# Recent Results on Domain-Specific Term Extraction from Online Chinese Text Resources

Lee-Feng Chien[1], Chun-Liang Chen[2], Wen-Hsiang Lu[3] and Yuan-Lu Chang[1]

1. Institute of Information Science, Academia Sinica
2. Dept. of CS and IE, National Taiwan University
3. Dept. of Computer Science and Information Engineering,
National Chiao Tung University, Taipei, Taiwan, R.O.C.

## Abstract

This paper is to introduce recent results of an ongoing research called *Live Dictionary Construction,* which is investigating a number of efficient techniques for IR systems to automatically acquire Chinese terminological knowledge including domain-specific terms and similar terms from online text resources. Such research effort is pursued to be able to build a dynamic dictionary with IR systems, in which most of the necessary dictionary information can be dynamically extracted and adapted with the change of the indexed online resources. According to the obtained experimental results so far, it is promising that a live dictionary can be established and automatically grow.

## 1. Introduction

Automatic extraction of domain-specific terminological knowledge, such as keyterms and similar terms from online text collections is significant but very challenging for developing more effective information retrieval and also natural language processing systems. In this paper, we intend to introduce recent results of an ongoing research called *Live Dictionary Construction,* which is investigating a number of efficient techniques including *corpus classification*, *term extraction*, named entity extraction, similar term extraction to automatically acquire Chinese terminological knowledge. The research is pursued to build a *live dictionary* with IR and also NLP systems, in which most of the necessary dictionary information can be dynamically extracted and adapted with collection change.

Whether the employed dictionary is rigid and suitable for the database domain is very crucial in designing an effective IR system. It is clear that a well-prepared dictionary can help to identify representative keyterms in document indexing, find relevant terms in query expansion and perform exact term translation in cross-language information retrieval[Lewis'96, Wan'97]. Unfortunately, online resources most increase very fast. To most of the existing IR systems, it is cost-ineffective and even unrealistic to manually construct a domain-specific dictionary for each searching database. To avoid too many unknown searching terms and term translations appearing in database retrieval, the construction of a live dictionary which can grow with the update of the database like Altavista's LiveTopic is believed an alternative solution.

Our ongoing research is known as one of a few works towards the systematic construction of live dictionary for IR applications. The approach proposed for this purpose is based on proper integration of linguistic knowledge acquisition and IR technologies. This approach has achieved several technical breakthroughs. Like the technique designed for domain-specific term extraction, it has been proven performing well in extracting new terms incrementally. Compared with conventional research on knowledge acquisition[Zernik'91], the proposed approach has carefully considered the incremental characteristics of online information service. The developed techniques are all capable of handling large and dynamic texts and also easily to be integrated with IR systems. According to the obtained experimental results so far, it is optimistic that a live dictionary can be established.

## 2. Previous Work - PAT-tree-based Term Extraction

Keyterm extraction is frequently used in document classification and many other information retrieval applications. Since in Chinese language there is no "blanks" between words serving as word boundaries in printed and written sentences and the words are actually not well-defined, keyterm extraction has been a much more difficult and challenging problem in Chinese language processing as compared to western languages. An efficient approach for keyword extraction from Chinese texts has been developed previously, in which the difficult problem of large numbers of out-of-vocabulary words outside of any given lexicon and the sophisticated problem of word segmentation from sentences can both be avoided, and keywords or concatenated keywords (key terms) of arbitrary length which are very useful in information retrieval can be successfully extracted [Chien'97, Chang'99]. This approach is statistics-based and efficient in extracting major "significant lexical patterns (SLP)" from the Chinese texts.

## 3. Overview of the Proposed Approach

The proposed approach is formed as an abstract diagram shown in Fig. 1, where an IR system is designed as a composition of a searching engine and a live dictionary subsystem. The purpose of the live dictionary subsystem is trying to dynamically produce domain-specific term lists, term associations, and low-frequency named entities for the use of the searching engine. Such a subsystem contains several working modules, i.e., corpus storage and classification module, term extraction module, similar term extraction module, and named entity extraction module.

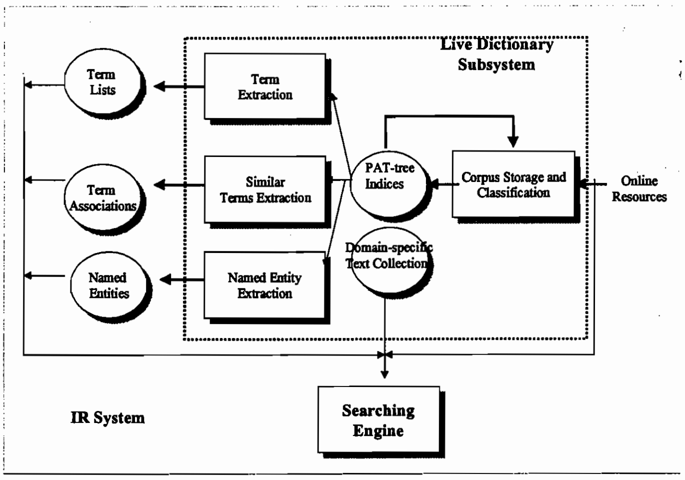

Fig. 1 An abstract diagram showing the proposed approach for live dictionary construction.

