A Comparative Study on Different Techniques for Thai Part-of-Speech Tagging

Jaruwat Pailai[#], Rachada Kongkachandra[#], Thepchai Supnithi[*], Prachya Boonkwan[*]

[#]*Department of Computer Science, Faculty of Science and Technology*
Thammasat University, Thailand 12121
[1]pomkab@gmail.com
[2]rdk@cs.tu.ac.th

[*]*Language and Semantic Technology Laboratory*
National Electronics and Computer Technology Center (NECTEC), Thailand 12120
[3]thepchai.supnithi@nectec.or.th
[4]prachya.boonkwan@nectec.or.th

***Abstract*— The natural language processing (NLP) for Thai language is rather complicated using in the real tasks because it has a complex sequential structure of the sentence. The POS tagging can improve the accuracy of syntactic analysis so it can support the improvement of many NLP tasks. We present the supervised machine learning that is suitable for annotate the POS type for Thai language by comparison between the Support Vector Machine (SVM) and the Conditional Random Fields (CRFs). The BEST 2012 News and Entertainments corpus is utilized in our experiments. However, the sequential characteristic of Thai language is the interesting point and we use it as our feature in training set. Our sequential features contain forward 3-gram, backward 3-gram and 5-gram. The best accuracy of our experiments is 93.638% from SVMs POS tagging that learning by word of forward 3-gram when the size of training data is ten thousand tokens. Moreover, with the same training data, the best accuracy of CRFs is very close with SVM that is 93.254% when the learning form is the word with POS of 5-gram.**

***Keywords*— Thai Part of Speech Tagging, Natural Language Processing (NLP), Support Vector Machine (SVM), Conditional Random Fields (CRFs), N-Gram**

I. INTRODUCTION

Part Of Speech (POS) tagging is a task of assigning words of texts with a proper part of speech tag. As the POS tagging can improve the exactness of syntactic analysis so many Natural Language Processing tasks that are applied POS tagging also get higher accuracy for the example machine translation, information extraction and linguistic research for corpora [1]. The satisfactory POS tagging can improve any NLP algorithm for more understanding language's meaning.

The Thai language is an analytic language so it has the problem of word segmentation and impact to POS tagging in morphological analysis. The general problem of POS tagging is "one word can have POS type more than one." In addition, the human language has the expressive power more than the general sequence of words in the sentence. At presents, there are a lot of approach that are purpose for POS tagging such as rule-based approach and statistical approach. The statistical method is popular in NLP field more than others because the statistical method can calculate the statistic value automatically. While the rule-based have to employ the hand-writing rules and the rules can not cover all pattern of sentence in the natural language. Many supervised machine learning techniques are popular to utilized in Thai POS tagging task.

Our approach is comparative on different supervised machine learning for Thai part of speech tagging. The comparative technique contains the Support Vector Machine (SVM) and The Conditional Random Fields (CRFs). The features of learning set are assigned into two forms that are the word without POS and the word with POS. For all training sets, we assign forward 3-gram, backward 3-gram and the combination of forward and backward 3-gram (5-gram) for our experiment respectively.

The remainders of this paper is as follows. The related techniques of part of speech tagging are described in Section II. In Section III, we explain about our Thai part of speech tagging, which include related Thai language corpus and the feature for learning process. Section IV represent the result of the experiment. Error Analysis and discussion will be explained in Section V. Conclusion is given in the last section.

II. RELATED WORK

For many techniques of part of speech tagging, the supervised machine learning is applied. The researchers try to find the suitably feature for train their classifier for the example rule-based tagger, stochastic, or transformation-based learning approaches. The rule-based taggers [2], [3], [4] assigned a tag to each word using a set of rules that created by human expert. These rules could specify such as the word that follows a determiner and an adjective must be a noun. However, the rule-based construction has many problems for the example time consuming, a lot of cost, the conflict among the rules and the constructed rules cannot cover all of contexts in the language. So the rule-based tagger is too hard to assign the all rules of Thai language sentences.

For the first time of part of speech tagging research, brill tagger was developed under the Error-driven transformation-based tagger [5], [6] to increase the performance of part of

978-1-4799-0545-4/13/$31.00 ©2013 IEEE

speech tagging. And it was compared with another related task such as N-gram and HMM. [7] purpose the Bangla POS tagging by using unigram, HMM and Brill respectively. Then they iterated their method with Brown corpus. The result was different. HMM got the best accuracy and more than Brill and unigram respectively.

One of interesting machine learning technique is Conditional Random Fields. Wallach04 presented the supervised learning CRFs that was developed from HMM had the performance of sequential data classification better than HMM. [8], [9], [10].

