Integrating Punctuation Rules and Naïve Bayesian Model for Chinese Creation Title Recognition

Conrad Chen and Hsin-Hsi Chen

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
drchen@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw
http://nlg.csie.ntu.edu.tw/

Abstract. Creation titles, i.e. titles of literary and/or artistic works, comprise over 7% of named entities in Chinese documents. They are the fourth large sort of named entities in Chinese other than personal names, location names, and organization names. However, they are rarely mentioned and studied before. Chinese title recognition is challenging for the following reasons. There are few internal features and nearly no restrictions in the naming style of titles. Their lengths and structures are varied. The worst of all, they are generally composed of common words, so that they look like common fragments of sentences. In this paper, we integrate punctuation rules, lexicon, and naïve Bayesian models to recognize creation titles in Chinese documents. This pioneer study shows a precision of 0.510 and a recall of 0.685 being achieved. The promising results can be integrated into Chinese segmentation, used to retrieve relevant information for specific titles, and so on.

1 Introduction

Named entities are important constituents to identify roles, meanings, and relationships in natural language sentences. However, named entities are productive, so that it is difficult to collect them in a lexicon exhaustively. They are usually "unknown" when we process natural language sentences. Recognizing named entities in documents is indispensable for many natural language applications such as information retrieval [2], summarization [3], question answering [7], and so on.

Identifying named entities is even harder in Chinese than in many Indo-European languages like English. In Chinese, there are no delimiters to mark word boundaries and no special features such as capitalizations to indicate proper nouns, which constitute huge part of named entities. In the past, various approaches [1, 4, 10] have been proposed to recognize Chinese named entities. Most of them just focused on MUC-style named entities [8], i.e., personal names, location names, and organization names. The extensive studies cover nearly 80% of named entities in real documents [1]. Although the performance of such kinds of named entity recognizers is satisfiable, the rest 20% of named entities are so far rarely mentioned and often ignored in previous studies.

These rarely mentioned ones belong to various sorts, such as terminologies, aliases and nicknames, brands, etc. These sorts may not occur as frequently as personal names or location names in a corpus, but the importance of the former in documents

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 838 – 848, 2005.

of specific domains is no less than that of the latter. For example, knowing names of dishes would be very important to understand articles about cooking. Among these rarely addressed named entities, titles of creations, such as book names, song titles, sculpture titles, etc., are one of the most important sorts. According to Chen & Lee (2004)'s study [1] of Academia Sinica Balanced Corpus (abbreviated ASBC corpus hereafter), about 7% of named entities are titles of creations. In other words, more than one-third of rarely mentioned named entities are titles of creations.

Chinese title recognition is challenging for the following reasons. There are no limitations in length and structures of titles. They might be a common word, e.g. "錯誤" (Mistakes, a Chinese poem), a phrase, e.g. "挪威的森林" (Norwegian Wood, a song), a sentence, e.g. "阿根廷別為我哭泣" (Don't Cry for Me Argentina, a song), or even like nothing, e.g. "摩擦・無以名狀" (Rub • Undescribable, a Chinese poetry collection). Besides, the choice of characters to name titles has no obvious preferences. Till now, few publications touch on Chinese title recognition. There are even no available corpora with titles being tagged.

Several QA systems, such as Sekine and Nobata (2004) [9], used fixed patterns and dictionaries to recognize part of titles in English or Japanese. Lee et al. (2004) [6] proposed an iterative method that constructs patterns and dictionaries to recognize English titles. Their method cannot be adapted to Chinese, however, because the most important feature employed is capitalization, which does not exist in Chinese.

In this paper, we propose a pioneer study of Chinese title recognition. An approach of integrating punctuation rules, lexicon, and naïve Bayesian models is employed to recognize creation titles in Chinese documents. Section 2 discusses some cues for Chinese title recognition. Section 3 gives a system overview. Punctuation rules and title gazetteer identify part of titles and filter out part of non-titles. The rest of undetermined candidates are verified by naïve Bayesian model. Section 4 addresses which features may be adopted in training naïve Bayesian model. Section 5 lists the training and testing materials, and shows experimental results. Section 6 concludes there marks.

2 Cues for Chinese Creation Title Recognition

Titles discussed in this paper cover a wide range of creations, including literature, music, painting, sculpture, dance, drama, movies, TV or radio programs, books, newspapers, magazines, research papers, albums, PC games, etc. All of these titles are treated as a single sort because they share the same characteristics, i.e., they are named by somebody with creativity, and thus there are nearly no regularity or limitations on their naming styles.

