AutoTagTCG : A Framework for Automatic Thai CG Tagging

**Thepchai Supnithi, Taneth Ruangrajitpakorn,
Kanokorn Trakultaweekool, and Peerachet Porkaew**

Human Language Technology
National Electronics and Computer Technology Center
112 Thailand Science Park, Phahonyothin Road, Klong 1,
Klong Luang Pathumthani, 12120, Thailand
+66-2-564-6900 Ext.2542, Fax.: +66-2-564-6772
{thepchai.sup, taneth.rua, kanokorn.tra, peerachet.por}@nectec.or.th

Abstract

Recently, categorical grammar has been focused as a powerful grammar. This paper aims to develop a framework for automatic CG tagging for Thai. We investigated two main algorithms, CRF and Statistical alignment model based on information theory (SAM). We found that SAM gives the best results both in word level and sentence level. We got the accuracy 89.25% in word level and 82.49% in sentence level. SAM is better than CRF in known word. On the other hand, CRF is better than SAM when we applied for unknown word. Combining both methods can be suited for both known and unknown word.

1. Introduction

Thai language resources [Asanee et al. 2002, Chai & Sadaoki 2007] are gradually developed. However, only few of them are practical and provided. There are several researches focusing on developing practical language resources in Thai, such as, BEST [BEST Corpus 2009] which develops a 7-million word corpus for word segmentation. Asian Wordnet, which focuses on developing an infrastructure in terms of word concept. Orchid corpus [Thatsanee 1997], a POS tagging corpus contains 36,457 sentences with 43 POS types in tagset. There are several dependencies Treebank under developing step [Vee & Asanee 2004]. BEST is only a corpus which distributes to various researchers with enough amounts of data. The most effective word segmentation using BEST is around 97% based on F-measure. Others are inadequate for developing reliable NLP applications for Thai.

Categorial Grammar (CG) [Bob 1992, Kazimierz 1935] and Combinatorial Categorial Grammar (CCG) [Mark 2000] are formalisms in natural language syntax which focuses on compositional principle of syntactic constituents. There are several researches on automatic CCG tagging [James et al, 2007, Stephen 2002]. All of them are originated from CCG bank [Mark 2000, Julia & Mark 2002] which is developed from Penn Treebank. Recently, CG dictionary [Taneth et al 2007] and CG Treebank [Taneth et al 2009] are developed in Thai.

With limited time and human resources, it is necessary to develop a framework for implementing an automatic CG tagging. We investigate two main algorithms. One is CRF which becomes a well-known algorithm for sequence labelling problem. Another is a statistical alignment model based on information theory, which adapted from Noisy channel model. It focuses on CG pattern and maps between word and CG sequentially.

This paper is organized as follows. Section 2 explains Thai CG dictionary and Treebank. Section 3 illustrates the overview of our system work flow. Section 4 explains two methodologies which are applied in this work. Section 5 shows experiments which compare between two main algorithms. Finally discussion, conclusion and future work are explained in section 6.

2. Thai CG

Categorial Grammar (CG) is a formalism which focuses on principle of syntactic behavior. It can be applied to solve word order issues in Thai. To apply CG for machine learning and statistical based approach, CG Treebank, is initially required. There are two main resources in Thai CG, CG dictionary and CG Treebank.

CG dictionary presently contains 70,441 lexical entries with 89 CG syntactic categories. For Thai language, six argument syntactic categories are determined. Thai CG arguments are listed with definition and examples in Table 1. Additionally, np, num, and spnum are Thai CG arguments that can directly tag to a word, but other can only be used as a combination for other arguments.

Recently, CG Treebank has been developed. Currently there are 20,824 sentences in our Treebank. It plays an important role as a fundamental resource for other Thai NLP processing, such as automatic CG tagging, chunking and parsing. Figure 1 shows an example of CG tree.

