Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information

Hutchatai Chanlekha **Asanee Kawtrakul**

The Specialty Research Unit of Natural Language Processing and
Intelligent Information System Technology
Department of Computer Engineering, Kasetsart University, Bangkok
{aim,ak}@vivaldi.cpe.ku.ac.th

Abstract

The role of Named entity (NE) extraction is very important in many NLP tasks, such as information extraction, etc. In Thai, the problems of NE extraction are much more difficult due to the characteristics of Thai language, that are lack of orthographical information to signal NEs, and no boundary indicator between words. In this paper, we present Thai NE extraction system by using Maximum Entropy model, with heuristic information and dictionary. Our system is divided into three steps. The first step is to identify the boundary of candidate NE that composes of many words by using heuristic rules, dictionary and statistic of word co-occurrence. The second step is NE extraction by using Maximum Entropy model. The final step is to extract the undiscovered NE by matching the extracted NEs against the rest of document. On Thai political news test data, the evaluation of the system shows that the F-measures of person, location, and organization names are 90.44%, 82.16% and 89.87% respectively.

1 Introduction

Named entity (NE) extraction task was first introduced in 1995 as a subtask of Message Understanding Conference (MUC). In MUC framework, NEs were defined as entity names (person, organization, and location name), temporal expressions (date and time), and numerical expression (monetary values and percentages).

Much of initial works in NE extraction usually based on rule-based approach (Appelt, et.al., 1993; Krupka and Hausman, 1998). Recent researches on NE extraction have focused on machine learning approach, such as, maximum entropy model (Borthwick, et.al., 1998; Chieu and Ng, 2002), Hidden Markov Model (Bikel, et.al., 1997), Support Vector Machine (Isozaki and Kazawa, 2002), decision tree (Sekine, 1998), etc.

NE extraction also has been applied to Asian language, such as Japanese, Chinese, etc. These systems usually use information from character, such as character type (Sassano and Utsuro, 2000; Sekine, 1998; Borthwick, 1999), or use character list to indicate NE (Sun, et.al., 2002); using sophisticate preprocessing step, include POS tagging, parsing (Zhang and Zhou, 2000); or use pattern or structure of NE (Ye, et. al., 2002), etc. However, these systems can not be directly applied to Thai NE extraction because of Thai language characteristics. In Thai, the NE extraction task is much more difficult due to the characteristics of Thai language. The difficulties come from:

- Thai does not have orthographical information, such as, uppercase character as used in English, or character type, such as Kanji, Katakana, as used in Japanese, to signal NE. So it is difficult to distinguish between NE and common word.
- Thai does not use space or special character as a delimiter between words. So the performance of the word segmentation system will definitely have the impact on NE extraction.

- A large portion of Thai NE doesn't have certain structure, and can construct from any word. So it is difficult to build a model or rules that extract NE by considering on an internal structure of NE (Wu, et.al., 2003).

In additional, there's also a problem from Thai writing style that usually refer to the previously introduced NE by using different form(s), and/or without clue word. This could cause the ambiguity between NE and common word.

The researches on Thai NE extraction are still at early stage. Charoenpornsawat, et. al. (1998) employed Winnow algorithm for Thai NE identification. The features that they use are context words and collocations as well as part of speech (POS) tag, and use heuristics information from dictionary and POS to generate NE candidates and solving NE boundary problem. The accuracy of their system is 92.17%. However, the corpus used in their experiment was manually segmented into words, and the POS was tagged by linguists. This required a lot of time and linguists skill.

In this paper, we will focus on extracting person, location, and organization name, since NEs in these categories are more ambiguity and usually cause the problems to document processing than temporal and numeric expressions. To extract Thai NE, we proposed the approach by applying Maximum Entropy model and incorporate knowledge, which are rules and dictionary to NE extraction system. Our maximum entropy model doesn't use POS information as a feature. Since we want to avoid the impact from correctness of POS tagging and to reduce time and effort in building training corpus. The system is divided into 3 steps:

- NE boundary identification: identifying a candidate set of multi-word NEs by using heuristics from rules, dictionary, and statistical of word co-occurrence.
- NE extraction: extracting NE that composes of one word (single-word NE), as well as verifying the candidate NEs from previous step by using Maximum Entropy Model. All extracted NEs will be kept as a list for using in the next step.
- Ambiguous NE discovering: discovering the rest of NE that appear ambiguously, or has a deviated form from the extracted NE by matching NE in the extracted list against the rest of document.