For Thai part of speech tagging, it is a research field that take an interest for a long time. Murata et al. [11], [12] tried to solve Thai part of speech tagging problem and compared the performance among eight techniques for machine learnings. Their conclusion represented the performance of Support Vector Machine (SVM) was more than the Hidden Markov Model (HMM).

Besides the supervised machine learning, there are researches that try to construct the POS tagger by using hybrid method. Bryan J. [13] purposed the use of a rule-based morphological component to extend traditional HMM techniques by the inclusion of lexical class probabilities and used dwdst system to reduce the search space of HMM model. The rule base system was utilized for hybrid method in many times. Sandipan D. et al. [1] build the rule based approach from the Bengali word's context that they claimed it is a highly ambiguous and relatively free word order language. Their rules based were combined with the modified HMM. Viterbi algorithm was applied to calculate the best probable path for a given word sequence for HMM. And the hybrid system took the accuracy more than the one of ruled based or unmodified HMM. The maximum entropy with HMM was purposed by Radu S. [14]. They used the statistical model with a rule based system to decrease the word ambiguity. Linguist experts constructed their rules.

From the conclusion of Murata et al., and Wallach04, we purpose the comparison between SVM and CRFs to find the appropriate technique and search for the proper training feature. The significant techniques are utilized for our preparation of hybrid method for Thai part of speech tagging.

III. METHODOLOGY

For our experiment, we have to prepare the Thai corpus for training set and find the feature of instance that use in the learning process. The supervised machine learning that we used contains the Support Vector Machine (SVM) and the Conditional Random Fields (CRFs). The features of learning set are assigned into two forms that are the word without POS and the word with POS.

A. Corpus

We used the BEST 2012 News and Entertainment Corpus, a Thai language corpus. It is developed by National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA). The contents in corpus relate with news and entertainments. The corpus is tagged part of speech 34 types and is annotated corpus by linguists and expert. The example of data in the corpus is shown in Figure 1.

<NE>บอร์ด กทพ.</NE>|ชี้/VV|ขาด/VV|ขึ้น/VV|ค่า/NN|ทาง/NN|ด่วน/JJA|24/OD|<AB>ก.ค.</AB>|
นี้/DDEM| |คณะ/NN|อนุกรรมการ/NN| |<NE>กทพ.</NE>| |เห็น/VV|ชอบ/VV|ขึ้น/VV|ค่า/NN|
ทาง/NN|ด่วน/JJA| |รถ/NN|ยนต์/NN| |4/CD| |ล้อ/CL| |จาก/P| |40/OD| |เป็น/P| |48/CD| |บาท/CL|
เตรียม/VV|เสนอ/VV|บอร์ด/NN|ใหญ่/JJA| |24/OD| |<AB>ก.ค.</AB>|นี้/DDEM|ชี้/VV|ขาด/VV|
เสนอ/VV|<NE>กระทรวงคมนาคม</NE>|อนุมัติ/VV|ต่อ/ADV|ไป/AUX|

Fig. 1 The example of data in the BEST 2012 News and Entertainment corpus

Figure 1 shows the content in Thai language that was tokenized by "|" symbol. Each word was tagged by POS which separate by slash "/" for the example "ชี้/VV" shows the word "ชี้" was tagged to "VV" (Verb), "ค่า/NN" shows the word "ค่า" was tagged to "NN" (Common noun) and "ด่วน/JJA" shows the word "ด่วน" was tagged to "JJA" (Adjective of noun).

B. Feature

The sequential data that we used in the learning process is N-Gram of words. T_i is out current state. For the feature of forward 3-gram, we assigned

$$T_i \quad T_{i+1} \quad T_{i+2}$$

where T_i is our current state. T_{i+1}, T_{i+2} are two subsequent words. And the feature of backward 3-gram, we assigned

$$T_{i-2} \quad T_{i-1} \quad T_i$$

where T_i is our current state. T_{i-1}, T_{i-2} are two previous words.

For Thai language, the position of modifier is almost back of noun or pronoun. The example is shown in Figure 2. The tag NN is noun and the tag JJA is adjective of noun. The sequence is different from English that the adjective's position is in front of noun.

คุณภาพ/NN | ดี/JJA
good/JJA | quality/NN

Fig. 2 The example position of modifiers

As for 5-gram, we combined the forward 3-gram and backward 3-gram together. So we chose the word sequence follow

$$T_{i-1} \quad T_{i-2} \quad T_i \quad T_{i+1} \quad T_{i+2}$$

where T_i is our current state. T_{i-1}, T_{i-2} are two previous words and T_{i+1}, T_{i+2} are two subsequent words.