Thai argument category	definition	example
np	a noun phrase	ช้าง (elephant), ผม (I, me)
num	A both digit and word cardinal number	หนึ่ง (one), 2 (two)
spnum	a number which is succeeding to classifier instead of proceeding classifier like ordinary number	นึง (one), เดียว (one)[1]
pp	a prepositional phrase	ในรถ (in car), บนโต๊ะ (on table)
s	a sentence	ช้างกินกล้วย (elephant eats banana)
ws	a specific category for Thai which is assigned to a sentence that begins with Thai word ว่า (that : sub-ordinate clause marker).	* ว่าเขาจะมาสาย [2] 'that he will come late'

Table 1: primitive CG in Thai

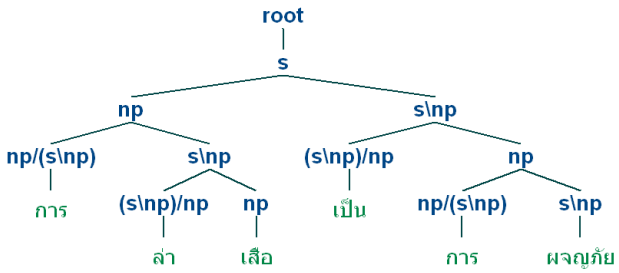

Figure 1: An example of CG tree

3. AutoTagTCG : An automatic Thai CG tagging

AutoTagTCG is an automatic Thai CG tagging. Figure 2 illustrates the flow of our system. Given a raw data, it is necessary to segmented in to word. We apply a Statistical approach for Word Segmentation Tool (SWS) [14] in this task. The output from SWS is constructed based on CG dictionary in order to obtain words with all possible CG information. The remaining words which are not included in CG dictionary will be recognized as unknown word. After we get a segmented corpus, we apply CG tagging process. This process is composed of two processes, CG tagging and CG tag prediction for unknown word algorithm. CG tree bank is applied to generate the all possible tagged sentences. Tagged sentences will be sent to CG tag verification to check the correctness of tagged results. Finally CG tagged corpus are constructed.

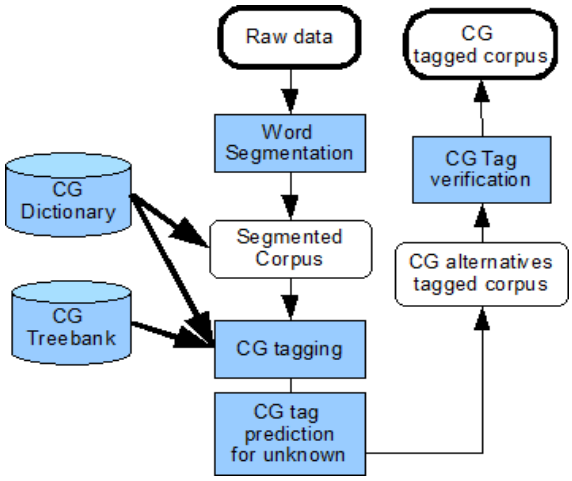

Figure 2: An AutoTagTCG system work flow

4. Tagging Methodology

We apply two methodologies, Conditional Random Field (CRF) [Lafferty 2001] and Statistical alignment modelling based on information theory (SAM).

4.1 Conditional Random fields (CRF)

CRF is an undirected graph model in which each vertex represents a random variable whose distribution is to be inferred, and edge represents a dependency between two random variables. CRF is widely used in sequential learning task. CRF are probabilistic models for computing the probability p (y|x) of a possible output y = (y1, …, yn) $\in Y^n$ given the input x = (x1,...,xn) $\in X^n$ which is also called observation. The general model formulation of CRF is derived as follow:

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(x_C, y_C)$$

4.2 Statistical alignment model based on information theory

This model is developed based on Channel model. We compute two parameters.

- Language model which represents the sequence of CG pattern
- Alignment Model which stands for the pairing of word and CG in both phrase level and sentence level.