The execution of the corpus storage and classification module is the first step to construct a live dictionary. To allow domain-specific dictionary information can be effectively extracted, each input online document needs to be classified into corresponding collection(s) and serves as a training corpus for subsequent information extraction. Considering the demand on both document retrieval and corpus utilization, a method which employs *PAT trees* as the working structure for corpus storage and classification is presented. Each classifying document will be generated a PAT tree which records the occurrences of all of the composed character strings as the feature vector of the document, and then compares with the corresponding PAT-tree indices of each text collection in the system, by means of vector-space-based similarity estimation. The classified document will be then appended into the belonging collection (s) and update the corresponding PAT tree(s). The updated PAT tree(s) can record the up-to-date information of the text collection(s), on which rigid linguistic information can be incrementally extracted.

Once an input document has been classified and indexed, the term extraction module will be performed. It will extract new keyterms from the document by estimating the completeness and significance of the composed strings with the corresponding PAT trees. The underlying technique for term extraction is an extension of the previous work [Chien'97, Chien'99]. The extended technique here emphasizes the incremental ability in new term extraction. Besides, in the similar term extraction module, it will find similar terms from the extracted keyterms. The basic concept for this processing is to extract keyterms with near context[Smadja'93]. To deal with the extraction of low-frequency keyterms especially on named entities such as personal names and organization names, a named entity extraction module is then developing. Since it is hard to extract low-frequency named entities simply based on the previously-developed PAT-tree-based approach, the proposed technique is tried to compare the contextual similarity between each named entity candidate and a set of extracted high reliable named entities.

This paper will only focus on the introduction of incremental term extraction and similar term extraction and brief description of named entity extraction. Further description about the corpus classification and named entity extraction can be referenced in [Chen'99, Chang'99].

## 4. Incremental Term Extraction

To many online NLP and IR systems such as voice browsers, web-based machine translation systems, Internet searching engines et al., it is cost-ineffective and even unrealistic to manually construct a domain-specific dictionary for each service domain. To capture up-to-date information and reduce unknown vocabulary, incremental extraction of domain-specific terms from online text resources are necessary.

This section is to define the considering problem and give an overview of the proposed tecnique for incremental extraction of domain-specific terms. A *domain-specific term* is defined as a string that consists of more than one successive characters in Chinese (or words in English) which has certain occurrences and is specific to a text collection with a distinct subject domain. Such a string has a complete meaning and lexical boundaries in semantics; it might be a word, compound word, phrase or linguistic template.

### 4-1 Overview of The Proposed Method

***Definition 1: The Incremental Term Extraction Problem***

Given a new document $D$, a set of incremental and domain-specific text collections $C_{1\sim n}$ and corresponding term lexicons $T_{1\sim n}$, the goal of this problem is to determine the most promising collection $C_i$ for $D$, extract *new terms* $X$ from $D$ and add to $T_i$, where $X = \{ x | x$ occurs in $D$; $x$ can be a domain-specific term of $C_i$ but missed in $T_i$ at present$\}$.

The above problem is defined to deal with the extraction of domain-specific terms with the increase of an online resource. The online resource will be divided into different text collections with specific subject domains in advance. Once a term is found specific and becoming important in a certain text collection, it is pursued can be extracted as soon as possible. Such a domain-specific term often indicates the occurrences of a certain event. If it can be identified immediately, some kinds of real-time reaction and information services like event detection of online news service can be implemented.

In fact, considering the reliability of term extraction, the extracted terms should have a certain occurrence and is expected will be used in a period of time, although some of them may not be used in a long term. So as to, a term which is a keyword in a single document but rarely occurs in other documents is not considered as a domain-specific term. Many low-frequency proper names are not taken into consideration in this way.