The challenging issue is that, unlike MUC-style named entities (MUC7, 1998), titles are usually composed of common words, and most of them have no internal features like surnames or entity affixes, e.g. "市" (City) in "台北市" (Taipei City). In other words, most titles might look just like common strings in sentences. Thus it is even more difficult to decide which fragment of sentences might be a title than to determine if some fragment is a title.

For the lack of internal features, external features or context information must be found to decide boundaries of titles. Table 1 shows some words preceding or following titles in one-tenth sampling of ASBC corpus with titles tagged manually. We can observe that quotation marks are widely used. This is because writers usually quote titles with punctuation marks to make them clear for readers. The most common used ones are the two pairs of quotation marks “ 「 」 ” and “ 『 』 ”. About 40% of titles are quoted in “ 「 」 ” or “ 『 』 ” in our test corpus. However, labeling proper nouns is only one of their functions. Quotation marks are extensively used in various purposes, like dialogues, emphasis, novel words, *etc*. In our analysis, only less than 7% of strings quoted in “ 「 」 ” or “ 『 』 ” are creation titles. It means the disambiguation of the usages of quotation marks is necessary.

Table 1. Preceding and Following Words of Titles in One-tenth Sampling of ASBC Corpus

Preceding Word	Frequency	Following Word	Frequency
「	450	」	442
《	216	》	216
（	44	、	56
。	31	，	32
、	25	（	26
，	24	』	16
的	19	的	13
『	16	。	11
是	9	中	7
在	7	）	6
·	6	】	6
【	6	裡	6

The most powerful external feature of creation titles is French quotes “《》”, which is defined to represent book names in the set of standard Simplified Chinese punctuation marks of China [5]. However, they are not standard punctuation marks in Traditional Chinese. Besides the usage of French quotes to mark book names, writers often use them to label various types of creation titles. According to our analysis on Web searching and the sampling corpus, about 20% of occurrences of titles in Traditional Chinese documents are quoted in “《》”, and nearly no strings other than titles would be quoted in “《》”. This punctuation mark shows a very powerful cue to deal with title recognition.

Nevertheless, there are still 40% of titles without any marks around. These unmarked titles usually stand for widely known or classic creations. In other words, these famous works are supposed to be mentioned in many documents many times. Such kinds of titles are extensively known by people like a common vocabulary. A lexicon of famous creations should cover a large part of these common titles.

3 System Overview

Based on the analyses in Section 2, we propose some punctuation rules that exploit the external features of titles to recognize possible boundaries of titles in Chinese documents. Most strings that cannot be titles are filtered by these rules. Titles with strong evidences like "《》" are also identified by these rules. The rest undecided strings are denoted as "possible titles." To verify these candidates is somewhat similar to solve word sense disambiguation problem. Naïve Bayesian classifier is adopted to tell whether a candidate is really a title or not. The overview of our system is shown in Figure 1.