This paper will be organized as follows. Section 2 describes the characteristics of Thai NE and the problem of Thai NE extraction. Thai NE extraction approach is described in Section 3. Section 4 is the evaluation and Section 5 is conclusion.

2 Problems in Thai NE extraction

In this section, we will present the definition of Thai word, characteristics of Thai NE, and the problems in Thai NE extraction.

2.1 Definition of Thai word

In Thai, word is defined as a sequence of characters that are consonant (c), implicit or explicit vowel (v), and/or tonal mark (t), representing a linguistic token. For example: "เก้าอี้" (chair; cvtcvt), "มด" (ant; cc), "มา" (come; cv), etc.

"Word" in this work are:

1) Morpheme (M): the smallest unit of word that has meaning, such as, "วิ่ง" (cvtc: run), etc.

2) Compound word (CW): the word that composes of two or more morphemes, which the meaning is changed from the original meaning of constituent morphemes. For example: "ลูกเสือ" (scout: $[cvc]_{M1}$ $[cv]_{M2}$) composes of 2 morphemes: "ลูก" (child: M1) and "เสือ" (tiger: M2)

3) Proper name (PN): proper name can be divided into two types, that are

- single-word NE, which composes of one word, either morpheme or compound word
- multi-word NE, which composes of sequence of morphemes or compound words.

The examples of NE are shown in table 1

Table1 Examples of single-word & multi-word NE

Type of NE	Example	Word unit
Single-word NE	[พัชรี]$_{M}$	1 word, composes of 1 M
	[สมชาย]$_{CW}$	1word, composes of 1 CW
Multi-word NE	[การ]$_{M}$[ไฟฟ้า]$_{CW}$ [นครหลวง]$_{CW}$	3 word, composes of 1 M and 2 CW

2.2 Characteristics of Thai NE

Thai NEs are formed by the combination of known words and unknown strings (Charoenpornsawat, et.al., 1998). Characteristics of Thai NE are:

- Thai NEs do not use special character, or any other orthographical information, such as uppercase character in English, to differentiate between NE and other word class. Furthermore, foreign NEs will be transliterated without using any special character, such as the use of Katakana character in Japanese, to signal the transliterated NE.

- A large portion of Thai NE has no specific construction rule, i.e. Thai NE doesn't have certain structure and can be constructed from any words.

2.3 Problems in Thai NE extraction

In order to extract Thai NE, two major problems must be solved, which are NE identification, and NE classification.

2.3.1 Problems in Thai NE Identification

- **Ambiguity between single-word NE and common word**

  The lack of orthographical information to signal position of NE makes the NE extraction more problematic when single-word NE has the same surface form as common word, for example:

สุภาพ	มี	กิริยามารยาท	สุภาพ	เรียบร้อย
Suparb	has	manner	polite	neat
[person NE]			[verb]	

  This characteristic makes Thai NE extraction have to rely heavily on contextual information.

- **Ambiguity between multi-word NE and common noun phrase**

  The lack of orthographical information also causes the problem in multi-word NE identification, especially when multi-word NE has the same structure as common noun phrase (NP), etc. For example:

	[ศูนย์	ส่งเสริม	อาชีพ	และ	พัฒนา	คนพิการ]$_{NE}$	
	[center	support	career	and	develop	the disabled]$_{NE}$	
เป็น	[ศูนย์	ฝึก	อาชีพ	และ	ช่วยเหลือ	ผู้พิการ]$_{NP}$	ให้
is	[center	train	career	and	help	the disabled]$_{NP}$	to
มี	ความรู้ความสามารถ						
have	ability						

Goodwill Industries of Thailand is a center that provides vocational training and assistance for the disabled.

- **Ambiguity in multi-word NE boundary identification**

  The same characteristic that causes an ambiguity between NE and common word or noun phrase also causes the problem in multi-word NE boundary identification, especially when NEs are formed by the combination of known words. From the following example, we can see that Thai NE does not have information from character, such as uppercase character in English, to help in boundary identification task.

