A Supervised Learning based Chunking in Thai using Categorial Grammar

Thepchai Supnithi, **Peerachet Porkaew, Taneth Ruangrajitpakorn, Kanokorn Trakultaweekool**
Human Language Technology, National Electronics and Computer Technology Center

{thepchai.sup, peerachet.por, taneth.rua, kanokorn.tra}@nectec.or.th

Chanon Onman, Asanee Kawtrakul

Department of Computer Engineering, Kasetsart University and National Electronics and Computer Technology Center

chanon.onman@gmail.com, asanee.kaw@nectec.or.th

Abstract

One of the challenging problems in Thai NLP is to manage a problem on a syntactical analysis of a long sentence. This paper applies conditional random field and categorical grammar to develop a chunking method, which can group words into larger unit. Based on the experiment, we found the impressive results. We gain around 74.17% on sentence level chunking. Furthermore we got a more correct parsed tree based on our technique. Around 50% of tree can be added. Finally, we solved the problem on implicit sentential NP which is one of the difficult Thai language processing. 58.65% of sentential NP is correctly detected.

1 Introduction

Recently, many languages applied chunking, or shallow parsing, using supervised learning approaches. Basili (1999) utilized clause boundary recognition for shallow parsing. Osborne (2000) and McCallum et al. (2000) applied Maximum Entropy tagger for chunking. Lafferty (2001) proposed Conditional Random Fields for sequence labeling. CRF can be recognized as a generative model that is able to reach global optimum while other sequential classifiers focus on making the best local decision. Sha and Pereira (2003) compared CRF to other supervised learning in CoNLL task. They achieved results better than other approaches. Molina et al. (2002) improved the accuracy of HMM-based shallow parser by introducing the specialized HMMs.

In Thai language processing, many researches focus on fundamental level of NLP, such as word segmentation, POS tagging. For example, Kruengkrai et al. (2006) introduced CRF for word segmentation and POS tagging trained over Orchid corpus (Sornlertlamvanich et al., 1998.). However, the number of tagged texts in Orchid is specific on a technical report, which is difficult to be applied to other domains such as news, document, etc. Furthermore, very little researches on other fundamental tools, such as chunking, unknown word detection and parser, have been done. Pengphon et al. (2002) analyzed chunks of noun phrase in Thai for information retrieval task. All researches assume that sentence segmentation has been primarily done in corpus. Since Thai has no explicit sentence boundary, defining a concrete concept of sentence break is extremely difficult.

Most sentence segmentation researches concentrate on "space" and apply to Orchid corpus (Meknavin 1987, Pradit 2002). Because of ambiguities on using space, the accuracy is not impressive when we apply into a real application.

Let consider the following paragraph which is a practical usage from news:

Proceedings of the 8th Workshop on Asian Language Resources, pages 129–136,
Beijing, China, 21-22 August 2010.

"สำหรับการวางกำลังของคนเสื้อแดง ได้มีการวางบังเกอร์โดยรอบพื้นที่ชุมนุม และใช้น้ำมันราด | รวมทั้งมียางรถยนต์ | ขณะการจราจรยังเปิดเป็นปกติ"
lit: "The red shirts have put bunkers around the assembly area and put oil and tires. The traffic is opened normally."

We found that three events are described in this paragraph. We found that both the first and second event do not contain a subject. The third event does not semantically relate to the previous two events. With a literal translation to English, the first and second can be combined into one sentence; however, the third events should be separated.

As we survey in BEST corpus (Kosawat 2009), a ten-million word Thai segmented corpus. It contains twelve genres. The number of word in sentence is varied from one word to 2,633 words and the average word per line is 40.07 words. Considering to a News domain, which is the most practical usage in BEST, we found that the number of words are ranged from one to 415 words, and the average word length in sentence is 53.20. It is obvious that there is a heavy burden load for parser when these long texts are applied.

Example 1:

คน	ขับ	รถแท็กซี่	พบ	กระเป๋าสตางค์
man(n)	drive(v)	taxi(n)	find(v)	wallet(n)

lit1: A man drove a taxi and found a wallet.

lit2: A taxi chauffeur found a wallet.

Example 2:

น่า	จะ	ต้อง	สามารถ	พัฒนา	ประเทศ
should	will	must	can	develop(v)	country(n)

lit: possibly have to develop country.

