CICLing 2017

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

Khin War War Htike[†], Ye Kyaw Thu[λ,φ], Zuping Zhang[†],
Win Pa Pa[‡], Yoshinori Sagisaka[φ], Naoto Iwahashi[λ]
[†]Central South University, Changsha, China
[λ]AI Lab., Okayama Prefectural University (OpU), Okayama, Japan
[‡]NLP Lab.,University of Computer Studies Yangon (UCSY), Yangon, Myanmar
[φ] Language and Speech Science Research Lab., Waseda University, Tokyo, Japan

1. Introduction

- Part-of-Speech (POS) tagging is an important issue in natural language processing (NLP)
- A robust Myanmar POS tagger is necessary for Myanmar NLP research and not available publicly yet
- We developed a manually annotated ten thousand (10K) sentences POS tagged corpus for the general domain
- Evaluated with six POS tagging approaches, CRFs, HMM, MaxEnt, SVM, Ripple Down Rules-based (RDR) and Two hours of annotation approach (i.e. combination of HMM and MaxEnt)

3. Statistic of POS Tag-set

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%

5. Accuracy on Training Data Size

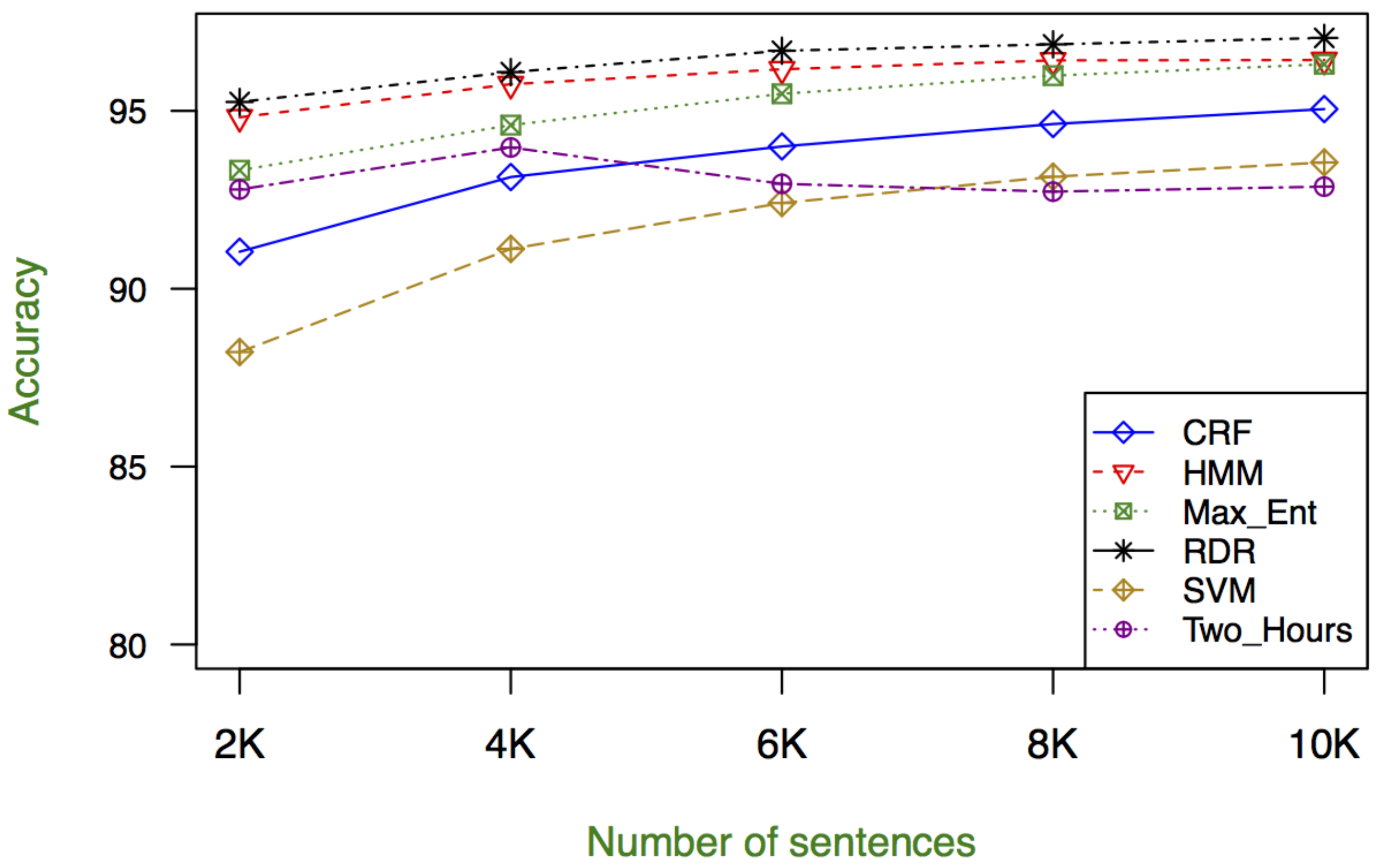

Fig. Accuracies of six POS tagging methodologies on varying training data sizes

2. Proposed POS Tag-set

- Based on 10 POS tag-set defined by Myanmar Language Commission
- 16 POS are used to meet futher NLP processing such as semantic processing
- abb (Abbreviation), adj (Adjective), adv (Adverb), conj (Conjunction), fw (Foreign Word), num (Number), int (Interjection), n (Noun), part (Particle), part_neg (Negative Particle), ppm (Post Positional Marker), pron (Pronoun), punc (Punctuation), sb (Symbol), tn (Text Number), v (Verb)

4. Result of Six Methodologies

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%

6. Error Analysis

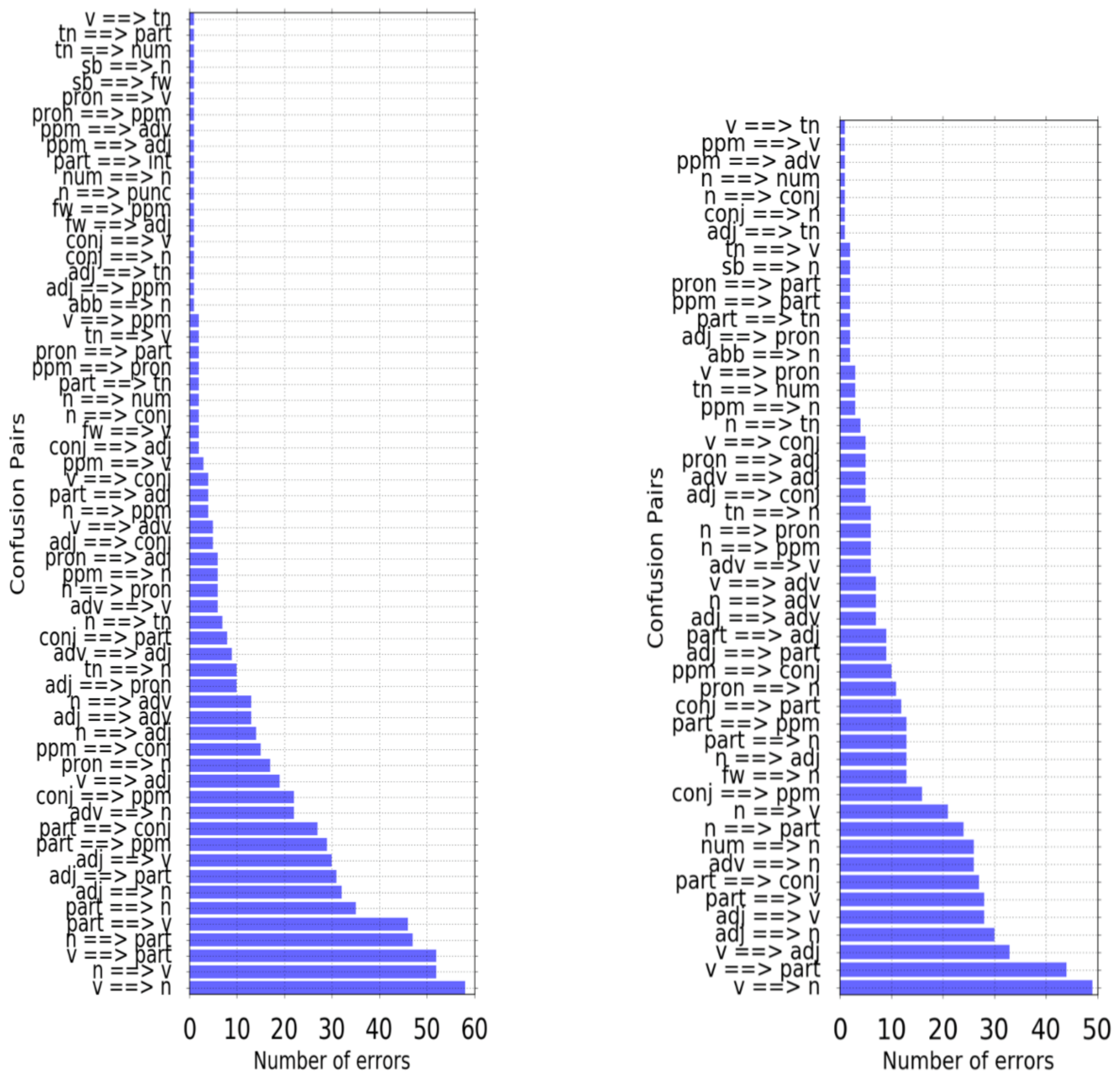

Fig. Confusion pairs with 10K model. Left: 3gHMM, Right: RDR