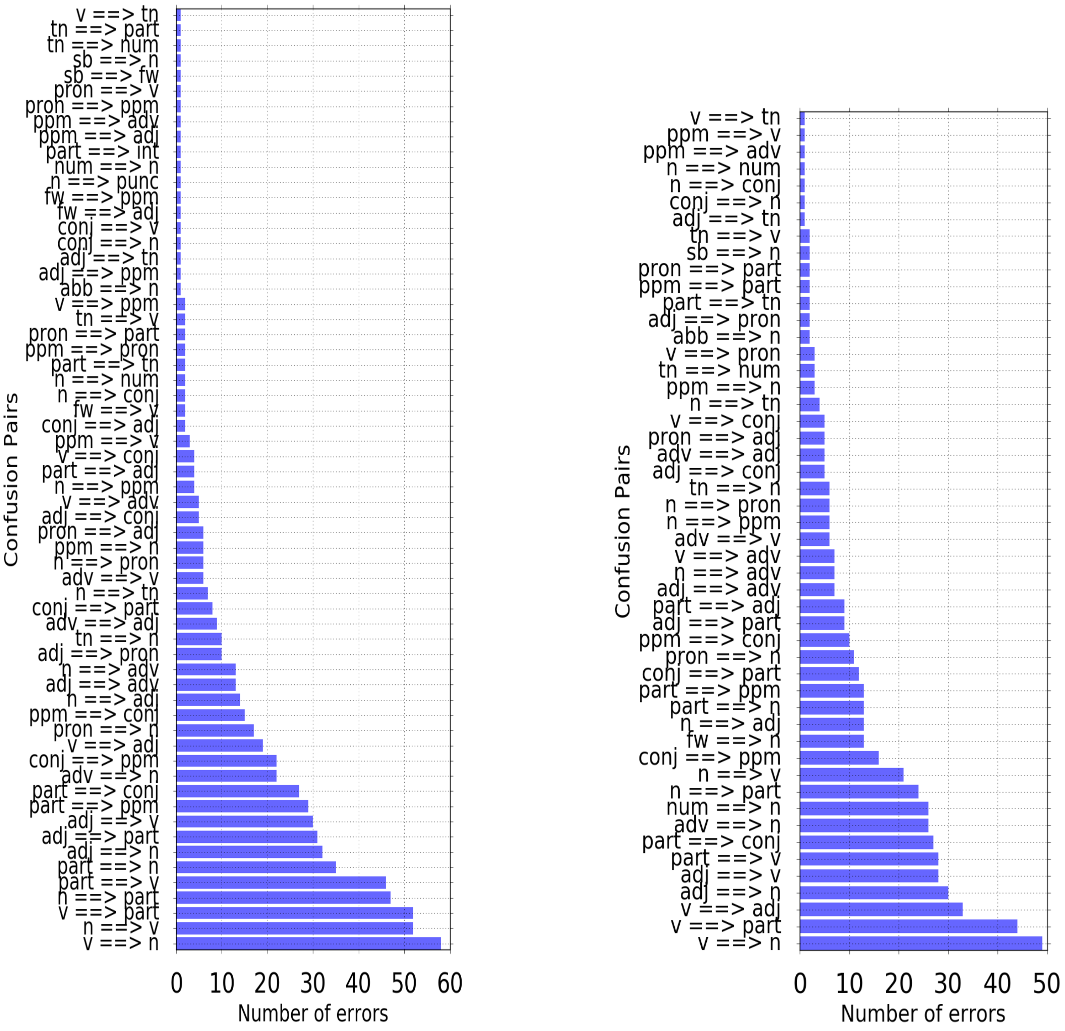


Fig. 3. Accuracy of six POS tagging methods on varying training data sizes with 1K open test data set.



(a) 3gHMM

(b) RDR

Fig. 4. Confusion pairs with 10K training model on 1K open test data set.

9 Discussion

We calculated OOV (Out of Vocabulary) for all incremental training data with open test sets and the results are 1,900 OOVs for 2K training data, 1,334 OOVs for 4K training data, 1,083 OOVs for 6K training data, 960 OOVs for 8K training data and 878 OOVs for 10K training data. From the experimental results we confirmed that RDR model is able to achieve 97.05% accuracy even with 10K training data and 878 OOVs on current open test-set.

Error analysis of six POS tagging approaches has been done with the help of confusion matrices produced by the SCLITE program (see Section 7). Confusion pairs of 3gHMM (in total: 62 pairs) and RDR (in total: 50 pairs) 10K models on 1K open test data are shown in Figure 4. Since the POS tag "n" (noun) has the most highest frequency and proportion in the corpus (see Table 2), unknown words have a tendency of being assigned "n" tags. The top confusion pairs for both 3gHMM and RDR are "v ==> n". We studied all confusion pairs for six models (including CRFs, MaxEnt, SVM and Two-Hours) and found that confusion pairs are different based on the POS tagging methodologies. For example, the top confusion pairs of Two-Hours approach for both closed and open test data are "punc ==> conj". One more example is unk (Unknown) tags are containing in running with KyTea toolkit (see Section 6.3) for SVM approach as shown in Table 4.