NE: "กรมอุทยานแห่งชาติ สัตว์ป่า และพันธุ์พืช"
(National Park, Wildlife, and Plant Conservation Department)

กรม	อุทยาน	แห่งชาติ	สัตว์ป่า	และ	พันธุ์พืช
Department	Park	National	wildlife	and	Plant

could be detected as:

[กรมอุทยานแห่งชาติ]$_{NE}$ สัตว์ป่า และพันธุ์พืช
[กรมอุทยานแห่งชาติ สัตว์ป่า]$_{NE}$ และพันธุ์พืช
[กรมอุทยานแห่งชาติ สัตว์ป่า และพันธุ์พืช] $_{NE}$

2.3.2 Problems in Thai NE Classification

- **Weak context**

  Problems in NE classification arise when contexts around NE are not strong clue to help in classification task, or the useful context is not in the position that can be captured by the extraction model. Furthermore, NE in different category can appear in similar context, for example:

การบินไทย เปิดเผยว่า... VS. อภิสิทธิ์ เปิดเผยว่า...
[org. NE] [announces]... [per. NE] [announces]...

- **Category ambiguity**

  NE in different category can have the same surface form, for example "พิจิตร" can be person name, or location name, depend on context.

To solve these problems, we developed Thai NE extraction system by using Maximum Entropy model together with rules and dictionary to extract Thai NE from word segmented documents.

3 Thai NE extraction System

3.1 System Overview

The system is divided into 2 modules: training module by using Maximum Entropy model, and NE extraction module.

- Training module

  Steps of system training process consists of:

  1) Corpus preparation, which includes segmentation and manually NE annotation.

  2) Extracting features of each token in training corpus. The features that are used in this work are, for example, word feature, dictionary feature, etc.

  3) System training by using Maximum Entropy Model. The results of this step are weights of each feature function.

- NE Extraction module

  The process of NE extraction composes of 3 steps (see figure1) that are:

1) Pre-processing step: For identifying the boundary and position of multi-word NE by using rules, dictionary, and statistics of word co-occurrence.
2) NE extraction step: For extracting single-word NE and verifying the candidate NE from pre-processing step by using weights of each feature that is acquired from training process.
3) Post-processing step: For extracting some of the remaining NEs by matching the extracted NE from previous step against the rest of document.

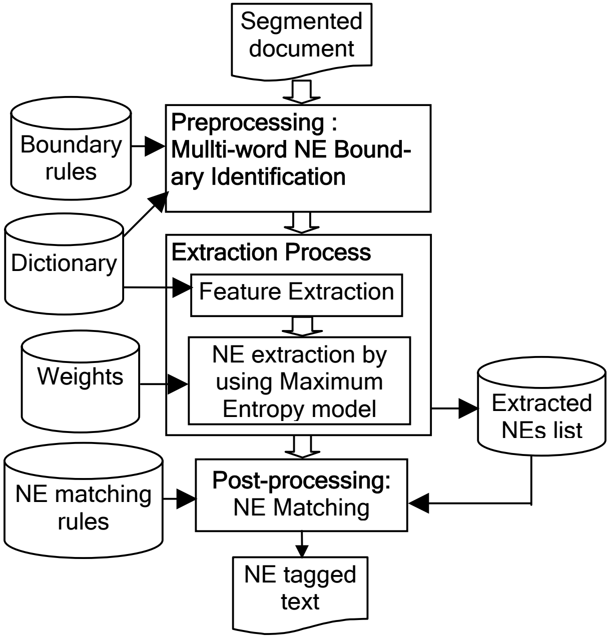

Figure 1 Thai Named Entity Extraction Process

3.2 Pre-processing step: Multi-word NE boundary identification

For most of the maximum entropy-based system, the problem of boundary identification is viewed as the problem of classification (Borthwick, 1998), which classifies each token into the tags that indicate the position of the token in NE. These tags are, the start of NE, a part of NE, the end of NE, and the NE that composes of one word. However, this consideration does not suit for Thai since Thai doesn’t have orthographical information to signal NE. Furthermore, Thai multi-word NE, especially organization name, usually has the same structure as common noun phrase. From this fact, we propose the approach for NE boundary identification task by using heuristics from rules, dictionary, and local statistical of word co-occurrence.

Multi-word NE boundary identification consists of identifying the starting position of multi-word NE and determining the ending position, i.e., identifying the boundary of that multi-word NE.

