Myanmar POS Resource Extension Effects on Automatic Tagging Methods

Zar Zar Hlaing, Ye Kyaw Thu, Myat Myo Nwe Wai, Thepchai Supnithi, Ponrudee Netisopakul

iSAI-NLP 2020 **November 18 - 20**

NECTEC
a member of NSTDA

Outlines

1. Introduction
2. Word Segmentation
3. POS Tag Sets
4. POS Tagging Methdologies
5. Experimental Setup
6. Results and Conclusion

1. Introduction

- Essential basis of Natural Language Processing (NLP)
- Process of tagging each word in a sentence with a corresponding POS tag
- POS information is very useful in many areas of NLP applications
- POS tagging mainly depends on word segmentation especially in ASEAN languages because these languages have no explicit word boundary
- And thus, the availability of POS-tagged corpus is an important issue

1. Introduction (Cont.)

- There is only one available small POS-tagged corpus for Myanmar language, namely **myPOS corpus version 1.0**

- Therefore, advanced machine learning models cannot be applied to NLP tasks in Myanmar language

Objectives of this study:

- ❖ To manually extend the original **myPOS corpus version 1.0** to **myPOS corpus version 2.0**
- ❖ To study the effects of this larger size POS-tagged corpus

- Compared the tagging accuracies of four methods, such as, CRFs, HMM, RDR and NCRF++

2. Word Segmentation

- Myanmar words are composed of single or multiple syllables that are not separated by white spaces
- Spaces are used for easy reading and generally put between phrases
- There are no clear rules for using spaces in Myanmar language
- And thus, word segmentation is needed before doing the POS tagging process
- A Burmese word can be identified by the combination of a root word, prefix and suffix

2. Word Segmentation (Cont.)

- The example of segmented Burmese sentence for “မင်းစွန့်စားမှုမရှိပဲဘာမှမရဘူး။” (“You can’t get nothing without risk.”) is described as follow:

  - ❖ Unsegmented sentence: မင်းစွန့်စားမှုမရှိပဲဘာမှမရဘူး။
  - ❖ Segmented sentence: မင်း_စွန့်စား_မှု_မ_ရှိ_ပဲ_ဘာ_မှ_မ_ရ_ဘူး_။

- Most of the Myanmar words are formed by one to three syllables

3. POS Tag Sets

- In Myanmar language, there are 10 POS tags defined by Myanmar Language Commission

- These are Noun, Pronoun, Adjective, Adverb, Verb, Post-positional-marker, Particles, Conjunction, Interjection and Punctuation.

- We used 15 Myanmar POS tags in our tag sets

3. POS Tag Sets (Cont.)

Table : Part-of-Speech Tag Sets for Myanmar Language

POS Tag	Brief Definition	Examples
abb	Abbreviation	အိုင်တီ (Information Technology), အ.လ.က (Basic Education Middle School)
adj	Adjectives	လိမ္မာ(clever), ကြင်နာ(kind), ကျော်ကြား(famous)
adv	Adverb	လျင်မြန် (quick), တည်တည်ငြိမ်ငြိမ်(quietly)
conj	Conjuction	နှင့်(and), ထို့ကြောင့်(therefore), သို့မဟုတ်(or)
fw	Foreign word	Facebook, VOA, 1, 2, 3, Myanmar, ミャンマ (Myanmar in Japan)
int	Interjection	အမလေး (Oh my God!)
n	Noun	သစ်ပင်(tree), ခဲတံ(pencil), ခဲဖျက်(eraser), မိန်းကလေး(girl)
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 tenses “-ed”), သင့်(modal verb “shall”), လိမ့်(modal verb “will”), နိုင်(odal verb “can”)
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),
v	Verb	စား(eat), လေ့လာ(learn), နားထောင်(listen)

4. POS Tagging Methdologies

4.1. Conditional Random Fields (CRFs)

- Models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite state sequence models
- Can incorporate domain knowledge into segmentation
- Unlike heuristic methods
- Computes the following probability of a label sequence $Y = \{y1, \ldots, yT\}$ of a particular character string $W = \{w1, \ldots, wT\}$

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

4. POS Tagging Methdologies (Cont.)

4.2. Hidden Markov Model (HMM)

- Probabilistic sequence model of random variables and state variables
- Computes a joint probability distribution over possible sequences of labels and chooses the best label sequence
- Describes the joint state and observation sequence by the following equation:

