# Knowledge Extraction for Identification of Chinese Organization Names

*Keh-Jiann Chen & Chao-jan Chen*
kchen@iis.sinica.edu.tw richard@iis.sinica.edu.tw
Institute of Information Science, Academia Sinica, Taipei

## ABSTRACT

In this paper, a knowledge extraction process was proposed to extract the knowledge for identifying Chinese organization names. The knowledge extraction process utilizes the structure property, statistical property as well as partial linguistic knowledge of the organization names to extract new organizations from domain texts. The knowledge extraction processes were experimented on large amount of texts retrieved from WWW. With high standard of threshold values, new organization names can be identified with very high precision. Therefore the knowledge extraction processes can be carried out automatically to self improve the performance in the future.

## 1. INTRODUCTION

The occurrences of unknown words cause difficulties in natural language processing. The word set of a natural language is open-ended. There is no way of collecting every words of a language, since new words will be created for expressing new concepts, new inventions, newborn babies, new organizations. Therefore how to identify new words in a text will be the most challenging task for natural language processing. It is especially true for Chinese. Each Chinese morpheme (usually a single character) carries meanings and most are polysemous. New words are easily constructed by combining morphemes and their meanings are the semantic composition of morpheme components. However there are also semantically non-compositional compounds, such as proper names. In Chinese text, there is no blank to mark word boundaries and no inflectional markers nor capitalization markers to denote the syntactic or semantic types of new words. Hence the unknown word identification for Chinese became one of the most difficult and demanding research topic.

There are many different types of unknown words and each has different morph-syntactic and morph-semantic structures. In principle their syntactic and semantic categories can be determined by their content and contextual information, but there are many difficult problems have to be solved. First of all it is not possible to find a uniform representational schema and categorization algorithm to handle different types of unknown words due to their different morph-syntactic structures. Second, the clues for identifying different type of unknown words are also different. For instance, identification of names of Chinese people is very much relied on the surnames, which is a limited set of characters. The statistical methods are commonly used for identifying proper names (Chen & Lee 1996, Chang et al. 1994, Sun et al. 1994). The identification of general compounds is more relied on the morphemes and the semantic relations between morphemes (Chen & Chen 2000). The third difficulty is the problems of ambiguities, such as structure ambiguities, syntactic ambiguities and semantic ambiguities. For instances, usually a morpheme character/word has multiple meaning and syntactic categories and may play the roles of common words or proper names. Therefore the ambiguity resolution became one of the major tasks.

In this paper we focus our attention on the identification of the organization names. It is considered to be a hard task to identify organization names in comparing with the identification of other types of unknown words, because there are not much morph-syntactic and morph-semantic clues to indicate an organization name. There is no significant preference on the selection of morphemes/ characters and the semantic of the morphemes, which gives no clue leading toward the identification. For instance, '微軟, micro-soft' (Microsoft) has the character by character (morpheme by morpheme) translation of 'slightly soft' and there is no marker, such as capitalization, to indicate that it is a proper name. The only reliable clue is its context information. However an organization's full names usually occur at its first mention, unless it is a well-known organization. A full name contains its proper name and organization type, such as ‘宏碁 電腦公司 , Acer Computer-Company’. The organization types became the major clue of identifying a new organization name. However abbreviated shorter names usually will be used, such as a) omit part of the organization type, for instances ‘宏碁 電腦, Acer Computer’, ‘宏碁公司, Acer Company’, b) omit the organization type totally, for instance ‘宏碁, Acer’, or c) the abbreviation, for instance ‘宏電, global-electric (Acer-computer)’. Therefore the task became not only the identification of organization names in different forms but also finding their meaning equivalence classes. To achieve the above goal, the knowledge of 1) proper names of organizations, 2) different lines of the businesses, and 3) different organization types, should be equipped. Unfortunately there is no well-prepared knowledge sources containing the above information. Therefore a knowledge extraction model is proposed to extract the above mentioned knowledge from the dictionary and domain texts.