3.2.1 Identifying the Starting Position of Multi-word NE

From corpus observation in domain politic news (corpus size 100,000 words), we found statistic of multi-word NE appearance as shown in table2.

Table 2 Statistic of multi-word NE appearance

Multi-word NE Appearance	Per. (%)	Loc. (%)	Org. (%)
With clue front or clue back	41.44	1.29	19.96
Sequence of unk. words /start with unk. Word	12.48	3.80	4.39
In NE dictionary	4.72	0.94	3.55
No clue but appear with clue in another position	0.98	0.13	0.25
Without clue	1.58	0.73	0.75
All percentage of multi-word NE in corpus	61.2	6.89	28.90

Remark that these statistics are computed in the order listed, so that in the case of joint situation, such as "multi-word NE appear with clue word" and "multi-word NE is in dictionary", the statistic will count for the former situation.

From table 2, the number of multi-word NE that appeared without any clue is very small. In this work, we will ignore this small portion and use just simple heuristics to generate the candidates starting position of multi-word NE. These heuristics are:

1) Candidate NE appears after NE starting clue.
2) Candidate NE appears before NE ending clue.
3) Candidate NE is in NE dictionary.
4) Candidate NE is/start with unknown word.

In this work, we consider NE starting clue as:

PER clue ≡ {title}
ORG clue ≡ {company designator, common
noun semantic organization}
LOC clue ≡ {common noun semantic location}

3.2.2 Identifying the Boundary of candidates for Multi-word NE

There are 3 modules in this step, which will process in sequence. These modules are:

1) Rule-based boundary identification

Many multi-word NEs in Thai have pattern that can be easily captured by rules. In this work, we try to construct rules for each NE category, which have high precision and least depending on domain. Examples of rules for organization are shown below:

(1) If : <clue_org start1> W^+ <clue_org end1>
Then : group W^+ as 1 NE with category ORG
where; clue_org start1 ≡ {“บริษัท”(Company), etc.
clue_org end1 ≡ {“จำกัด”(ltd.), etc.}

(2) If : <clue_org start2> W^+ <org_end pattern>
Then : group [W^+<org_end pattern>] as 1 org. NE
where; clue_org start2 ≡ {“สภา” (council), etc.
org_end pattern ≡ {“แห่งชาติ” (national),
”แห่งประเทศW_c” (of W_c), etc.}
$W_c \in$ Country name

Since these rules are high-precision rules, the sequence of tokens that conform to these rules will be extracted as NE without using Maximum Entropy as verification module.

2) Dictionary-based boundary identification

In our work, the well-known NEs will be kept in NE dictionary for the usefulness in reducing the complexity and difficulty in boundary identification task. This module use NE dictionary to identify boundary of candidate NE by comparing each NE in dictionary to the sequence of token in the document. However, there is a problem when multi-word NE composes of known words, for example: “สองพี่น้อง” which can be both location name and noun phrase means two brothers or sisters. So Maximum Entropy model is needed as verification system for this module.

3) Statistical-based boundary identification

Boundary of multi-word NE is determined by considering statistic of word co-occurrence. In this module, we classify the problem into 2 cases: NE appears only once in document, and NE appears more than once in document.

- NE appear only once in document

At any possible cues of starting (or ending) position of multi-word NE, its predicted boundary will be repeatedly extended along with the calculation of its boundary probability. The calculation will be done by using the formula below:

$$P_{boundary}(w_0) = P_{in\ NE}(w_{-1}) * P_{NE\ end}(w_0) * P_{after\ NE}(w_1)$$

where, $P_{in\ NE}(w)$ is probability from training corpus that w will be NE constituent.

$P_{NE\ end}(w)$ is probability from training corpus that w will be NE ending word.

P_{after} NE(w) is probability from training corpus that w will appear after NE.

NE boundary extension will be terminated if the succeeded word is a member of predefined stop words or extended NE boundary is more than ten words. NE boundary whose probability is the highest will be selected.

- NE appear more than once in document

Our NE identification will be processed as follows. First, the possible cues of starting (or ending) points of multiple-word NE are recognized. Afterward, for each cue point, the system will expand the NE boundary to the longest-matched sequences that appear at least twice within the document. Herewith, these sequences are considered NE candidates. NE boundary extension will be terminated if at least one of the following criteria is satisfied.