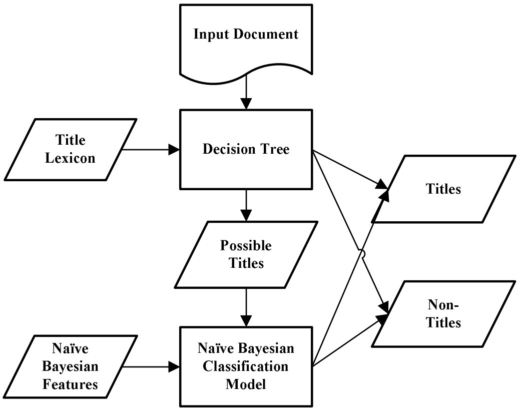

Fig. 1. System Overview

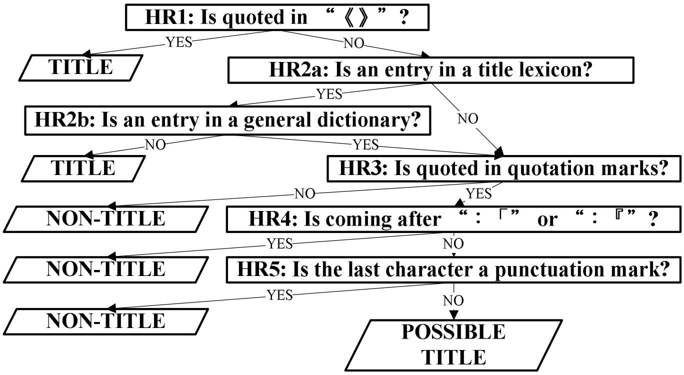
HR1: Is quoted in "《》"?
YES
NO
TITLE
HR2a: Is an entry in a title lexicon?
YES
NO
HR2b: Is an entry in a general dictionary?
NO
YES
TITLE
HR3: Is quoted in quotation marks?
NO
YES
NON-TITLE
HR4: Is coming after "：「" or "：『"?
YES
NO
NON-TITLE
HR5: Is the last character a punctuation mark?
YES
NO
NON-TITLE
POSSIBLE TITLE

Fig. 2. Decision Tree of Punctuation Rules and Lexicon

Figure 2 shows the applications of the punctuation rules and the title lexicon, which are illustrated as a decision tree. HR1 exploits French quotes "《》" to identify titles like "《桃花扇》" (Taohua Shan, a traditional Chinese drama by Kong, Shang-Ren) and "《迷宮中的將軍》" (El General En Su Laberinto, a novel by Garcia Marquez). HR2a and HR2b then look up the title lexicon to find famous titles like "百年孤寂" (Cien Anos de Soledad) and "三國演義" (Romance of Three Kingdoms). HR3 limits our recognition scope to strings quoted in quotation marks, and HR4 and

HR5 filter out a major sort of non-titles quoted in quotation marks, dialogues, such as "我說：「觀眾小心了！」" (I said, "the audience should be careful!") and "笛卡兒說：「我思，故我在。」" (Rene Descarte said, "I think, therefore I am.").

The title lexicon we use is acquired from the catalogues of the library of our university. These titles are sent to *Google* as query terms. Only the ones that have ever been quoted in "《》" in the first 1,000 returned summaries are kept. The remained titles are checked manually and those ones that possibly form a fragment of a common sentence are dropped to avoid false alarms. After filtering, there are about 7,200 entries in this lexicon. Although the lexicon could cover titles of books only, it is still useful because books are the major sort of creations.

The punctuation rules and lexicon divide all the strings of a document into three groups – say, titles, non-titles, and possible titles. All strings that cannot be definitely identified by the punctuation rules and lexicon are marked as "possible" titles. These possible titles are then verified by the second mechanism, the naïve Bayesian model. The naïve Bayesian model will be specified in the next section.

4 Naïve Bayesian Model

Naïve Bayesian classifier is widely used in various classification problems in natural language processing. Since it is simple to implement and easy to train, we adopt it in our system to verify the possible titles suggested by the decision tree.

Naïve Bayesian classification is based on the assumption that each feature being observed is independent of one another. The goal is to find the hypothesis that would maximize the posterior probability $P(H|F)$, where H denotes the classifying hypotheses and F denotes the features that determine H. According to Bayesian rule, the posterior probability can be rewritten as:

$$P(H \mid F) = P(H)\, P(F \mid H) \,/\, P(F) \tag{1}$$

Since $P(F)$ is always the same under different hypotheses, we only need to find which hypothesis would obtain the maximal value of $P(H)P(F|H)$. Besides, under the independence assumption, Equation (1) is rewritten into:

$$P(H \mid F) = P(H) \prod P(f_i \mid H) \quad \text{where } F = \{\, f_1, f_2, \ldots, f_n \,\} \tag{2}$$

In our system, we have two hypotheses:

H1: candidate S is a title
H2: candidate S is not a title

Four features shown below will be considered. The detail will be discussed in the subsequent paragraphs.

F1: Context
F2: Component
F3: Length
F4: Recurrence

Context. To exploit contextual features, our system adopts a word-based, position-free unigram context model with a window size of 5. In other words, our context model can be viewed as a combination of ten different contextual features of the naïve Bayesian classifier, five of them are left context and the other five are right context. It can be represented as:

$$P(F_{context}|H) = P(L_5, L_4, L_3, L_2, L_1, R_1, R_2, R_3, R_4, R_5 \mid H) \tag{3}$$

Where L_i and R_i denote preceding and following words of the possible title we want to verify, and H denotes the hypothesis.