Table 4. Confusion pairs of SVM POS tagging approaches with 10K model on 1K open test data.

Closed Test-set		**Open Test-set**	
Frequency	REF ==> HYP	Frequency	REF ==> HYP
7	v ==> part	506	n ==> unk
6	part ==> v	132	fw ==> unk
5	ppm ==> conj	110	v ==> unk
4	conj ==> ppm	57	num ==> unk
4	part ==> adj	50	v ==> part
3	adj ==> part	36	part ==> v
3	adj ==> v	30	adv ==> unk
3	part ==> conj	29	adj ==> unk
2	part ==> ppm	29	conj ==> ppm
1	conj ==> part	28	v ==> adj

RDRPOSTagger employs an error-driven approach to automatically construct tagging rules in the form of a binary tree and it obtains fast performance in both learning and tagging process. One of the merit points of the RDR approach is producing human readable SCRDR rules as a trained model and that will be useful for analysis on word-category disambiguation of the Myanmar language. The following is a part of SCRDR tree that we trained with 10K:

```
True : object.conclusion = ”NN”
  object.tag == ”part_neg” : object.conclusion = ”part_neg”
  object.tag == ”adv” : object.conclusion = ”adv”
    object.word == “တရားဝင်” and object.nextTag1 == ”n” : object.conclusion = ”adj”
    object.word == “သီးခြား” and object.nextTag1 == ”n” : object.conclusion = ”adj”
    object.word == “အများဆုံး” and object.nextTag1 == ”n” : object.conclusion = ”n”
    object.word == “အထူး” and object.nextTag1 == ”n” : object.conclusion = ”adj”
```

10 Conclusion and Future Work

In this paper, we conducted six POS tagging experiments on Myanmar language with our developing POS-tagged corpus. We found that RDR approach can consistently achieved accuracy 97.05% on open data set and best among six POS tagging methods. In further work, we plan to check errors of manual segmentation and POS tagging of the whole corpus and re-evaluate with cross-validation. We plan to release our POS-tagged corpus including trained models for Myanmar language NLP research in early 2017.

Acknowledgments. We would like to express our gratitude to Dr. Chenchen Ding and Dr. Andrew Finch for their valuable comments.

References

[Berger et al., 1996a] Berger, Adam L., Stephen A. Della Pietra, & Vincent J. Della Pietra 1996a. A Maximum Entropy approach to Natural Language Processing. COMPUTATIONAL LINGUISTICS, 22:39–71.

[Berger et al., 1996b] Berger, Adam L., Vincent J. Della Pietra, & Stephen A. Della Pietra 1996b. A Maximum Entropy Approach to Natural Language Processing. Comput. Linguist., 22(1):39–71.

[Brants, 2000] Brants, Thorsten 2000. TnT: A Statistical Part-of-speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing, ANLC ’00, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Compton & Jansen, 1990] Compton, P., & R. Jansen 1990. A philosophical basis for knowledge acquisition. Knowledge Acquisition, 2(3):241 – 258.

[de Kok, 2014] de Kok, Daniël 2014. Jitar: A simple Trigram HMM part-of-speech tagger. [accessed 2016].

[Denis & Sagot, 2012] Denis, Pascal, & Benoît Sagot 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. Language Resources and Evaluation, 46(4):721–736.

[Garrette & Baldridge, 2012] Garrette, Dan, & Jason Baldridge 2012. Type-supervised Hidden Markov Models for Part-of-speech Tagging with Incomplete Tag Dictionaries. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 821–831, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Garrette & Baldridge, 2013] Garrette, Dan, & Jason Baldridge 2013. Learning a Part-of-Speech Tagger from Two Hours of Annotation. pages 138–147.

[Garrette et al., 2013] Garrette, Dan, Jason Mielens, & Jason Baldridge 2013. Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. pages 583–592.

[Hnin et al., 2016] Hnin, Hay Mar, Win Pa Pa, & Ye Kyaw Thu 2016. Back-Propagation Neural Network Approach to Myanmar Part-of-Speech Tagging. In Genetic and Evolutionary Computing - Proceedings of the Tenth International Conference on Genetic and Evolutionary Computing, ICGEC 2016, November 7-9, 2016, Fuzhou City, Fujian Province, China, pages 212–220.

[Jaynes, 1957] Jaynes, E. T. 1957. Information Theory and Statistical Mechanics. Phys. Rev., 106:620–630.

[Jurafsky & Martin, 2000] Jurafsky, Daniel, & James H. Martin 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

[Khine, 2009] Khine, Khine Zin 2009. Hidden Markov Model with Rule Based Approach for Part-of-Speech Tagging of Myanmar Language. In Proceedings of the 3rd International Conference on Communications and Information Technology, pages 123–128.