Figure 1. Examples of compounds in Thai

Two issues are raised in this paper. The first question is "How to separate a long paragraph into a larger unit than word effectively?" We are looking at the possibility of combining words into a larger grain size. It enables the system to understand the complicate structure in Thai as explained in the example. Chunking approach in this paper is closely similar to the work of Sha and Pereira (2003). Second question is "How to analyze the compound noun structure in Thai?"

Thai allows a compound construction for a noun and its structures can be either a sequence of nouns or a combination of nouns and verbs. The second structure is unique since the word order is as same as a word order of a sentence. We call this compound noun structure as a "sentential NP".

Let us exemplify some Thai examples related to compound word and serial construction problem in Figure 1. The example 1 shows a sentence which contains a combination of nouns and verbs. It can be ambiguously represented into two structures. The first alternative is that this sentence shows an evidence of a serial verb construction. The first word serves as a subject of the two following predicates. Another alternative is that the first three word can be formed together as a compound noun and they refer to "a taxi driver" which serve as a subject of the following verb and noun. The second alternative is more commonly used in practical language. However, to set the "N V N" pattern as a noun can be very ambiguous since in the example 1 can be formed a sentential NP from either the first three words or the last three words.

From the Example 2, an auxiliary verb serialization is represented. It is a combination of auxiliary verbs and verb. The word order is shown in Aux Aux Aux Aux V N sequence.

The given examples show complex cases that require chunking to reduce an ambiguity while Thai text is applied into a syntactical analysis such as parsing. Moreover, there is more chance to get a syntactically incorrect result from either rule-based parser or statistical parser with a high amount of word per input.

This paper is organized as follows. Section 2 explains Thai categorial grammar. Section 3

illustrates CRF, which is supervised method applied in this work. Section 4 explains the methodology and experiment framework. Section 5 shows experiments setting and result. Section 6 shows discussion. Conclusion and future work are illustrated in section 7.

2 Linguistic Knowledge

2.1 Categorial Grammar

Categorial grammar (Aka. CG or classical categorial grammar) (Ajdukiewicz, 1935; Bar-Hillel, 1953; Carpenter, 1992; Buszkowski, 1998; Steedman, 2000) is formalism in natural language syntax motivated by the principle of constitutionality and organized according to the syntactic elements. The syntactic elements are categorised in terms of their ability to combine with one another to form larger constituents as functions or according to a function-argument relationship. All syntactic categories in CG are distinguished by a syntactic category identifying them as one of the following two types:

1. Argument: this type is a basic category, such as s (sentence) and np (noun phrase).
2. Functor (or function category): this category type is a combination of argument and operator(s) '/' and '\'. Functor is marked to a complex constituent to assist argument to complete sentence such as s\np (intransitive verb) requires noun phrase from the left side to complete a sentence.

CG captures the same information by associating a functional type or category with all grammatical entities. The notation α/β is a rightward-combining functor over a domain of α into a range of β. The notation α\β is a leftward-combining functor over β into α. α and β are both argument syntactic categories (Hockenmaier and Steedman, 2002; Baldridge and Kruijff, 2003).

The basic concept is to find the core of the combination and replace the grammatical modifier and complement with set of categories based on the same concept with fractions. For example, intransitive verb is needed to combine with a subject to complete a sentence therefore intransitive verb is written as s\np which means

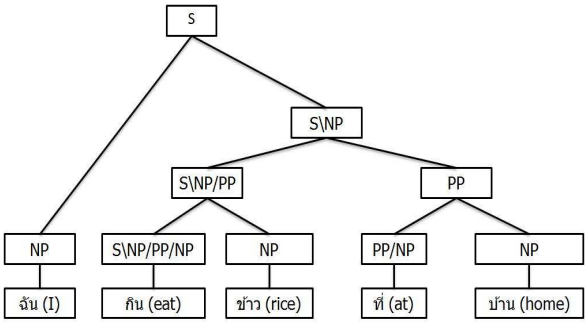

Figure 2 Example of Thai CG-parsed Tree.

it needs a noun phrase from the left side to complete a sentence. If there is a noun phrase exists on the left side, the rule of fraction cancellation is applied as np*s\np = s. With CG, each constituent is annotated with its own syntactic category as its function in text. Currently there are 79 categories in Thai. An example of CG derivation from Thai is shown in Figure 2.