## 2. STRUCTURES OF ORGANIZATION NAMES

There is no rigid structure for an organization name as mentioned in the previous section. Roughly speaking an organization name is composed by two major components. The first part is the proper name and the second part is the organization type. The second part contains the major key words lead toward the identification of an organization, since the organization types, such as ‘公司, company’, ‘基金會 foundation’, ‘小組 group’, ‘集團 enterprise’ etc, tells what kind of organizations they are. If it is a company, to be more informative the line of business usually goes with the key word ‘公司 company’, for instances ‘食品公司 food company’, ‘電腦公司 computer company’, ‘投資顧問公司 investment consultant company’, but in most cases the keyword ‘公司 company’ will be ignored, such as 統一食品( President food). Sometimes the line of business and the organization type go together to become a single word, such as ‘中學 middle school’，’銀行 bank’，’醫院 hospital’. By observing the structure of the organization name, it seems that once a complete list of the organization types is well prepared, then it is not hard to identify the organizations by their full names. The only complication is that abbreviated names occur more frequently than full names. The identifier ‘公司 company’ is usually ignored in real text. The lines of business became the major identifier for a company and many business lines are common words, such as ‘食品 food’, ‘電腦 computer’, ‘水泥 cement’. Therefore it is necessary to make the distinction between a common compounds and a company name, for examples, ‘健康食品 health food’ vs. ‘統一食品 President food’, ‘個人電腦 personal computer’ vs. ‘宏碁電腦 Acer computer’. Although they are two-way ambiguous, usually they have only one preference reading.

In conclusion, the types and the proper names of organizations will be the major clues lead toward the identification of the organizations. In addition, it is also better to have a list of well known organization names, such that the well known company names, like ‘微軟 Microsoft’, can be identified immediately. Most of the knowledge preparation works should be done by offline approaches. The prepared knowledge would be utilized to online identification of newly coined organizations. The equivalent classes of the well-known organizations are also classified by a similarity-based approach.

## 3. KNOWLEDGE EXTRACTION

There are two knowledge sources. One is the CKIP Chinese lexicon and another is the Chinese text from WWW. The lexicon provides a partial list of important organizations and the information extracted from them will be the initial knowledge of the identification system. The texts from WWW provide ample of new organization names implicitly. The problem is how to extract some, if not all, of them from the texts. Once we have a list of organization names. The proper names for organizations and the organization types will be extracted by analyzing the morphological structures of the organization names. However an effective morphological analyzer depends upon the availability of the knowledge of the organization types, but the lists of the organization types are not available yet.

As we mentioned before the complete organization names have two parts. The first part is the proper name and the second part is the organization type. The number of different proper names is unlimited and on the other hand the number of different organization types is limited. This property will be utilized to separate the variable parts, i.e. the proper name, and the constant parts, i.e. the organization type, from the organization names.

The numbers of organization names in the lexicon is very limited, since only the important organizations in the common domain will be collected. Therefore the initial knowledge extracted from lexicon is also very limited. To make the sources of knowledge more adequate, vast amount of new organization names should be extracted from each different domain corpus. Unfortunately none of the existing corpora had tagged the organization names. Therefore we are going to design a semi-automatic method to extract the high frequency organization names from text corpora.

The locality of occurrences of keywords in a text will be utilized for keyword extraction. Once an organization name occurring in a text it is very probably reoccurred in the same text. The recurrence property had been utilized to extract keywords or key-phrases from text (Chien 1999, Fung 1998, Smadja 1993). However not all keywords are organization names. The knowledge extracted from the lexicon, i.e. the list of the organization types will be the initial knowledge for identifying organization names. In addition to the initial knowledge, the structure property of the organization names will be also utilized in classifying extracted keywords into organization names and non-organization names. The extraction processes will be repeated for extracting new organizations and therefore extracting new organization types. The more knowledge would have been extracted the more accurate of the organization identification will achieve.