If we postulate that the contextual features are independent of each other, then equation (3) can be transformed to:

$$P(F_{context}|H) = \prod P(L_i \mid H) \prod P(R_i \mid H) \tag{4}$$

Equation (4) assumes that the distance from a contextual word to a possible title is not concerned both in training and testing. The reason is that we do not have a realistic, vast, and well-tagged resource for training. On the other hand, if we want to exploit it in testing, we need a well-tagged corpus to learn the best weights we should assign to contextual words of different distances.

Component. *Context* deals with features surroundings titles. In contrast, *Component* further considers the features within titles. Similar to the above discussion, our component model is also a word-based, position-free unigram model. A possible title will be segmented into a word sequence by standard maximal matching. The words in the segmentation results are viewed as the "components" of the possible title, and the component model can be represented as:

$$P(F_{comp}|H) = P(C_1 \ldots C_n \mid H) = \prod P(C_i \mid H) \tag{5}$$

Where C$_i$ denotes the component of the possible title we want to verify, and H denotes the hypothesis.

Similar to the context model, the position of a component word is not concerned both in training and testing. Besides the availability issue of large training corpus, the lengths of possible titles are varied so that positional information is difficult to be exploited. Different titles consist of different number of component words. There are no straightforward or intuitive ways of using positional information.

Length. The definition of *Length* feature is the number of characters that constitute the possible title. It can be represented as:

$$P(F_{length}|H) = P(\textit{the length of } S \mid H) \tag{6}$$

Where S denotes the possible title to be verified and H denotes the hypothesis that S is a title.

Recurrence. The definition of *Recurrence* feature is number of occurrences of the possible title in the input document. It can be represented as:

$$P(F_{Rec}|H) = P(\textit{the appearing times of } S \mid H) \tag{7}$$

Where S denotes the possible title to be verified and H denotes the hypothesis that S is a title.

5 Experiment Results

The estimation of $P(H)$ and $P(F|H)$ is the major issue in naïve Bayesian model. There are no corpora with titles being tagged available. To overcome this problem, we used two different resources in our training process. The first one is a collection of about 300,000 titles, which is acquired from library catalogues of our university. This collection is used to estimate *Component* and *Length* features of titles. Besides, these titles are regarded as queries and submitted to *Google*. The returned summaries are segmented by maximal matching and then used to estimate *Context* features of titles. Since titles are usually composed of common words, not all query terms in retrieved results by *Google* are a title. Therefore, only the results with query terms quoted in French quotes "《》" are adopted, which include totally 1,183,451 web page summaries. Recall that French quotes are a powerful cue to recognize creation titles, which was discussed in Section 2.

The second resource used in training is ASBC corpus. Since titles in ASBC corpus are not specially tagged and we are short-handed to tag them by ourselves, a compromised approach is adopted. First, the decision tree shown in Figure 2 is used to group all strings of the training corpus into titles, non-titles, and possible titles. All titles thus extracted are used to estimate the *Recurrence* feature of titles, and all possible titles are treated as non-titles to estimate all features of non-titles. Since the probability of possible titles being titles are much less than being non-titles, the bias of the rough estimation is supposed to be tolerable.

We separate one-tenth of ASBC corpus and tag it manually as our testing data. The rest nine-tenth is used for training. There are totally 610,760 words in this piece of data, and 982 publication or creation titles are found. During execution of our system, the testing data are segmented by maximal matching to obtain context and component words of possible titles. To estimate $P(H)$, we randomly select 100 possible titles from the training part of ASBC corpus, and classify them into titles and non-titles manually. Then we count the probability of hypotheses from this small sample to approximate $P(H)$.

Table 2 shows the performance of the decision tree proposed in Figure 2 under the testing data. If we treat HR2a and HR2b as a single rule that asks "Is the string an entry in the title lexicon and not in a general dictionary?", we could view our rules as an ordered sequence of decisions. Each rule tells if a part of undecided strings are titles or non-titles, which is denoted in the column "Decision Type" of Table 2. The column "Decided" shows how many strings can be decided by the corresponding rules, while the columns of "Undecided Titles" and "Undecided Non-Titles" denote how many titles and non-titles are remained in the testing data after applying the corresponding rule. The correctness of the decision is denoted in the columns of "Correct" and "Wrong".