C. Classifiers and Parameter

We utilized two machine learnings for our experiment. There are

1) The support vector machine (SVM): SVM is a machine learning that separates group of data by maximum margin. In our task, we used a Library for Support Vector

Machines (LIBSVM) [15] as the first classifier. We used two forms for the parameter of SVM. First is the word without POS. The current word's parameter is assigned by the probability of the word occurs in the answer class. The parameters of other words are set to one.

The second form is the word with POS. All word's parameter is assigned by the probability of the word occurs in their class.

2) Conditional Random Fields (CRFs): CRFs is a statistical model to segment and label sequence data. In our experiment, we used CRFsuite [16] library as the second classifier. We used two from for the CRF's learning. For the first the word without POS, the learning observes only words in the sequence. And the second the word with POS, the learning also observes words and labeling of word in the sequence.

IV. Experiment Result

For all our experiment, the 10-fold cross-validation is applied with the BEST 2012 corpus. From the Table I, the accuracy results of SVM and CRFs are close. But the best result of SVM gets by the word without POS parameter when we use forward 3-gram feature. It take the best accuracy of SVM at 93.638% that resembling with the best of CRFs. CRFs technique get the highest accuracy at 93.254% when we set the learning is the word with POS form and used 5-gram feature.

From the results, the comparative techniques between the word parameter shows SVM is suitable for the word without POS while CRFs is proper for the word with POS. In the comparative by feature, 3-gram is appropriate with SVM but CRFs get the more accuracy when we used 5-gram.

From the Figure 3, when the training set contains 40,000 tokens, the accuracy of both techniques are slightly increase. We investigated this case by plot the learning curve. The SVM's step the accuracy is more than CRFs which learned by the word without POS form. However CRFs took the accuracy higher than SVM when we compare them with the same data set and feature.

From Table II, The most the number of mistake is NN and VV respectively that their CRFs's results are better than SVM. NN is POS type that has the most of proportion in corpus that is 22.016% of the BEST corpus. Nevertheless, NN is also POS type which has the most of error percent 2.763% from the best SVM's result. While the error of NN tagging on CRFs is 0.979%. Unlike the most of error from CRFs is VV 1.071%. This POS type is the second large number of BEST corpus that is 20.866% but the error percent of VV on SVM still more than CRFs that are 1.833%. On the other hand, the error percent of ADV, JJV and JJA on SVM are less than the error rate from CRFs.

TABLE I
THE ACCURACY RESULTS

Tokens	Words						Words and POS					
	3-gram				5-gram		3-gram				5-gram	
	Backward		Forward				Backward		Forward			
	SVM	CRFs	SVM	CRFs	SVM	CRFs	SVM	CRFs	SVM	CRFs	SVM	CRFs
10,000	85.040%	80.550%	86.170%	82.230%	83.220%	83.260%	86.370%	81.960%	86.300%	82.670%	84.270%	84.900%
20,000	87.995%	83.815%	89.845%	86.290%	85.930%	87.025%	88.190%	85.650%	88.835%	86.895%	86.960%	86.645%
30,000	89.923%	86.337%	91.697%	88.770%	88.223%	89.567%	89.770%	87.980%	90.413%	89.227%	89.277%	90.727%
40,000	90.703%	87.348%	92.395%	89.476%	89.080%	90.223%	90.425%	88.823%	90.783%	89.983%	90.355%	91.480%
50,000	91.486%	88.136%	92.806%	90.198%	89.568%	90.862%	90.864%	89.443%	91.296%	90.632%	90.460%	92.068%
60,000	91.700%	88.510%	93.070%	90.577%	89.687%	91.235%	90.900%	89.755%	91.190%	90.968%	90.825%	92.202%
70,000	91.687%	88.811%	92.966%	90.891%	89.943%	91.569%	91.134%	90.130%	91.373%	91.293%	91.260%	92.526%
80,000	91.850%	89.055%	93.123%	91.101%	90.144%	91.868%	91.180%	90.450%	91.444%	91.449%	91.428%	92.775%
90,000	92.200%	89.472%	93.453%	91.337%	90.109%	92.252%	91.407%	90.901%	91.592%	91.700%	91.717%	93.130%
100,000	92.372%	89.796%	93.638%	91.560%	90.204%	92.521%	91.533%	91.663%	91.705%	91.916%	91.868%	93.254%