### 3.1 Morphological Analysis for Organization Names

There are 1391 number of words in the CKIP lexicon classified as organizations. Table 1 shows some of the examples.

一女中 人代會 人民公社 人事行政局 人事室
人事處 人事部 十信 八號分機 三軍總醫院
三商隊 三商銀 三專 三輕 三總 下院 下議院
上議院 土木系土地局 土地銀行 土銀 大仁藥專
大同工學院 大同商專 大同盟 大使館 大英國協
大英博物館 大專 大通銀行 大華工專 大漢工商
大學 女青年會

Table 1. The samples of organizations from the CKIP dictionary

As we observed, the morphological structure of an organization name usually is a compounding of a proper name and a organization type. The organization type might be a compounding of a line of business and a type, for instances 電腦公司 (computer company), 銀行(bank), 中學 (middle school), or simply a line of business, for instances 食品(food), 電腦(computer), 水泥 (cement). The proper names are variables, since each organization type may have many different institutions with different names. The types are constants. There is a limited number of constants attached with many different proper names to form different organization names. Therefore to extract the organization types is equivalent to extract the high frequency ending morphemes. Table 2 shows the top 20 high frequency ending morphemes extracted from the 1391 organization names and in deed they are organization types.

| | | | | |
|---|---|---|---|---|
| 52 會 | 38 局 | 36 部 | 30 社 | 29 大學 |
| 21 黨 | 21 處 | 20 院 | 17 科 | 16 署 |
| 16 校 | 16 委會 | 15 所 | 15 系 | 15 工專 |
| 13 隊 | 12 協 | 12 司 | 12 公司 | 11 大 |

Table 2. The top 20 organization types ranked with their occurrence frequencies extracted from the 1391 organization names

### 3.2 Automatic Extraction of Organization Names

A Web spider can extract text from each different domain through WWW. Then keyword extraction technique is applied on domain texts to retrieve possible keywords. The keyword set includes organization names, personal names, general compounds, and also error extraction. Most of which are not organization names. It is supposed that the available list of the organization types will be the source of knowledge to identify candidates of organizations. However such a method only identifies the organizations of the known organization types and provides new proper names only. It will not identify new types of organizations. Therefore we use a new method to extract the organization names by using the structure property of organization names.

**Extraction Algorithm for Organization Types:**

Step 1. Using a Web spider to collect Chinese texts of a fixed domain, such as domain of finance and business, from WWW.

Step 2. Extract high frequency keywords in the text (Smadja 93, Chang & Su 97, Chien 99).

Step 3. For the keywords of length 3,4, and 5, each keyword is divided into two parts X and Y. X is a candidate of proper name and

Y is a candidate of organization type. The X is the initial two-characters of the keyword and Y is the remained characters. (Since most proper names of organizations have two characters, we can extract the organization types of the lengths 1, 2 and 3 from three different groups of keywords with lengths 3, 4 and 5 respectively.) Extract the organization type Y, if for some keywords X+Y, the following conditions hold.

a) X satisfies one of the following cases.
   1. X is not in the lexicon, i.e. X is an unknown word.
   2. X has the categories of Nb or Nc, i.e. it is a known proper name (Nb) or a location name (Nc).

b) For each Y, assumed to be the organization type, there must have more than n number of different X, such that X+Y in the extracted keyword list. In practice, the threshold value n was set to 2.

In general, Chinese company names like most proper names are non-common word (unknown words). However sometimes they are place names (Nc), but rarely they are common nouns, adjectives, or verbs. Therefore in order to avoid too many false alarms, such as "超級電腦 super computer", to be considered as a company name, the condition a) of step 3 is set. The reason to setup the condition b) is that each organization type Y should have many different organizations which have the same organization type Y, such as ‘宏碁電腦 Acer computer’, ‘國眾電腦 Leo computer’, ‘藍天電腦 Blue-sky computer’, ...etc.. The real implementation shows the different threshold value n gives the different precision and recall for identification.