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

Transition probability **Emission probability**

4. POS Tagging Methdologies (Cont.)

4.2. Hidden Markov Model (HMM)

- Graphical representation of a HMM can be seen in the following figure:

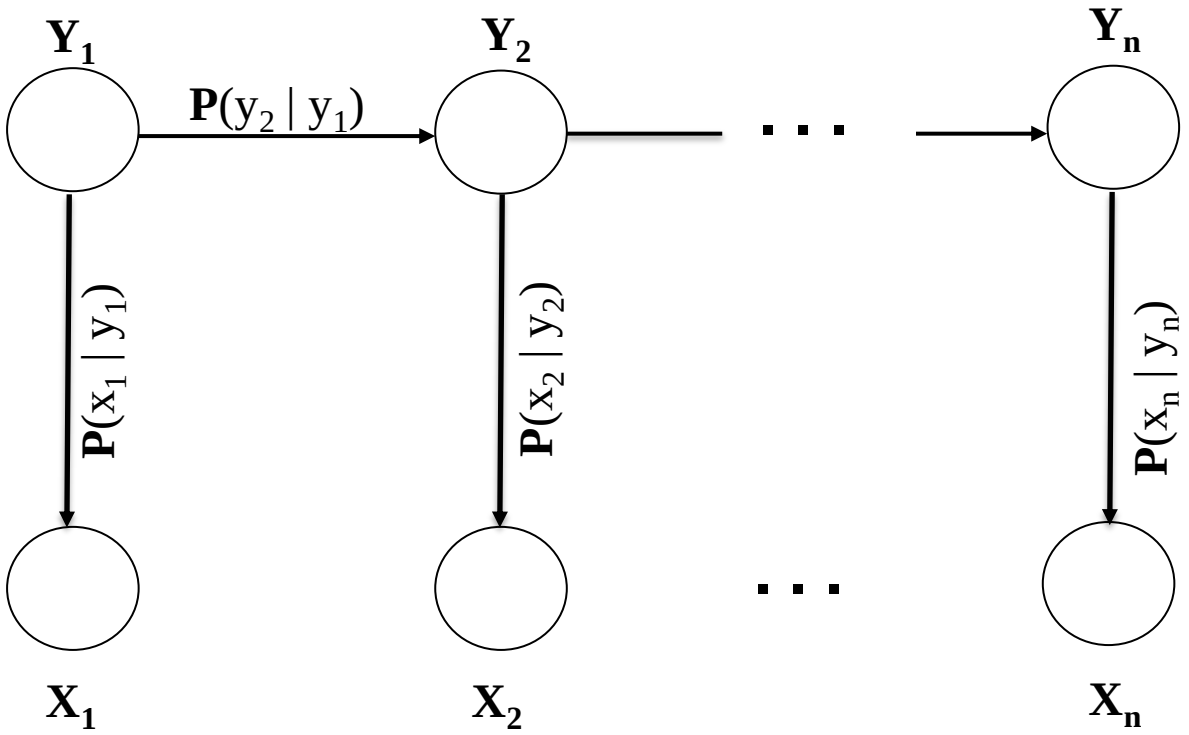

Figure: A graphical representation of a hidden markov model

4. POS Tagging Methdologies (Cont.)

4.3. Ripple Down Rules-based (RDR)

- Strategy of building knowledge-based systems
- RDRPOSTagger provides a failure-driven approach to automatically reconstruct transformation rules in the form of a Single Classification Ripple Down Rules (SCRDR) tree
- A SCRDR can be written as *if X then Y* where *X* is **condition** and *Y* is **conclusion**