2.2 CG-Set

CG-Set are used as a feature when no CG are tagged to the input. We aim to apply our chunker to a real world application. Therefore, in case that we have only sentence without CG tags, we will use CG-Set instead.

Cat-Set Index	Cat-Set	Member
0	np	คุณสมบัติ
2	s\np/pp,s\np/np,s\np/pp/np,s\np	เก็บ, กรอง
3	(np\np)/(np\np), ((s\np)\(s\np))/spnum, np, (np\np)\num,np\num, (np\np)/spnum, ((s\np)\(s\np))\num	วงจร, สัญญาณ
62	(s\np)\(s\np),s\s	มั้ย, มั้ง, ล่ะ
134	np/(s\np), np/((s\np)/np)	การ, ความ

Table 1 Example of CG-Set

The concept of CG-Set is to group words that their all possible CGs are equivalent to the other. Therefore every word will be assigned to only one CG-Set. By using CG-Set we use the lookup table for tagging the input. Table 1 shows examples of CG-set. Currently, there are 183 CG set.

3 Conditional Random Field (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. It is a supervised framework for labeling a sequence data such as POS tagging and chunking. Let X is a random variable of observed input sequence, such as sequence of words, and Y is a random variable of label sequence corresponding to X , such as sequence of POS or CG. The most probable label sequence ($\hat{y}$) can be obtain by

$$\hat{y} = \arg\max p(y \mid x)$$

Where $x = x_1, x_2, ..., x_n$ and $y = y_1, y_2, ..., y_n$ $p(y \mid x)$ is the conditional probability distribution of a label sequence given by an input sequence. CRF defines $p(y \mid x)$ as

$$P(y \mid x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{n} F(y, x, i) \right)$$

where $Z = \sum_y \exp\left(\sum_{i=1}^{n} F(y, x, i)\right)$ is a normalization factor over all state sequences. $F(y, x, i)$ is the global feature vector of CRF for sequence x and y at position i . $F(y, x, i)$ can be calculated by using summation of local features.

$$F(y, x, i) = \sum_i \lambda_i f_i(y_{i-1}, y_i, t) + \sum_j \lambda_j g_j(x, y, t)$$

Each local feature consists of transition feature function $f_i(y_{i-1}, y_i, t)$ and per-state feature function $g_j(x, y, t)$. Where λ_i and λ_j are weight vectors of transition feature function and per-state feature function respectively.

The parameter of CRF can be calculated by maximizing the likelihood function on the training data. Viterbi algorithm is normally applied for searching the most suitable output.

4 Methodology

Figure 3 shows the methodology of our experiments. To prepare the training set, we start with our corpus annotated with CG tag. Then, each sentence in the corpus was parsed by

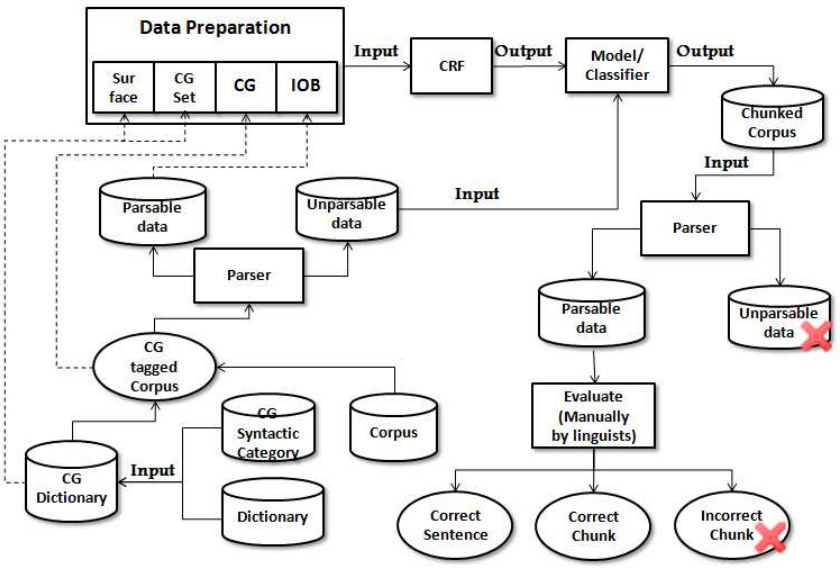

Figure 3 Experimental Framework

our Thai CG parser, developed by GLR tech-technique. However, not all sentences can be parsed successfully due to the complexity of the sentence. We kept parsable sentences and unparsable sentences separately. The parsable sentences were selected to be the training set.