For the first iteration of knowledge extraction, we suggest to have higher recall rate. Set the threshold value low and manually select the final list of the organization types. For the future automatic knowledge extraction, in order to increase the precision of the information extraction higher threshold values are suggested.

## 4. EXPERIMENTAL RESULTS

The knowledge extraction processes for Chinese organization names are carried out by different stages. At the first stage, the words marked with semantic category of organization were accessed from the CKIP dictionary. There are 1391 word organization types. As mentioned in section 3.1, a pseudo morphological analysis process was carried out, which try to find the high frequency ending morphemes. Since the structure of an organization name is a composition of X+Y, where X is a proper name and Y is a organization type. There are 546 different ending morphemes. The high frequency ending morphemes are exactly to be the morphemes for common organization types. Many of them are monosyllabic words and they are polysemous, as shown in Table 1. For the future identification, the disambiguation process has to be carried out for those polysemous ending morphemes (Chen & Chen 2000). The extracted morphemes and list of organizations will be the first collection of the organization types.

At the second stage, we try to extract new organizations names from different domain text. Each different domain has many new organization types. For instance in the domain of finance and business, there are many company names, which have completely different word strings for the organization types as in the extracted list by the first stage.

The algorithm shown in the section 3.2 was carried out. At the step 1, 31787 texts of news of the finance and business domain were extracted from http://www.cnyes.com. At step 2, 40675 keywords were extracted from the news corpus. At step 3, organization names were identified and the organization types were extracted. If the threshold value n = 2, 92 types were extracted and among them 83 are correct organization types. The precision is 90%. If the threshold was set to 3, only 56 types were extracted and all of them happen to be correct. The precision increased to 100%, but of course the recall rate dropped. We don't know the exact recall rate, since there are too many keywords in the training set. However the recall rate is not important, since the whole knowledge extraction process is a recurrent process. The knowledge extraction procedures should be repeatedly applied on the different set of text and at each iteration more information will be extracted. Hence the precision is much more important than the recall. The knowledge sources for future identification of organizations are the accumulated lists of the organization names, the proper names of organizations and the organization types.

Table 3 contains the extracted organization types while the threshold value n=3. The organization types are classified by their lengths and sorted by their frequencies of uses. Table 4 contains the extracted organization types which associated with exactly two different names and the last line shows the error extractions. Among newly extracted organization types only 23 of them are already in the old list.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 證 | 24 | 銀 | 14 | 電 | 9 | 市 | 6 | 網 | | | |
| 證券 | 81 | 科技 | 74 | 集團 | 46 | 銀行 | 41 | 電子 | 40 | 公司 | 36 |
| 電腦 | 19 | 建設 | 18 | 電訊 | 17 | 企業 | 16 | 汽車 | 15 | 國際 | 15 |
| 資訊 | 15 | 電信 | 15 | 航空 | 14 | 投信 | 13 | 實業 | 11 | 日報 | 10 |
| 工業 | 9 | 基金 | 9 | 光電 | 8 | 通信 | 7 | 控股 | 6 | 開發 | 6 |
| 電機 | 6 | 精密 | 6 | 人壽 | 5 | 紡織 | 5 | 商銀 | 5 | 網絡 | 5 |
| 興業 | 5 | 化學 | 4 | 石化 | 4 | 企銀 | 4 | 百貨 | 4 | 重工 | 4 |
| 時報 | 4 | 電力 | 4 | 電工 | 4 | 工程 | 3 | 化工 | 3 | 瓦斯 | 3 |
| 石油 | 3 | 金屬 | 3 | 電話 | 3 | 電器 | 3 | 網路 | 3 | 數碼 | 3 |
| 線上 | 3 | | | | | | | | | | |
| 半導體 | 7 | 交易所 | 6 | | | | | | | | |

Table 3. The extracted organization types associated with the number of different names >=3

中心 五金 水泥 委員會 房屋 信託 保全 保險
旅遊網 海運 紙業 動力 商事 啤酒 組織 貨運 貨櫃
郵報 塑膠 新聞 精工 寬頻 廣場 數碼港 聯網
證券報 證監會
*大西洋 *先生 *先進 *明珠 *指數 *添惠 *概念股
*總統 *創業板

Table 4. The extracted organization types associated with two different names and the last lines show the error extractions.