Table 2 shows that these five rules are very good clues to recognize titles. HR1, HR2, HR4 and HR5 have precisions of 100%, 94.01%, 99.15%, and 100% respectively. Because the number of non-titles is much larger than that of titles, the actual precision of HR3 is comparatively meaningless. These rules could efficiently solve a large part of the problem. The rest possible titles are then classified by the naïve Bayesian classifier. The performance is listed in Table 3. We try different combinations of the four features. F1, F2, F3, and F4 denote *Context*, *Component*, *Length*, and *Recurrence*, respectively. The number of True Positives, True Negatives, and False Positives are listed. Precision, recall and F-measure are considered as metrics to evaluate the performance.

Table 2. Performance of Decision Tree in Figure 2

	Decision Type	Decided	Correct	Wrong	Undecided Titles	Undecided Non-Titles
HR1	Title	216	216	0	766	$\sim\|corpus\|^2/2$
HR2	Title	167	126	41[1]	640	$\sim\|corpus\|^2/2$
HR3	Non-Title	$\sim\|corpus\|^2/2$	$\sim\|corpus\|^2/2$	186	454	5812
HR4	Non-Title	1997	1980	17	437	3832
HR5	Non-Title	372	372	0	437	3458

Note that there are two different numbers in the False Positive, Precision, and F-measure columns in Table 3. The left number shows the total number of false positive errors, and the right one ignores the errors caused by other sorts of named entities. This is because many false positive errors come from other types of named entities. For example, in the sentence "參加「一九九四年第三十五屆國際數學奧林匹克競賽」" (attend 1994 35th International Mathematical Olympiad), "一九九四年第三十五屆國際數學奧林匹克競賽" ("1994 35th International Mathematical Olympiad") is a contest name, however, ill-recognized as a title by our system. Because there are various sorts of ill-recognized named entities and most of them have not been thoroughly studied, there are no efficient ways available to solve these false alarms. Fortunately, in many applications, there would be little harm incorrectly recognizing these named entities as titles.

The other major source of false positive errors is appearances of monosyllabic words. For example, in the sentence "「以德報怨」是老子的話" ("Render Good for Evil" is Lao Tzu's speech), "以德報怨" ("Render Good for Evil") are ill-recognized as titles. The reason might be that many context and component words of titles are named entities or unknown words. During training, these named entities are neither tagged nor recognized, so that most of these named entities are segmented into sequences of monosyllabic words. Therefore, while the naïve Bayesian classifier encounters monosyllabic context or component words, it would prefer recognizing the possible title as a title.

From Table 3, we could observe that *Context* and *Component* are supportive in both precision and recall. *Length* boosts precision but decreases recall while *Recurrence* is on the contrary. The combination of F1+F2+F3 obtains the best F-measure, but the combination of all features might be more useful in practical applications,

[1] Total 31 of them can be easily corrected by a maximal-matching-driven segmentation. For example, "心經" (xīn jīng, Heart Sutra, a Buddha book) in "用心經營" (yòng xīn jīng yíng) is an entry in the title lexicon. However, maximal matching prefers the segmentation of "用心 / 經營" (yòng xīn/jīng yíng) than "用 / 心經 / 營" (yòng/xīn jīng/yíng), so that this false alarm would be recovered.

since it only sacrifices 1.4% of precision but gains 3% of recall in comparison with the former. Table 4 summaries the total performance of our creation title recognition system. It achieves the F-measure of 0.585.

Table 3. Performance of the Naïve Bayesian Classifier Using Different Features

	True Positive	True Negative	False Positive	Precision	Recall	F-measure
F1	277	160	959 / 772	0.224 / 0.264	0.634	0.331 / 0.373
F2	153	284	532 / 332	0.223 / 0.315	0.350	0.273 / 0.332
F1 + F3	273	164	859 / 676	0.241 / 0.288	0.625	0.348 / 0.394
F2 + F3	148	289	453 / 247	0.246 / 0.375	0.339	0.285 / 0.356
F1 + F2	288	149	976 / 722	0.228 / 0.285	0.659	0.339 / 0.398
F1 + F4	289	148	1067 / 867	0.213 / 0.250	0.661	0.322 / 0.363
F2 + F4	169	268	695 / 467	0.196 / 0.266	0.387	0.260 / 0.315
F1 + F2 + F3	286	151	888 / 631	0.244 / 0.312	0.654	0.355 / 0.422
F1 + F3 + F4	285	152	946 / 750	0.232 / 0.275	0.652	0.342 / 0.387
F2 + F3 + F4	164	273	542 / 320	0.232 / 0.339	0.375	0.287 / 0.356
All	299	138	967 / 703	0.236 / 0.298	0.684	0.351 / 0.416