There are four features – surface, CG, CG-set and chunk marker – in our experiments. CRF is applied using 5-fold cross validation over combination of these features. Accuracy in term of averaged precision and recall are reported.

We select the best model from the experiment to implement the chunker. To investigate performance of the chunker, we feed the unparsable sentences to the chunker and evaluate them manually.

After that, the sentences which are correctly chunked will be sent to our Thai CG parser. We calculate the number of successfully-parsed sentences and the number of correct chunks.

5 Experiment Settings and Results

5.1 Experiment on chunking

5.1.1 Experiment setting

To develop chunker, we apply CG Dictionary and CG tagged corpus as input. Four features are provided to CRF. Surface is a word surface. CG is a categorial grammar of the word. CG-set is a combination of CG of the word. IOB represents a method to mark chunk in a sentence. "I" means "inner" which represents the word within the chunk. "O" means "outside" which represents the word outside the chunk. "B" means "boundary" which represents the word as a boundary position. It accompanied with five chunk types. "NP" stands for noun phrase, "VP" stands for verb phrase, "PP" stands for preposition phrase, "ADVP" stands for adverb phrase and S-BAR stands for complementizer that link two phrases.

Surface and CG-set are developed from CG dictionary. CG is retrieved from CG tagged corpus. IOB is developed by parsing tree. We apply Thai CG parser to obtain the parsed tree. Figure 4 shows an example of our prepared data. We provide 4,201 sentences as a training data in CRF to obtain a chunked model. In this experiment, we use 5-fold cross validation to evaluation the model in term of F-measure.

surface	cg_set	cg	chunk_label
ใน	74	s/s/np	B-ADVP
วัน	3	np	I-ADVP
ที่	180	(np\np)/(s\np)	I-ADVP
ไม่	54	(s\np)/(s\np)	I-ADVP
หนาว	7	s\np	I-ADVP
หรือ	130	((s/s)\(s/s))/(s/s)	I-ADVP
ใน	74	s/s/np	I-ADVP
ฤดูร้อน	0	np	I-ADVP
เขา	0	np	B-NP
สวม	8	s\np/np	B-VP
เสื้อ	0	np	B-NP
มา	148	(s\np)/(s\np)	B-VP
เข้าเฝ้า	2	s\np	I-VP

Figure 4 An example of prepared data

model	surface	cg-set	cg	NP			VP			PP			OVERALL			
				precision	recall	FB1	precision	recall	FB1	precision	recall	FB1	accuracy	precision	recall	FB1
1	Yes	No	No	60.32	55.83	57.99	77.06	67.69	72.07	93.42	83.64	88.26	83.28	68.57	62.08	65.16
2	No	Yes	No	65.74	62.67	64.17	79.42	76.24	77.80	92.94	89.24	91.05	86.02	72.68	69.47	71.04
3	Yes	Yes	No	66.02	63.34	64.65	80.21	77.46	78.81	94.19	89.54	91.80	86.42	73.16	70.3	71.7
4	No	No	Yes	81.84	80.46	81.15	89.56	92.39	90.96	99.56	99.56	99.56	93.24	86.09	86.31	86.2
5	Yes	No	Yes	76.19	75.13	75.65	87.30	89.12	88.20	99.56	99.41	99.48	91.38	82.14	82.13	82.14
6	No	Yes	Yes	76.65	75.52	76.08	87.38	89.45	88.41	99.56	99.48	99.52	91.45	82.44	82.47	82.46
7	Yes	Yes	Yes	76.17	75.09	75.63	87.41	89.08	88.24	99.56	99.34	99.45	91.34	82.16	82.09	82.12

Table 2 Chunking accuracy of each chunk

model	surface	CG-set	CG	average	
				word	sent
1	Yes	No	No	83.28	41.37
2	No	Yes	No	86.02	49.95
3	Yes	Yes	No	86.42	50.12
4	No	No	Yes	93.24	74.17
5	Yes	No	Yes	91.38	66.74
6	No	Yes	Yes	91.45	67.41
7	Yes	Yes	Yes	91.34	66.68

Table 3 Chunking accuracy based on word and sentence.