### 4.1 Strategies for On-line Identification of Organization Names

The knowledge about organizations extracted from the dictionaries and domain texts will be used to identify organization names at on-line sentence processing. During the word segmentation process, an organization name is either identified immediately (if it is a known organization name), or it will be segmented into two segments of X+Y or several segments of (x1+x2+...+xn)+Y, where X is a proper names, Y is the organization type. When the proper name X is a new word, it will be segmented into shorter segments (x1+x2+...+xn). To simplify the experiment process, we assume the proper names X are either the words of categories Nb (i.e. proper names) or Nc (i.e. the place names) or a two-character unknown word. For the identification experiment, a corpus extract from a T.V. news（http://www.ttv.com.tw）

The patterns of X+Y in the testing corpus were searched. 117 different organizations were identified. Among them 56 are known organizations, i.e. they are in the organization name list. 61 of them are identified by the composition of X+Y and 52 of them are correct. It counts the precision of 52/61=85% for identifying new names. The total performance is the precision of 108/117=92%.

The knowledge-based approach for identifying organization names seems very promising. It outperforms the reports of the precision of 61.79% and the recall of 54.50% in (Chen & Lee 1996) and the experiment was carried out under the condition that the knowledge extraction process is in its initial stage. We expect that performance of the algorithm will become better and better while the knowledge extraction process continuously performs.

### 4.2 Automatic Extraction of Name Equivalent Classes

The abbreviated names are very frequently occurred in the real text especially in the domain of the stock market. By observing the abbreviation names, the heuristic rules for abbreviating a company name can be concluded as follows.

Abbreviation rule: If the proper name of a company is unique, then take the proper name as its abbreviation name, such as '微軟, Microsoft'. Otherwise the abbreviation will be a compound of key-characters from part of its proper name and part of its line of business, such as 中油 is the abbreviation of '中國石油, China petroleum'.

An experiment was carried out to find the full names of the abbreviations of company names shown in the price table of the Taiwan stock market. The purposes of this experiment are a) to find the equivalent classes of company names and b) to have some idea about the recall rate of the current knowledge extraction process.

The matching process between the abbreviations and the extracted organization name lists is as follows.

1. For each abbreviation name matches the organization names in the organization name list. Find all the organization names containing the abbreviation name.
2. Rank the matched organization names according to the following criterion.
   The first rank: The proper name of the organization name is exactly matched with the abbreviation name.
   The second rank: The abbreviation is compounding of key-characters from part of the proper name and part of the line of business of the matched organization names.
   If there are many candidates with the same rank, then rank them according to their frequencies occurring in the training corpus.

There are 471 abbreviated company names in the price list of the stock market. 302 of them have matched candidates. Each abbreviation name may match many different organization names. The recall rate for the top ranked candidate is 282/471=60%. The precision of the first rank candidate is 282/302=93%. Table 5 shows some of the results.

| Abbr. | Candidates arranged in the order of their ranks |
|---|---|
| 泰山 | **泰山企業** 8 泰山集團 3 泰山電子 2 |
| 福壽 | **福壽實業** 7 幸福人壽 9 宏福人壽 2 |
| 惠勝 | **惠勝實業** 11 |
| 南亞 | 南亞科技 101 南亞興業 2 |
| 國喬 | **國喬石化** 48 國喬石油 2 |
| 中石化 | **中國石化** 10 中間石化 2 中油石化 1 |
| 宏電 | **宏電集團** 5 宏碁電腦 22 宏泰電工 22 |
|  | 旺宏電子 21 |
| 台達電 | **台達電子** 6 |
| 華通 | **華通電腦** 15 華南通信 2 華智通信 2 |
| 台揚 | **台揚科技** 11 |
| ?大眾 | 大眾電信 67 **大眾電腦** 29 大眾銀行 17 |
|  | 大眾集團 6 大眾網路 2 大眾科技 2 |
|  | 大眾投信 2 |