The equation can be represented as follow:

$$t_{best} = \text{argmax}_t P_{AL}(t) * P_{LM}(t)$$

Language Model

$$P_{LM}(t) = P(t|t\text{-}1,t\text{-}2)$$

[1] This spnum category has a different usage from other numerical use, e.g. ม้า[noun,'horse'] ตัว[classifier] เดียว[spnum,'one'] 'lit: one horse'. This case is different from normal numerical usage, e.g. ม้า[noun,'horse'] หนึ่ง [num,'one'] ตัว[classifier] 'lit: one horse'

[2] This example is a part of a sentence ฉันเชื่อว่าเขาจะมาสาย 'lit: I believe that he will come late'

Alignment Model

$P_{AL}(t|w) = P(t_i|w_i)$ when i=1,2,3,4

when t, t_i represents tag. w_i represent the i^{th} word.

5. Experiment Setting and Results

We apply CG dictionary and CG Treebank to compare between CRF and SAM. In CRF model, we apply word as a feature set. 4-gram word is trained for tagging words. Example of CRF data is illustrated as follows

Sample of Input data in CRF

Word	**CG tag**
สิ่งของ	np/pp
ที่	pp/(s\np)
แสดง	(s\np)/pp
…	

In SAM model, bigram word is used in language model and 4-gram word aligns with CG tag is used in alignment model. Examples of SAM data are illustrated as follows. Pipe represents a word boundary.

Sample of Input data in SAM

Alignment Model

4-gram word:	4-gram CG tag
สิ่งของ|ที่| แสดง |ใน :	np/pp |pp/(s\np)| (s\np)/pp | pp/np
เขา |ขอร้อง |ให้| ฉัน :	np |(s\np)/pp| pp/s | np

…

Word based level		
Training Set	SAM	CRF
1	88.53	87.10
2	89.10	87.65
3	90.00	87.82
4	89.33	87.77
5	89.07	87.20
6	89.86	87.68
7	88.96	87.66
8	89.65	87.87
9	88.70	87.93
10	89.26	87.40
Average	89.25	87.61
Sentence based level		
	SAM	CRF
1	74.41	72.95
2	75.22	75.17
3	77.00	75.07
4	76.00	75.41
5	75.79	74.96
6	76.56	74.58
7	75.26	74.63
8	76.61	75.30
9	74.11	75.06
10	75.94	72.95
Average	75.69	74.61

Table 2: primitive CG in Thai

We evaluate the results using 10-fold cross validation. Table 2 shows the output result comparing between CRF and SAM in terms of word and sentence. It is obviously seen that SAM model gives the better results (89.25% for word level and 75.69% for sentence level tagging) when compare with CRF (87.61% for word level and 74.61% for sentence level tagging).

Training Set	SAM
1	85.75
2	84.39
3	82.72
4	82.19
5	85.35
6	81.56
7	80.06
8	80.26
9	80.64
10	81.99
Average	82.49

Table 3: Accuracy based on Best alternatives list

We investigate the potential to increase the accuracy of automatic CG tagging by analyzing output results (sentence level) of SAM in alternatives (vary from 1 to n) list. Table 3 Shows the output results based on n-fold cross validation.

We analyze the accuracy based on the correct alternatives. We found that 90.00% is founded at the first alternative, 7.50% is founded at the second alternative, 1.75% is founded at the third alternative and 0.75% for other alternatives, respectively.

Focusing on unknown word, as shown in table 4, we found that CRF gives the better result than SAM. This shows evidence that SAM produce better solution for known data, but CRF give a better results for unknown data.

6. Discussion

There are three major issues for the incorrect result. First, most of Thai vocabularies have various usages. A word is possibly designed into ten categories.

For instance, Thai word "เสมอ" can perform as
"s\np",
"(s\np)/pp",
"np\np",
"(np\np)/pp",
"(s\np)\(s\np)",
"((s\np)/pp)\((s\np)/pp)",
"(np\np)\(np\np)",
"((np\np)/pp)\((np\np)/pp)",
"((s\np)\(s\np))\((s\np)\(s\np))",
"((np\np)\(np\np))\((np\np)\(np\np))".