TABLE II
THE DETAIL TOP 5 ERROR IN THE BEST OF SVM AND THE BEST OF CRFS

POS	Proportion in corpus	Words, Forward 3-gram				Words and POS, 5-gram			
		Error	Precision	Recall	F-measure	Error	Precision	Recall	F-measure
VV	20.866%	1.833%	94.382%	87.380%	90.746%	1.071%	90.538%	94.865%	92.651%
NN	22.016%	2.763%	96.138%	91.485%	93.754%	0.979%	89.689%	95.551%	92.527%
ADV	2.942%	0.275%	75.951%	89.048%	81.980%	0.703%	86.827%	76.121%	81.122%
JJV	1.422%	0.138%	60.084%	86.103%	70.778%	0.708%	87.302%	50.246%	63.782%
JJA	1.511%	0.109%	77.778%	91.518%	84.090%	0.455%	87.791%	69.907%	77.835%

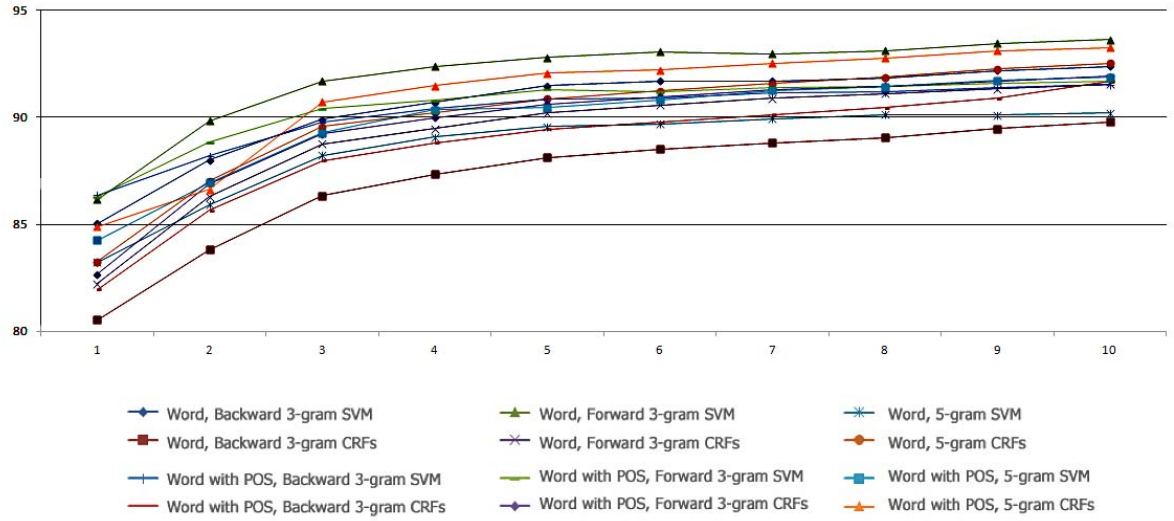

Fig. 3 Learning curve

V. Error Analysis and Discussion

From the results of Table II, the quantity of error of NN and VV has a direct effect to their precision and accuracy. So we might solve this problem by applied with the rule based method to improve the results. On the other perspective, the quality of JJV should be get the improvement. Because the JJV recall of SVM and CRFs are low. From the proportion of JJV in the corpus, the size of training data affect to the quality of the learning of POS tagging.

Table III, IV and Figure 4 represent the error with two directions. The two most proportion errors are NN and VV that are VV is tagged to NN and NN is tagged to VV. This mistake affects to their precisions. And one of interest error is JJV which is often confused to NN and VV. So the recall of JJV is low.

From the results, the best accuracy of two techniques are very close. When we compare among the running time. LIBSVM used the time around nine hours that more than CRFsuit approximately 40 minutes. So the CRFsuite learning time is faster than LIBSVM and take the resemble accuracy.

Table V shows POS types that are error tagged from tagging process. From all 34 tag set, the error tag set of CRFs is less than SVM's. For the example, CRFs has error tag of VV 16 tags from 17 possible tags but SVM's has 25 tags from 26 possible tags. We can conclude the CRFs is suitable for apply in the hybrid method more than SVM since the size of CRFs possible error tag set is less than SVM. It can support to decrease the necessary rule based and take easier when we group the error tag for build the local learning.

VI. Conclusions

Because the human language has the expressive power more than the general sequence of words in the sentence and one word can has a more than one part of speech type. It is a challenge for POS tagging on Natural Language Processing research field for a long time.