Table 5. Some examples of the abbreviations and the matched candidates (the correct answer is highlighted by the boldface characters)

## 5. CONCLUSIONS

The knowledge extraction process will be continuously carried on in the future. The accumulated knowledge will be utilized for the on-line unknown word identification as well as for the off-line knowledge extraction. The proposed knowledge extraction processing model can be generalized to extract other types of linguistics or morphological knowledge, for instances, to extract the transliterate foreign names, to extract the titles of people.

Some of the errors are caused by that the titles of the people are wrongly identified as organization types, since the patterns of people's name followed by their title are commonly occurred in real text. These patterns are similar to the structures of organization names. Such kind of errors can be avoid, if the titles of people are known and in fact the titles of people can be extracted by the same extraction model except that most of people's names have three characters instead of two.

In the future, the knowledge extraction processes will be automatically carried out. We expect that it will be one of the major building blocks for automatic learning systems for Chinese morphology and sentence processing.

## REFERENCES

[1] Bai, M.H., C.J. Chen & K.J. Chen, 1998, "POS tagging for Chinese Unknown Words by Contextual Rules" *Proceedings of ROCLING*, pp.47-62.

[2] Chang, J. S.,S.D. Chen, S. J. Ker, Y. Chen, & J. Liu,1994 "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts", *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 1, pp. 75-85.

[3] Chang, Jing Shin and Keh-Yih Su, 1997," An Unsupervised Iterative Method for Chinese New Lexicon Extraction," Computational Linguistics and Chinese Language Processing, Vol. 2 #2, pp97-147.

[4] Chen, Keh-Jiann, Ming-Hong Bai, 1997, "Unknown Word Detection for Chinese by a Corpus-based Learning Method." *Proceedings of the 10th Research on Computational Linguistics International Conference*, pp159-174.

[5] Chen, K.J. & Chao-jan Chen, 2000," Automati Semantic Classification for Chinese Unknown Compound Nouns," Coling 2000.

[6] Chen, K.J. & S.H. Liu, 1992,"Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling*, pp. 101-107.

[7] Chen, Hsin-His & Jen-Chang Lee, 1996," Identification and Classification of Proper Nouns in Chinese Texts," Proceedings of Coling-96, Vol. 1., pp. 222-229.

[8] Chien, Lee-feng, 1999," PAT-tree-based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval," Information Processing and Management, Vol. 35, pp. 501-521.

[9] Fung P., 1998," Extracting Key Terms from Chinese and Japanese Texts," Computer Processing of Oriental Languages, Vol. 12, #1, pp 99-122.

[10] Lee, J.C. , Y.S. Lee and H.H. Chen, 1994, "Identification of Personal Names in Chinese Texts." *Proceedings of 7th ROC Computational Linguistics Conference.*

[11] Lin, M. Y., T. H. Chiang, & K. Y. Su, 1993," A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation" *Proceedings of Rocling VI*, pp 119-137.

[12] McDonald D., 1996, " Internal and

External Evidence in the Identification and Semantic Categorization of Proper Names", in *Corpus Processing for Lexical Acquisition*, J. Pustejovsky and B. Boguraev Eds, MIT Press 1996.

[13] Smadja, Frank, 1993,"Retrieving Collocations from Text: Xtract," Computational Linguistics, vil. 19, #1, pp. 143-177.

[14] Sun, M. S., C.N. Huang, H.Y. Gao, & Jie Fang, 1994, "Identifying Chinese Names in Unrestricted Texts", *Communication of COLIPS*, Vol.4 No. 2. 113-122.