With such many possibilities, it can be ambiguous for system to determine the accurate tag for a word. Second, there some words without correct tag in the test set. Since the vocabularies can have several available tags mentioning previously, these lead a problem of insufficient data in train set which does not cover the data in test set. For example,

test data	กระดุม ที่ แขนเสื้อ
reference tag	np/pp pp/np np
result	np pp/np np
test data	เขา แคร์ ความ รู้สึก ของ เรา จริงๆ
reference tag	np (s\np)/np np/(s\np) (s\np) (np\np)/np np (s\np)\(s\np)
result	np (s\np)/ np/(s\np) (s\np) (np\np)/np np (np\np)\(np\np)

All of Thai word "กระดุม" was annotated as "np" in the training set, but in the test set the right one for the correct result is "np/pp" since the word in the test sentence are accompanied with a preposition as well as the word "แคร์" and "จริงๆ". Last, we discovered that the sentence with six words or more has much better chance to return incorrect result because of its variety of combinations.

7. Conclusion and Future Work

We proposed an AutoTCG framework for automatically tagging CG categories to word. The accuracy on tagging sentence is around 82.49% using SAM model with alternative list. In the future, we plan to apply this framework in order to reduce manual tagging. An unknown word tagging should be included as our challenge task. Moreover, the result will be used as a new data to increase the coverage and amount of training set.

8. References

Asanee Kawtrakul, Mukda Suktarachan, Patcharee Varasai and Hutchatai Chanlekha, (2002). "A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling", Coling 2002 post-conference workshops: the 3rd Workshop on Asian Language Resources and International Standardization, Taipei, Taiwan

BEST corpus (2009). Available online from: http://thailang.nectec.or.th/2009/index.php?option=com_content&task=view&id=1&Itemid=34

Bob Carpenter (1992). "Categorial Grammars, Lexical Rules,and the English Predicative", In R. Levine, ed., Formal Grammar: Theory and Implementation. OUP.

Chai Wutiwiwatchai, Sadaoki Furui (2007) Corrigendum to: "Thai speech processing technology: A review" Speech Communication 49 (1) 8-27.

Kazimierz Ajdukiewicz (1935). "Die Syntaktische Konnexitat", Polish Logic.

Mark Steedman (2000). "The Syntactic Process", The MIT Press, Cambridge Mass.

James R. Curran, Stephen Clark and David Vadas (2007). Multi-Tagging for Lexicalized-Grammar Parsing. Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL-06), pp.697-704, Sydney, Australia.

Julia Hockenmaier and Mark Steedman (2002). Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), pp. 1974–1981, Las Palmas.

John Lafferty, Andrew McCallum and Fernando Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of 18th International Conference on Machine Learning, pp. 282–289, Morgan Kaufmann, San Francisco, CA, USA.

Stephen Clark (2002). Supertagging for Combinatory Categorial Grammar. In Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), pp. 19-24, Venice, Italy.

Taneth Ruangrajitpakorn, Wasan na Chai, Prachya Boonkwan, Montika Boriboon, and Thepchai Supnithi (2007). The Design of Lexical Information for Thai to English MT, In Proceeding of 7th International Symposium on Natural Language Processing (SNLP 2007), Pattaya, Thailand.

Taneth Ruangrajitpakorn, Kanokorn Trakultaweekoon, and Thepchai Supnithi (2009). A Syntactic Resource for Thai: CG Treebank. In Proceedings of the 7th Workshop on Asian Language Resources (ALR 09), Singapore.

Thatsanee Charoenporn (1997). Technical Report: TR-NECTEC-1997-001, ISBN: 9747576-98-8, printed and published by National Electronics and Computer Technology Center, Thailand.

Phiradet Bankcharoensap, Peerachet Porkaew, and Thepchai Supnithi (2009). A Statistical Machine Translation Approach to Word Boundary Identification: A project Analogy of Bilingual Translation, InterBest Workshop in the Proceedings of 8th International Symposium on Natural Language Processing (SNLP 2009), Bangkok, Thailand.

Vee Satayamas, and Asanee Kawtrakul (2004). Wide-Coverage Grammar Extraction from Thai Treebank. In Proceedings of Papillon 2004 Workshops on Multilingual Lexical Databases, Grenoble, France.