5.1.2 Experiment result

From Table 2, considering on chunk based level, we found that CG gives the best result among surface, CG-set, CG and their combination. The average on three types in terms of F-measure is 86.20. When we analyze information in detail, we found that NP, VP and PP show the same results. Using CG shows the F-measure for each of them, 81.15, 90.96 and 99.56 respectively.

From Table 3, considering in both word level and sentence level, we got the similar results, CG gives the best results. F-measure is 93.24 in word level and 74.17 in sentence level. This shows the evidence that CG plays an important role to improve the accuracy on chunking.

5.2 Experiment on parsing

5.2.1 Experiment setting

We investigate the improvement of parsing considering unparsable sentences. There are 14,885 unparsable sentences from our CG parser. These sentences are inputted in chunked model to obtain a chunked corpus. We manually evaluate the results by linguist. Linguists evaluate the chunked output in three types. 0 means incorrect chunk. 1 means correct chunk and 2 represents a special case for Thai NP, a sentential NP.

5.2.2 Experiment result

From the experiment, we got an impressive result. We found that 11,698 sentences (78.59%) are changed from unparsable to parsable sentence. Only 3,187 (21.41%) are unparsable. We manually evaluate the parsable sentence by randomly select 7,369 sentences. Linguists found 3,689 correct sentences (50.06%). In addition, we investigate the number of parsable chunk calculated from the parsable result and found 37,743 correct chunks from 47,718 chunks (78.47%). We also classified chunk into three types NN VP and PP and gain the accuracy in each type 79.14% ,74.66% and 92.57% respectively.

6 Discussion

6.1 Error analysis

From the experiment results, we found the following errors.

6.1.1 Chunking Type missing

Some chunk missing types are found in experiment results. For example, [PP บันทึก (record)][NP ตัวอักษรได้ประมาณ (character about)]. [PP

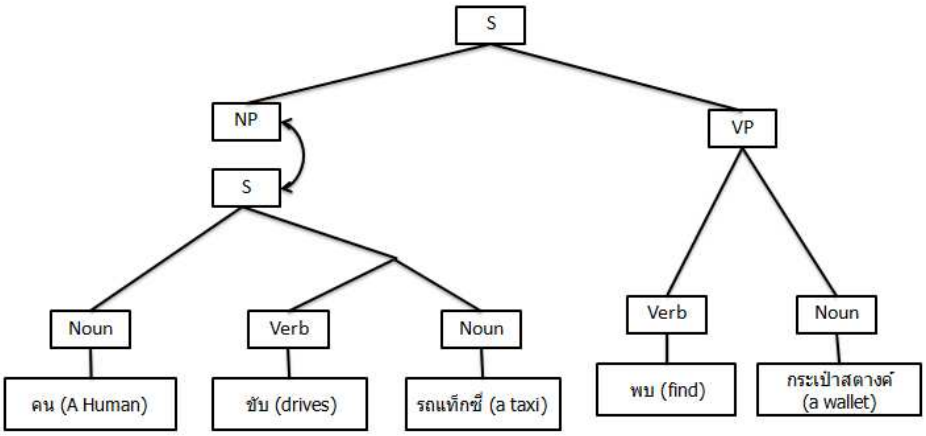

Figure 4 An Example of sentential NP

บันทึก (record)] should be defined as VP instead of PP.

6.1.2 Over-grouping

In the sentence "[VP ใช้ (Using)][NP (medicine)][VP รักษา (treat)][NP โรคแต่ละครั้งต้องเป็นไป (each disease have to)][PP ตาม (follow)][NP คำแนะนำของแพทย์ (doctor's instruction)] ", we found that "NP โรคแต่ละครั้งต้องเป็นไป (each disease have to) " has over-grouping. IT is necessary to breakdown to NP โรคแต่ละครั้ง(each disease) and VP ต้องเป็นไป(have to). The reason of this error is due to allow the sentential structure NP VP NP, and then NP and VP are combined.

6.1.3 Sentential NP

We investigated the number of sentential NP. If the number of chunk equal to 1, sentence should not be recognized as NP. Other cases are defined as NP. We found that 929 from 1,584 sentences (58.65 % of sentences) are correct sentential NP. This evidence shows the impressive results to solve implicit NP in Thai. Figure 4 shows an example of sentential NP.