4. POS Tagging Methdologies (Cont.)

4.3. Ripple Down Rules-based (RDR)

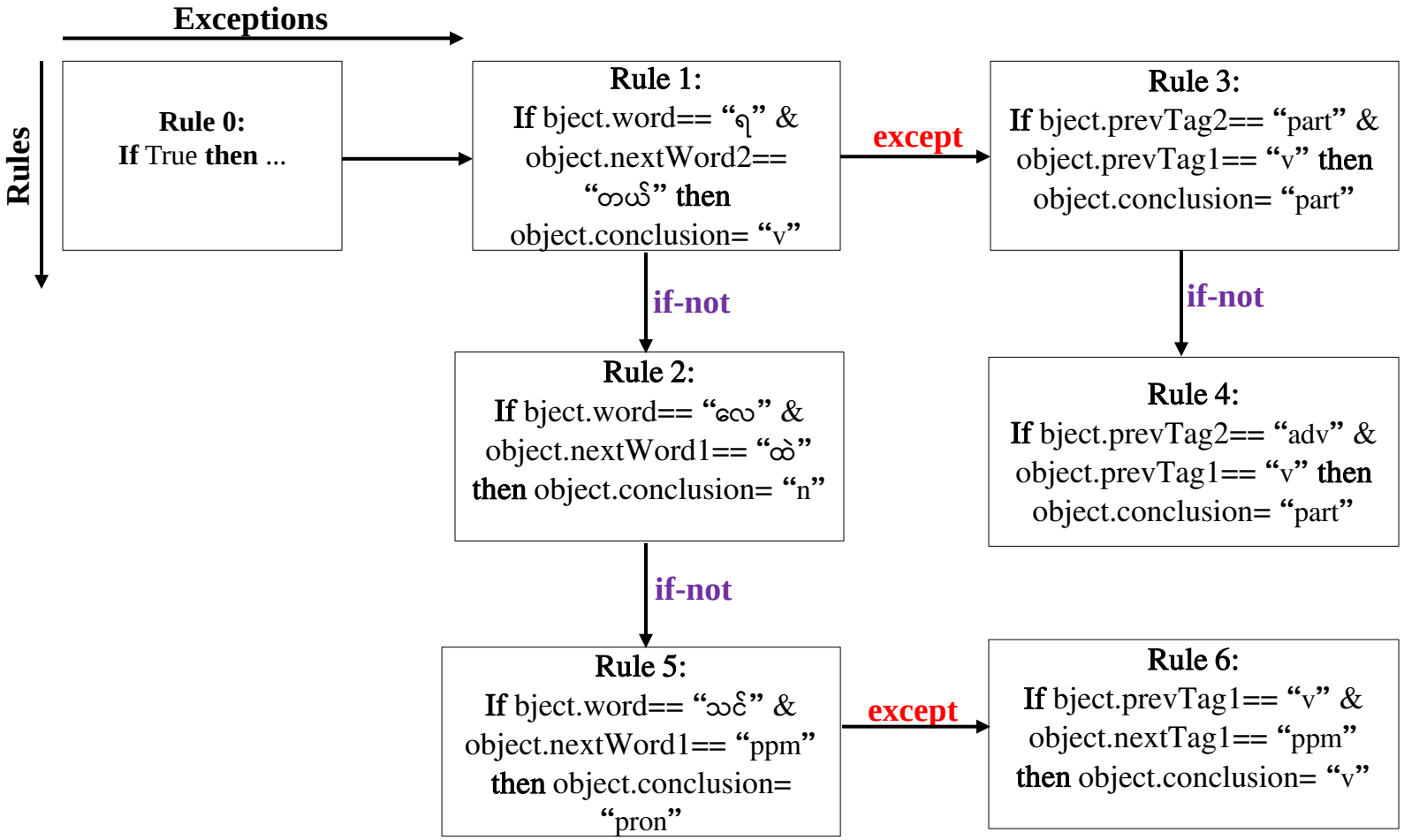

Figure: A binary tree of single classification ripple down rules

4. POS Tagging Methdologies (Cont.)

4.4. Open-source Neural Sequence Labeling Toolkit (NCRF ++)

- Toolkit for neural sequence labeling
- Provides the flexible running time and state-of-the-art results
- Designed for quick implementation of different neural sequence labeling models with a CRF inference layer
- Need the basic requirements of Python version 2 or 3 and PyTorch version 1.0
- Supports different structure combinations on **three levels** : character sequence representation, word sequence representation and inference layer.
  - ❖ Character sequence representation: character LSTM, character GRU, character CNN and handcrafted word features
  - ❖ Word sequence representation: word LSTM, word GRU, word CNN
  - ❖ Inference layer: Softmax, CRF