1) The succeeded word of each matched position is different. Or the succeeded word is a member of predefined stop words.

2) The succeeded word has a probability not to occupy in NE.

3.3 NE extraction by Maximum Entropy Model

Main role of Maximum Entropy model in this step is to identify the position and classify NE. In this work, we will classify each token into one of the set of NE tag categories. Introduction to Maximum Entropy and the details of each kind of features are explained below.

Maximum Entropy

Given a test corpus and a set of tags, which define NE categories, the problem of NE extraction can be reduced to the problem of assigning one of tags in tags set to each token (Borthwick, et. al., 1998). Maximum entropy model allows the computation of p(f|h) for f from the space of possible futures (possible tags), and for h from the space of possible histories, i.e. all information that enables system to make decision among the space of f.

The computation of p(f|h) in maximum entropy depends on a set of binary-valued “feature” functions, g_i(h,f). Given a set of feature functions and training data, the maximum entropy estimation process produces a model by associated a parameter α_i with every feature g_i(h,f). This allows us to compute the conditional probability as follow:

$$P(f \mid h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\alpha(h)} \quad (1)$$

$$Z_\alpha(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)} \quad (2)$$

Features

- Word Features

This type of feature considers special characteristic of each token. Word features in this work are designed mainly for extracting person, location and organization name. These features are shown in Table 3.

Table 3 Word features

Word feature	Example
Contain ‘ฯ’	การไฟฟ้าฯ
Contain ‘&’	เอส & พี
Contain ‘.’	ททท.
Contain ‘-‘	ถนนตาก-เถิง
Contain blank	เทเลคอม เอเชีย
Contain parentheses	การบินไทย จำกัด (มหาชน)
Contain any digit	ถนนพระราม 3
Contain English character	EU

- Lexical Features

To create lexical feature, the token in range ±2 words (where the current token is denoted as w0) are compared with the vocabulary and their vocabulary indices are recorded as features.

- Dictionary Features

NEs that are kept in our dictionary are well-known NE. In Thai, well-known NEs are usually stated without any clue word. So dictionary is an important source of knowledge for extraction system, that can help when NE appears in an ambiguous context. Dictionaries that are used in this work are shown in Table 4.

Table 4 Dictionary

Dictionary		Number of word
Common word		15142
Person name	First name	14000
	Last name	14000
Location name		7000
Organization name		1084
Person starting clue word		300
Location starting clue word		20
Org. starting clue word		33
Org. ending clue word		27

- Blank Features

In Thai, there is no explicit boundary indicator between words or sentences. However, blank is usually used as a separator between sentences, phrases, especially between consecutive NEs. For example:

ประเทศในกลุ่มอาเซียน ได้แก่ **ไทย ลาว พม่า** เป็นต้น

Asian countries, such as, Thai, Lao, Myanmar, etc.

From this fact, we include behavior of blank-using as a feature to this model. In this work, we consider two types of blank features, which are: “*blank exists in front of W_0*” and “*blank exists after W_0*”

3.4 Post-processing: NE matching

The main role of this step is to discover the remaining NEs in the document by matching the list of extracted NEs against each token in the document. This step also uses rules for matching task. Example of rule is:

```
If (w’= org. name && Part(w’,x) && Start(w’,x))
   then x = org. name
where:  Part(w’,x) is true when w’ composes of or
        equal to x
        Start(w’,x) is true when w’ start with or
        equal to x

For example :
(1) “เซ็นทรัล ประกาศ…” (Central announces…)
suppose we have :
“เซ็นทรัล พัฒนา” (Central Pattana) ∈ extracted list.
then “เซ็นทรัล” (Central) in (1) will be org. name
```

Incidentally, common words can be incorrectly extracted by our model. In this situation, instead of improvement, the NE matching will degrade the system performance. To prevent this, we establish a notion of suspected NE, which is a token whose appearance is more than three times whereas it can be extracted as NE only once. These suspected NEs will be excluded from NE matching task.

4 Experimental result

In our experiment, we use corpus domain political news, with size 110,000 words for training task, and 25,000 words for testing. The performances of NE extraction system that considers context in range ±1 words and ±2 words are shown in table 5 and 6, respectively.