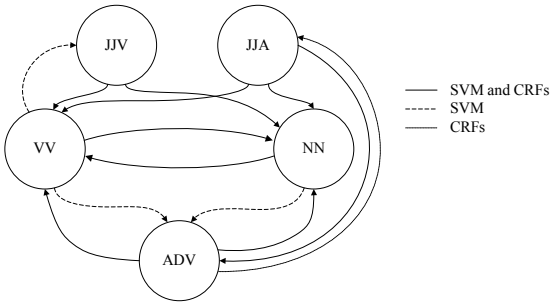

Fig. 4 The error's directions of SVM and CRFs

We propose the Support Vector Machine (SVM) to solve this problem by using 3-gram and 5-gram as a feature and compare with the Conditional Random Fields (CRFs). The results of our experiment prove the SVM get the accuracy more than CRFs a bit.

TABLE III
The example of confusion matrix of the best of SVM

POS	NN	VV	JJV	JJA	ADV
NN	19,145	679	182	159	238
VV	359	19,705	353	59	270
JJV	23	76	855	9	7
JJA	25	13	9	1,176	45
ADV	47	65	0	76	2,236

TABLE IV
The example of confusion matrix of the best of CRFs

POS	NN	VV	JJV	JJA	ADV
NN	21,049	490	24	12	35
VV	652	19,806	52	3	63
JJV	166	520	715	8	6
JJA	180	125	16	1,057	103
ADV	222	325	2	84	2,241

Moreover, the number of training data also has effect to the accuracy. When we train the classifier model with the more training data, the accuracy result is more than old. So our experiment shows the SVM model get the best accuracy 93.638% with 3-gram feature on the corpus that contains 100,000 tokens. In the future, we plan to apply the locality learning to the group of POS type that has the error with two directions. The locality learning can improve precision and recall of the confusing tagging. In the other way, our approach can apply to develop the hybrid approach. The hybrid method may use CRFs with the rule based from linguist expert or a data-driven error. Moreover, the increasing of the corpus size can take the more precision for the learning process. And we try to develop Thai POS tagging for utilization of any Natural Language tasks.

TABLE V
THE NUMBER OF ERROR TAG IN THE BEST OF SVM AND THE BEST OF CRFS

POS	SVM	CRFs
NN	26	17
VV	25	12
ADV	21	14
JJV	13	11
JJA	16	9

ACKNOWLEDGMENT

BEST 2012 News and Entertainment corpus is supported by Language and Semantic Technology Laboratory (LST), National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA).

REFERENCES

[1] S. Dandapat, S. Sarkar, A. Basu. A hybrid model for Part-of-Speech tagging and its application to Bengali, in Proc. IJCI - IJIT - IJSP – Conferences, 2004, p. 169-172.

[2] B. Greene and G. Rubin, “Automatic Grammatical Tagging of English”, Technical Report, Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

[3] S. Klein and R. Simmons, “A computational approach to grammatical coding of English words”, Journal of the ACM (JACM), vol. 10, p. 334-347, July. 1963.

[4] Z. Harris, String Analysis of Language Structure, Mouton and Co., The Hague, 1962.

[5] E. Brill, Transformation based error driven parsing, in Proc. the Third International Workshop on Parsing Technologies, 1993, p.1-13.

[6] E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging, Computational Linguistics, vol. 21, p. 543-565, December. 1995.

[7] F. Muhammad Hasan, N. UzZaman and M. Khan. Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill’s tagger) for Bangla, in Proc. Advances and Innovation in Systems, Computing Sciences and Software Engineering. 2007, p. 121-126.

[8] M. Murata, Q. Ma, H. Isahara, Part of Speech Tagging in Thai Language Using Support Vector Machine. in Proc. the Second Workshop on Natural Language Processing and Neural Networks (NLPNN), 2001, p. 24–30.

[9] M. Murata, Q. Ma, H. Isahara, Comparison of three machine-learning methods for Thai part-of-speech tagging. ACM Transactions on Asian Language Information Processing (TALIP), vol. 1 issue 2, p. 145-158, June. 2002.

[10] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation, in Proc. International Conference on Machine Learning, 2000, p. 591-598.

[11] D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. in Proc. the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, p. 235-242.

[12] F. Sha and F. Pereira. Shallow parsing with conditional random fields. in Proc. the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, p. 134-141.

[13] B. Jurish. A Hybrid Approach to Part-of-Speech Tagging. Final Report Berlin-Brandenburgische Akademie der Wissenschaften, 2003.

[14] R. Simionescu. Hybrid POS Tagger, in Proc. Language Resources and Tools with Industrial Applications, 2011, p. 21-28.

[15] (2012) LIBSVM A Library for Support Vector Machines website. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[16] (2012) CRFsuite A fast implementation of Conditional Random Fields (CRFs).[Online]. Available: http://www.chokkan.org/software/crfsuite/