4. POS Tagging Methdologies (Cont.)

4.4. Open-source Neural Sequence Labeling Toolkit (NCRF ++)

- Users can compare twelve neural sequence labeling models (charLSTM, charCNN, None × wordLSTM, wordCNN × softmax, CRF)
- We compared two models : (1) CharLSTM+WordCNN+CRF (2) CharCNN+WordLSTM
- When building the network, users need only to edit the configuration file to configure the model structure, training settings and hyper-parameters
- Users can extend the toolkit by defining their own structure in any layer and integrate it into NCRF++

4. POS Tagging Methdologies (Cont.)

4.4. Open-source Neural Sequence Labeling Toolkit (NCRF ++)

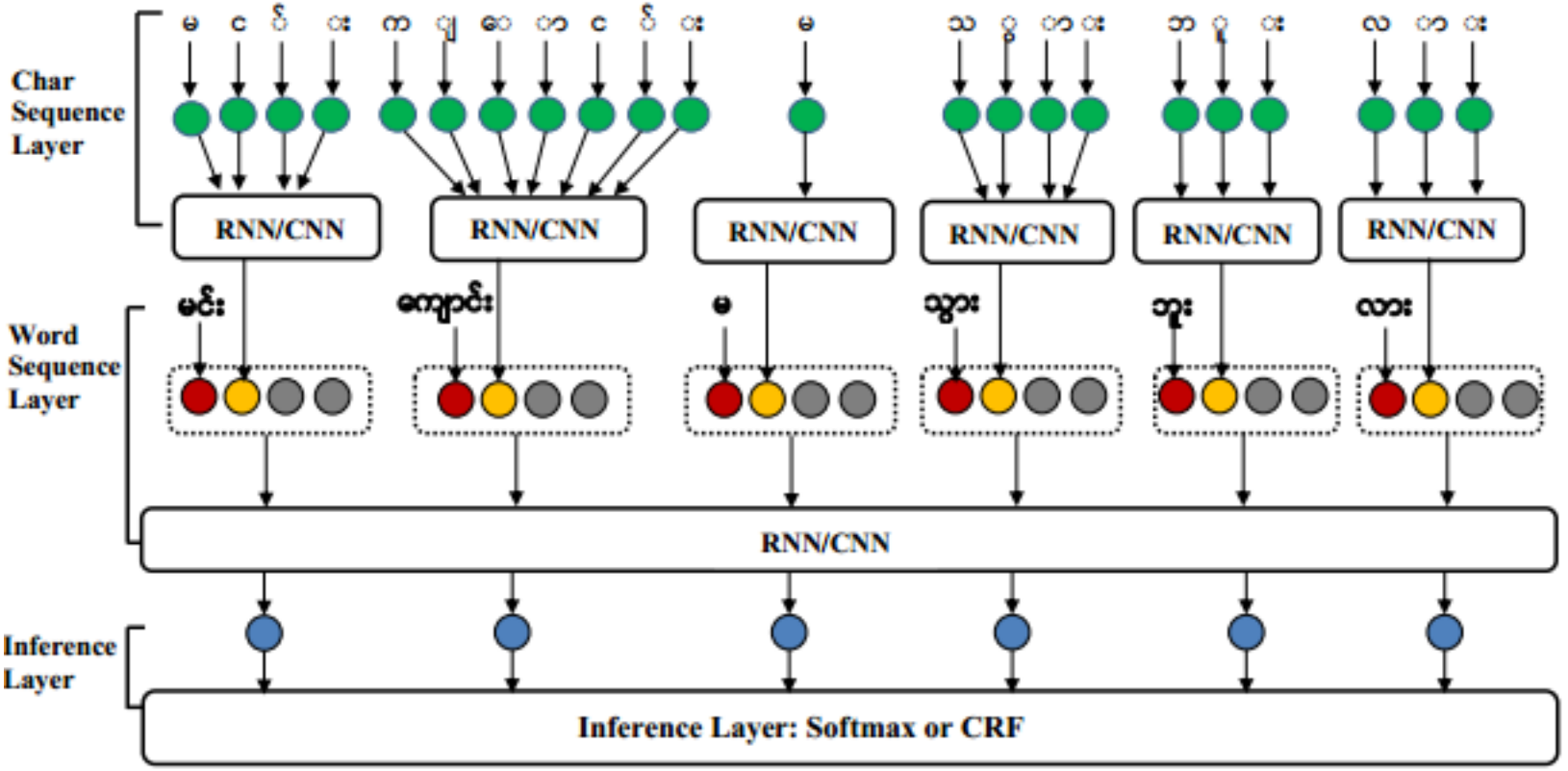

Figure: NCRF ++ for Burmese sentence “မင်းကျောင်းမသွားဘူးလား” (Don't you go to school?)