Table 5 NE extraction performance; consider context in range ±1W.

NE category	P (%)	R (%)	F (%)
Person	88.33	92.66	90.44
Organization	92.08	87.76	89.87
Location	80.15	84.28	82.16
Total	87.60	87.80	87.70

Table 6 NE extraction performance; consider context in range ±2W.

NE category	P (%)	R (%)	F (%)
Person	84.56	84.27	84.43
Organization	76.84	78.70	77.76

Location	82.83	77.06	79.84
Total	80.14	79.42	79.78

The results show that, for every category, F-score, as well as precision and recall, of case ±1W is higher than case ±2W. However, for location, precision of case ±1W is lower than case ±2W, since location name usually appear in more ambiguous context than other categories. By incorporate more contextual information, it will improve the precision of the system.

The results also indicate that organization name is the most intervened by the data sparseness problem. By considering the organization F-scores, f-score of case ±2W is 77.76%, which is significantly improved to 89.87% in the model that considers less contextual information, i.e., less feature functions. The person name has the highest F-score, since person name in political news testing data usually appear with clue word, which enable the system to correctly extract NE in this category.

5 Conclusion

In this work, we propose Thai NE extraction approach by using Maximum Entropy model with knowledge from dictionary and rules and solve the boundary problem of multi-word NE by using heuristic from rules, dictionary and statistic of word co-occurrence. In our work, we extract Thai NE from segmentation document without using information from POS tag, since we want to avoid the impact from correctness of POS tagging and to reduce time and effort in building training corpus.

While our results have been quite acceptable, there is still much that can be done to improve performance of the systems. We would like to incorporate the following into the current system:

- More powerful NE boundary detection approach.
- Longer-distance information, to find names that are not captured by our model.

References

D. Appelt, J. Hobbs, D. Israel, and M. Tyson. FASTUS: A finite-state processor for information extraction from real-world text. In Proc. of IJCAI-93, 1993.

D. Bikel, S. Miller, R. Schwartz, and R.Weischedel, “Nymble: a high-performance learning name-finder”. In Proc. of the 5th Conference on Applied Natural Language Processing, 1997.

A. Borthwick, “A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese”, In Proc. of the IREX Workshop, Tokyo, Japan, 1999.

A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, “NYU: Description of the MENE Named Entity System as used in MUC-7”. In Proc. of the 7th Message Understanding Conference, Fairfax, USA, 1998.

P. Charoenpornsawat, B. Kijsirikul, and S. Meknavin, “Feature-based Proper Name Identification in Thai”, In Proc. of National Computer Science and Engineering Conference: NCSEC’98, Thailand, 1998.

H. Isozaki, H. Kazawa, “Efficient Support Vector Classifiers for Named Entity Recognition”, In Proc. of COLING-2002, Taipei, Taiwan, 2002.

G. R. Krupka, and K. Hausman, “IsoQuest: Description of the Ne-tOwl(tm) extractor system as used in MUC-7”. In Proc. of the 7th Message Understanding Conference, Fairfax, USA, 1998.

H. Leong C. and H. T. Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information". In Proc. of COLING-2002. Taipei, Taiwan, 2002.

M. Sassano and T. Utsuro. "Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition", In Proc. of COLING 2000, 2000.

S. Sekine. “NYU: Description of the Japanese NE System Used for MET-2”. In Proc. of 7th Message Understanding Conference, Fairfax, USA, 1998.

J. Sun, J. Gao, L. Zhang, M. Zhou, and C. Huang, “Chinese Named Entity Identification using Class-based Language Model”, In Proc. of COLING-2002, Taipei, Taiwan, 2002.

Y. Wu, J. Zhao, and B. Xu, “Chinese Named Entity Recognition Combining Statistical Model and Human Knowledge”, In Proc. of the Workshop on Multilingual and Mixed-Language Named Entity Recognition, Sapporo, Japan, 2003.

S. Ye, Tat-Seng Chua, and J. Liu, “An Agent-based Approach to Chinese Named-entity Recognition”, In Proc. of COLING-2002, Taipei, Taiwan, 2002.

Y. Zhang, J. F. Zhou, “A Trainable Method for Extracting Chinese Entity Names and Their Relations”, In Proc. of the second Chinese Language Processing Workshop, ACL, Hong Kong, 2000.