6.1.4 CG-set

Since CG-set is another representation of word and can only detect from CG dictionary. It is very easy to develop a tag sequence using CG-set. We found that CG-set is more powerful than surface. It might be another alternative for less language resource situation.

6.2 The Effect of Linguistic Knowledge on chunking

Since CG is formalism in natural language syntax motivated by the principle of constitutionality and organised according to the syntactic elements, we would like to find out whether linguistic knowledge effects to the model. We grouped 89 categorial grammars into 17 groups, called CG-17.

It is categorized into Noun, Prep, Noun Modifier, Number modifier for noun, Number modifier for verb, Number, Clause Marker, Verb with no argument, Verb with 1 argument, Verb with 2 or more arguments, Prefix noun, Prefix predicate, Prefix predicate modifier, Noun linker, Predicate Modification, Predicate linker, and Sentence Modifier.

We found that F-measure is slightly improved from 74.17% to 75.06%. This shows the evidence that if we carefully categorized data based on linguistics viewpoint, it may improve more accuracy.

7 Conclusions and Future Work

In this paper, we stated Thai language problems on the long sentence pattern and find the novel method to chunk sentence into smaller unit, which larger than word. We concluded that using CRF accompanied with categorical grammar show the impressive results. The accuracy of chunking in sentence level is 74.17%. We are possible to collect 50% more on correct tree. This technique enables us to solve the implicit sentential NP problem. With our technique, we found 58% of implicit sentential NP. In the future work, there are several issues to be improved. First, we have to trade-off between over-grouping problem and implicit sentential problem. Second, we plan to consider ADVP, SBAR, which has a very small size of data. It is not adequate to train for a good result. Finally, we plan to apply more linguistics knowledge to assist more accuracy.

References

Abney S., and Tenny C., editors, 1991. *Parsing by chunks*, *Priciple-based Parsing*. Kluwer Academic Publishers.

Awasthi P., Rao D., Ravindram B., 2006. *Part of Speech Tagging and Chunking with HMM and CRF*, Proceeding of the NLPAI Machine Learning Competition.

Basili R., Pazienza T., and Massio F., 1999. *Lexicalizing a shallow parser,* Proceedings of

Traitement Automatique du Langage Naturel 1999. Corgese, Corsica.

Charoenporn Thatsanee, Sornlertlamvanich Virach, and Isahara Hitoshi. 1997. Building A Large Thai Text Corpus - Part-Of-Speech Tagged Corpus: ORCHID. Proceedings of Natural Language Processing Pacific Rim Symposium.

Kosawat Krit, Boriboon Monthika, Chootrakool Patcharika, Chotimongkol Ananlada, Klaithin Supon, Kongyoung Sarawoot, Kriengket Kanyanut, Phaholphinyo Sitthaa, Purodakananda Sumonmas,Thanakulwarapas Tipraporn, and Wutiwiwatchai Chai. 2009. *BEST 2009: Thai Word Segmentation Software Contest*. The Eigth International Symposium on Natural Language Processing : 83-88.

Kruengkrai C., Sornlertlumvanich V., Isahara H, 2006. *A Conditional Random Field Framework for Thai Morphological Analysis*, Proceedings of 5th International Conference on Language Resources and Evaluation (LREC-2006).

Kudo T., and Matsumoto Y., 2001. *Chunking with support vector machines*, Proceeding of NAACL.

Lafferty J., McCallum A., and Pereira F., 2001. *Conditional Random Fields : Probabilistic models for segmenting and labeling sequence data*. In Proceeding of ICML-01, 282-289.

McCallum A., Freitag D., and Pereira F. 2000. *Maximum entropy markov model for information extraction and segmentation.* Proceedings of ICML.

Molina A., and Pla F., 2002. *Shallow Parsing using Specialized HMMs*, Journal of Machine Learning Research 2,595-613

Nguyen L. Minh, Nguyen H. Thao, and Nguyen P., Thai. 2009. *An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models*, Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009,9-16

Osborne M. 2000. *Shallow Parsing as Part-of-Speech Tagging.* Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal.

Pengphon N., Kawtrakul A., Suktarachan M., 2002. *Word Formation Approach to Noun Phrase Analysis for Thai*, Proceedings of SNLP2002.

Sha F. and Pereira F., 2003. *Shallow Parsing with Conditional Random Fields*, Proceeding of HLT-NAACL.