[Lafferty et al., 2001] Lafferty, John D., Andrew McCallum, & Fernando C. N. Pereira 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Lwin, 1993] Lwin, San 1993. Myanmar - English Dictionary. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[Myanmar Language Commission, 2005] Myanmar Language Commission, Department 2005. Myanmar Thdda (2005). Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

[Nagano et al., 2005] Nagano, Tohru, Shinsuke Mori, & Masafumi Nishimura 2005. A stochastic approach to phoneme and accent estimation. In INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005, pages 3293–3296. ISCA.

[Neubig & Mori, 2010] Neubig, Graham, & Shinsuke Mori 2010. Word-based Partial Annotation for Efficient Corpus Construction. In The seventh international conference on Language Resources and Evaluation (LREC 2010), pages 2723–2727, Malta.

[Nguyen et al., 2014] Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham, & Son Bao Pham 2014. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 17–20, Gothenburg, Sweden. Association for Computational Linguistics.

[Nguyen et al., 2016] Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham, & Son Bao Pham 2016. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. AI Communications, 29(3):409–422.

[Okazaki, 2007] Okazaki, Naoaki 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

[Pa et al., 2016] Pa, Win Pa, Ye Kyaw Thu, Andrew Finch, & Eiichiro Sumita 2016. Word Boundary Identification for Myanmar Text Using Conditional Random Fields, pages 447–456. Springer International Publishing, Cham.

[Phyu et al., 2011] Phyu, Hninn Myint, Myat Htwe Tin, & Thein Ni Lar 2011. Bigram Part-of-Speech Tagger for Myanmar Language. In Proceedings of 2011 International Conference on Information Communication and Management, pages 147–152.

[Pietra et al., 1997] Pietra, Stephen Della, Vincent J. Della Pietra, & John D. Lafferty 1997. Inducing Features of Random Fields. IEEE Trans. Pattern Anal. Mach. Intell., 19(4):380–393.

[Rabiner, 1989] Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286.

[Ratnaparkhi, 1996] Ratnaparkhi, Adwait 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In Conference on Empirical Methods in Natural Language Processing, pages 133–142.

[Ravi et al., 2010] Ravi, Sujith, Ashish Vaswani, Kevin Knight, & David Chiang 2010. Fast, Greedy Model Minimization for Unsupervised Tagging. In Huang, Chu-Ren, & Dan Jurafsky (eds), COLING, pages 940–948. Tsinghua University Press.

[Richards, 2009] Richards, Debbie 2009. Two decades of Ripple Down Rules research. Knowledge Eng. Review, 24(2):159–184.

[Ringger et al., 2007] Ringger, Eric, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, & Deryle Lonsdale 2007. Active Learning for Part-of-speech Tagging: Accelerating Corpus Annotation. In Proceedings of the Linguistic Annotation Workshop, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Scheffer, 1996] Scheffer, Tobias 1996. Algebraic Foundation and Improved Methods of Induction of Ripple Down Rules. In In, pages 23–25.

[Talukdar & Crammer, 2009] Talukdar, Partha Pratim, & Koby Crammer 2009. New Regularized Algorithms for Transductive Learning. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, ECML PKDD '09, pages 442–457, Berlin, Heidelberg. Springer-Verlag.

[Thant et al., 2012] Thant, Win Win, Tin Myat Htwe, & Ni Lar Thein 2012. Parsing of Myanmar sentences with function tagging. CoRR, abs/1205.1603.

[The National Institute of Standards and Technology, 2015] The National Institute of Standards and Technology, (NIST) 2015. Speech Recognition Scoring Toolkit (SCTK), Version: 2.4.10.

[The United Nations, 1948] The United Nations 1948. Universal Declaration of Human Rights.

[Thu et al., 2016] Thu, Ye Kyaw, Win Pa Pa, Masao Utiyama, Andrew Finch, & Eiichiro Sumita 2016. Introducing the Asian Language Treebank (ALT). In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, & Stelios Piperidis (eds), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA).

[Toutanova & Manning, 2000] Toutanova, Kristina, & Christopher D. Manning 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-speech Tagger. In Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the

38th Annual Meeting of the Association for Computational Linguistics - Volume 13, EMNLP '00, pages 63–70.

[Tsuboi et al., 2008] Tsuboi, Yuta, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, & Yuji Matsumoto 2008. Training Conditional Random Fields Using Incomplete Annotations. In Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1, COLING '08, pages 897–904, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wikipedia, 2014] Wikipedia 2014. The Wikipedia Myanmar.

[Ye et al., 2014] Ye, Kyaw Thu, Tamura Akihiro, Finch Andrew, Sumita Eiichiro, & Sagisaka Yoshinori 2014. Unsupervised POS Tagging of Low Resource Language for Machine Translation. In Proceedings NLP2014, pages 590–593.

[Zhang, 2003] Zhang, Le 2003. Maximum Entropy Modeling Toolkit for Python and C++. [accessed 2016].