Table 4. Performance of the Title Recognition System

	True Positive	True Negative	False Positive	Precision	Recall	F-measure
Decision Tree	342	203	41 / 10	0.915 / 0.978	0.685	0.783 / 0.806
Naïve Bayesian	299	138	967 / 703	0.236 / 0.298	0.684	0.351 / 0.416
Total	641	341	1008 / 713	0.424 / 0.510	0.685	0.524 / 0.585

6 Conclusion

This paper presents a pioneer study of Chinese title recognition. It achieves the precision of 0.510 and the recall of 0.685. The experiments reveal much valuable information and experiences for further researches.

First, the punctuation rules proposed in this paper are useful to recognize creation titles with a high precision. They can relief our burdens in building more resources, make supervised learning feasible, and give us some clues in similar studies like recognition of other sorts of named entities. These useful rules are also helpful for those applications needing high accuracies. For example, we can exploit these rules on an information retrieval system to filter out noises and show only the information about the requested creation or the publication.

Second, naïve Bayesian classifier could achieve a comparable recall on the verification of possible titles. Since we only adopt simple features and use a rough estimation in feature model building, the result shows that naïve Bayesian classifier is practicable in recognizing creation titles. In future works, we may find other useful features and adopt more sophisticated models in naïve Bayesian classifier to seek a higher performance, especially in precision.

Third, our result shows that recognizing rarely seen sorts of named entities is practicable. Because un-recognized named entities might significantly affect subsequent applications in Chinese, in particular, segmentation, we should not ignore the problems introduced by Non-MUC style named entities. Our study suggests that the recognition of these rarely mentioned named entities is promising. The performances of many applications, such as natural language parsing and understanding, might be boosted through adding the mechanism of recognizing these rare named entities.

Finally, our research can also be extended to other oriental languages, such as Japanese, in which there are no explicit features like specialized delimiters or capitalizations to mark creation titles. Just as Chinese, un-recognized named entities in these languages might affect the performances of natural language applications. Recognizing Non-MUC style named entities is an indispensable task to process these languages.

Acknowledgement

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC94-2752-E001-001-PAE.

References

1. Chen, Conrad and Lee, Hsi-Jian. 2004. A Three-Phase System for Chinese Named Entity Recognition, Proceedings of ROCLING XVI, 2004, 39-48.
2. Chen, Hsin-Hsi, Ding, Yung-Wei and Tsai, Shih-Chung. 1998. Named Entity Extraction for Information Retrieval, Computer Processing of Oriental Languages, Special Issue on Information Retrieval on Oriental Languages, 12(1), 1998, 75-85.
3. Chen, Hsin-Hsi, Kuo, June-Jei, Huang, Sheng-Jie, Lin, Chuan-Jie and Wung, Hung-Chia. 2003. A Summarization System for Chinese News from Multiple Sources, Journal of American Society for Information Science and Technology, 54(13), November 2003, 1224-1236.
4. Chen, Zheng, W. Y. Liu, and F. Zhang. 2002. A New Statistical Approach to Personal Name Extraction, Proceedings of ICML 2002, 67-74.
5. Gong, Chian-Yian and Liu, Yi-Ling. 1996. Use of Punctuation Mark. GB/T15834-1995. http://202.205.177.129/moe-dept/yuxin-/content/gfbz/ managed/020.htm
6. Lee, Joo-Young, Song, Young-In, Kim, Sang-Bum, Chung, Hoojung and Rim, Hae-Chang. 2004. Title Recognition Using Lexical Pattern and Entity Dictionary, Proceedings of AIRS04, 342-348.
7. Lin, Chuan-Jie, Chen, Hsin-Hsi, Liu, Che-Chia, Tsai, Ching-Ho and Wung, Hung-Chia. 2001. Open Domain Question Answering on Heterogeneous Data, Proceedings of ACL Workshop on Human Language Technology and Knowledge Management, July 6-7 2001, Toulouse France, 79-85.

8. MUC7. 1998. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 1998, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.
9. Sakine, Satoshi and Nobata, Chikashi. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy, Proceedings of LREC04.
10. Sun, Jian, J. F. Gao, L. Zhang, M. Zhou, and C. N. Huang. 2002. Chinese Named Entity Identification Using Class-based Language Model, Proceedings of the 19th International Conference on Computational Linguistics, Taipei, 967-973