5. Experimental Setup

5.1. Corpus Extension

- We extended the original **myPOS** corpus from ***Myanmar-Chinese*** parallel corpus (Extension-1) and ***Myanmar-Korea*** parallel corpus (Extension-2) as **myPOS corpus version 2.0**.

- Word segmentation, POS tagging for each word and error checking was done manually

- In the corpus, the shortest sentence contains two words (for example: “သူခိုး။” ,“ thief” in English)

- The longest sentence contains 423 words

5. Experimental Setup (Cont.)

5.1. Corpus Extension

Myanmar-Chinese parallel corpus:

- “ Happy” ကို တရုတ် လို ဘယ်လို ပြော လဲ ။ (How to say“ Happy” in Chinese?)
- ပညာသင် ကာလ ၁ ခု ကို ကျောင်း လခ ဘယ်လောက် လဲ ။ (How much does it cost the tuition fee per semester?)
- ခဏ နေ မှ ပြန် တွေ့ မယ် ။ (See you soon.)

Myanmar-Korean parallel corpus:

- အလုပ် ပြီး ရင် မီး ပိတ် ပါ ။ (Turn off the light after doing work.)
- သတင်းစာ က နေ တဆင့် ကိုရီးယား ရဲ့ သတင်း ကို နားစွင့် နေ တယ် ။ (I am listening to Korean news through newspapers.)
- သူ့ ကို ကျေးဇူး ပြု ၍ ဒီ သတင်း ပို့ ပေး ပါ ။ (Please send this message to him.)

5. Experimental Setup (Cont.)

5.1. Corpus Extension

- Firstly, checked and corrected manually the segmented Myanmar sentences from our developing Myanmar-Chinese and Myanmar-Korean parallel corpora
- These two-segmented corpora are manually tagged and checked repeatedly
- We defined 10K POS tagged Myanmar sentences of **Myanmar-Chinese Parallel Corpus** as **Extension-1**
- Another 11K POS tagged Myanmar sentences of **Myanmar-Korean Parallel Corpus** as **Extension-2**
- These two different domains POS tagged sentences are combined with the original **myPOS** corpus to obtain the extended **myPOS** corpus **version 2.0**
- The extended corpus becomes 31,052 sentences, which approximately triple size of the original **myPOS** corpus

5. Experimental Setup (Cont.)

5.1. Corpus Extension

Table: Data Comparison between Original Baseline Corpus and Extended Corpus (myPOS corpus version 2.0)

Count	Original Corpus	Extended Corpus		
	myPOS	Extension-1	Extension-2	myPOS (version-2)
Total no. of Sentences	11,000	10,000	10,052	31,052
Total no. of Words	239,598	103,909	106,864	450,371
Average Words per Sentence	21.78	10.39	10.63	14.50

5. Experimental Setup (Cont.)

5.2. Training Data and Test Sets

- For the comparison of the original **myPOS** model and the extended **myPOS corpus version 2.0** model, we used two training data sets
- One for the baseline myPOS model which is taken from the whole **myPOS** corpus
- Another one is taken from the whole extended **myPOS corpus version 2.0** for the extended model
- Two types of test sets: closed test-set and open test-set
- 50% of training data from original **myPOS** corpus (5,500) is taken for closed test-set (ctest)
- Defined two open test sets: otest-1, otest-2
- For otest-1 and otest-2, we used sentences from the travel and tour domain **ASEAN-MT** corpus

5. Experimental Setup (Cont.)

5.2. Training Data and Test Sets

- Randomly selected 6,072 sentences (50 % of ASEAN-MT corpus) for otest-1
- The whole 12,144 POS tagged sentences of **ASEAN-MT** domain data are used for otest-2 (i.e. big size open test set)
- Myanmar sentences from Thai-Myanmar parallel corpus of **ASEAN-MT** corpus are manually tagged
- We thoroughly checked the manual POS tagged sentences to get the optimal open test sets

5. Experimental Setup (Cont.)

5.3. Evaluation

- To evaluate the effect of our extended POS corpus on automatic tagging task, we conduct the experiments on several test sets

- We measured the POS tagging performance by using the **Accuracy** defined in the following equation:

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

5. Experimental Setup (Cont.)

5.4. Evaluation

- For our experiments, we used the following open source POS Taggers :

  - ❖ CRFsuite tool (version 0.12) (https://github.com/chokkan/crfsuite)
  - ❖ *NCRF++* toolkit for neural sequence labeling (https://github.com/jiesutd/NCRFpp)
  - ❖ Jitar (version 0.3.3) for HMM (https://github.com/danieldk/jitar)
  - ❖ RDRPOSTagger (version 1.2.3) for RDR model (https://github.com/datquocnguyen/RDRPOSTagger)

6. Results and Conclusion

Table: Comparison of accuracies between the original myPOS corpus and the extended myPOS corpus version 2.0 among various automatic tagging methods

Methods	ctest: Closed Test-set		otest-1		otest-2	
	Original myPOS	myPOS (version 2.0)	Original myPOS	myPOS (version 2.0)	Original myPOS	myPOS (version 2.0)
CRFs	**98.40 %**	98.19 %	91.69 %	**94.61 %**	91.79 %	**94.75 %**
NCRF++ (CharLSTM+WordCNN+CRF)	**96.50 %**	95.10 %	91.07 %	**91.65 %**	90.15 %	**94.64 %**
NCRF++ (CharCNN+WordLSTM)	**97.61 %**	97.40 %	93.14 %	**95.79 %**	93.23 %	**95.98 %**
HMM	**97.14 %**	96.44 %	93.66 %	**94.97 %**	93.78 %	**95.06 %**
RDR	**98.43 %**	98.35 %	95.10 %	**97.54 %**	95.15 %	**97.57 %**

6. Results and Conclusion (Cont.)

- We extended the original **myPOS corpus version 1.0** to **myPOS corpus version 2.0** that is approximately triple size of the original corpus

- When two models are applied to the closed test-set (ctest) from the original **myPOS** corpus, the model trained from the original **myPOS** corpus provides higher accuracies for all methods

- This is not surprised because the test-set and the training set are from the same corpus

6. Results and Conclusion (Cont.)

- The model trained from the **extended myPOS corpus version 2.0** provides comparable accuracies but slightly lower accuracies due to its wider scope of tagging ability

- However, when two models are applied to the open test-set outside of training corpus, the benefit of the **extended myPOS corpus version 2.0** becomes apparent

- The models trained from the **extended myPOS corpus version 2.0** provide the improvement about **1% to 3 %** higher accuracies for all automatic tagging methods trained from the very strong baseline **myPOS corpus**

- When the open test-set is larger, the model provides higher accuracies

6. Results and Conclusion (Cont.)

- In conclusion, the larger the training dataset for POS tagging, the better the model for any automatic tagging methods

- Therefore, the benefit of this manually **extended myPOS corpus version 2.0** is apparent from the evaluation experiments

- In the near future, we plan to do this **extended myPOS corpus version 2.0** as publicly available resource for NLP R&D

References

- K. W. W. Htike, Y. K. Thu, Z. Zhang, W. P. Pa, Y. Sagisaka and N. Iwahashi, “ Comparison of Six POS Tagging Methods on 10K Sentences Myanmar Language (Burmese) POS Tagged Corpus” , at 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary, April 17–23, 2017.
- K. Zin, “ 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, pp. 123–128, 2009.
- J. Lafferty, A. McCallum and F. C. N. Pereira, “ Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data” , In Proceedings of the Eighteenth International Conference on Machine Learning, ICML ’ 01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, pp. 282–289, 2001.
- J. Yang and Y. Zhang, “ *NCRF ++*: An Open-source Neural Sequence Labeling Toolkit” , Proceedings of ACL 2018, System Demonstrations, pp. 74–79, July 2018, http://aclweb.org/anthology/P18-4013.
- P. Boonkwan and T. Supnithi, “ Technical Report for The Network-based ASEAN Language Translation Public Service Project,”Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013.
- Okazaki, Naoaki, “ CRFsuite: a fast implementation of Conditional Random Fields (CRFs)” , 2007.
- T. Brants,“ TnT: A Statistical Part-of-speech Tagger” , In Proceedings of the Sixth Conference on Applied Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 224–231, April 2000.
- de Kok, Daniël, “ Jitar: A simple Trigram HMM part-of-speech tagger” , 2014

Thank you