The proposed technique is known as one of a few works considering such an incremental extraction problem. To deal with the problem, several difficulties need to be faced with. *The first difficulty is to identify new and meaningful terms with document inputs as soon as possible.* It is known that to extract meaningful terms in an automatic way is still a challenging problem in western languages, but it is more critical in Chinese and oriental language processing because of difficulties in word segmentation and unknown word identification[Wu'95]. Our idea to this difficulty is to develop an efficient algorithm which is able to monitor the frequency change and usage freedom of each candidate term in the text collections, with the input of the new documents. *The second difficulty is how to estimate the significance of the candidate terms.* In our solution each new document should be classified into corresponding text collection(s) and its composed candidate terms will be checked by observing their distributions in different collections in the system. The candidates which are "non-specific" will be removed. Moreover, *the third difficulty is the efficiency in handling large and dynamic texts.* Since real-time processing is required in many applications, the utilized techniques have to be efficient in execution. To reduce the difficulty, the PAT-tree-based working structure is adopted again.

The term extraction module, as shown in Fig. 2, consists of two elementary sub-modules: completeness analysis and significance analysis. The outputs obtained with the proposed technique will contain the classified text collections, the PAT-tree indices and the domain-specific term lexicons from online text resources.Because the words in Chinese are not well-defined anyway, in this technique all the character strings of any length in the texts are first taken as candidates of keyterms.

### (1). *Completeness Analysis*

The first step is to extract new complete terms from each examining document. Like the completeness analysis step of the previous approach, this step is mainly to check if the strings of candidate terms are complete in lexical boundaries. But in difference, the strings need to be checked here are only that occurred in the new document *D* which have certain occurrences in the corresponding text collection $T_i$ but not found in the term lexicon $K_i$ at present. For each string *X* in *D*, it will judge if *X* is complete in semantic by its distribution and context in the updated PAT tree $I_i$. *X* is defined as complete in semantic iff its *association norm* of the composed sub-strings is strong enough and has no *left and right context dependency.* The estimations defined below are the same with the previous work. Such a design really considers the characteristics of Chinese.

***Definition 2: The association norm estimation***

The association norm estimation *MI(X)* for each string *X* is defined below:

$$MI(X) = \frac{f(X)}{f(X_s) + f(X_e) - f(X)}$$

Where *MI(X)* is the mutual information of a target string *X*, $X_s$ is the longest starting sub-string of *X*, i.e., the sub-string which is exactly *X* except that the last character of *X* is deleted, $X_e$ is the longest ending sub-string of *X*, i.e., the sub-string which is exactly *X* except that the first character of *X* is deleted, and *f(X)*, *f(X*$_s$*)*, *f(X*$_e$*)*, are the frequency counts of *X*, $X_s$, and $X_e$, in the text collection respectively. Such a definition is based on the efficiency of calculation in real-time applications. Character stings with the above mutual information below a threshold are considered to be incomplete.

***Definition 3: Left Context Dependency (LCD)***

Each string *X* has left context dependency if $|L| < t1$ or $\mathrm{MAX}_{\alpha}\, f(\alpha X)/f(X) > t2$, where *t1, t2* are threshold values, *f(.)* is frequency, *L* is the set of left adjacent strings of *X*, $\alpha \in L$ and $|L|$ means the number of unique left adjacent strings.

***Definition 4: Right Context Dependency (RCD)***

Each string *X* has right context dependency if $|R| < t1$ or $\mathrm{MAX}_\beta\, f(X\beta)/f(X) > t2$, where *t1, t2* are threshold values, *f(.)* is frequency, *R* is the set of right adjacent strings of *X*, $\beta \in R$ and $|R|$ means the number of unique right adjacent strings. The stings with either left or right context dependency are considered to be incomplete.

In fact, the above metrics are actually used to check if *X* contains highly-associated composed strings and also has complete lexical boundaries, by judging the usage freedom of *X* according to its contextual information. The basic assumption is that if *X* has few unique left or right adjacent strings, or if it frequently occurs together with certain adjacent strings, it might be incomplete in semantics.

The above estimations are easy to be implemented using the PAT-tree indices [Gonnet'92]. To know if a candidate string in *D* is complete or not, it just needs to check its association norm of the composed sub-strings as well as left and right context dependency. All of the operations can be easily done with PAT-tree access.

**(2).** ***Significance Analysis***

The second step is to find out domain-specific new terms. Like in the previous approach, the significance analysis step is to extract specific and significant candidate strings as the domain-specific terms. Using the following procedures, all of the remaining candidates strings will be checked using a common-word lexicon, a set of lexical rules and the analysis of the significance estimation function *S*(*Y*) shown below[ Schutze'98]. If a candidate string appears either in the common-word lexicon or can be formed using the lexical rules, it is treated as a non-significant candidate and is removed. The remaining candidates will be further checked by observing their frequencies and distributions between the corresponding and different PAT trees in the system. The candidates which are also frequently appear in the different PAT trees are treated as non-specific and are removed too. The strings which satisfy the estimation (larger than a threshold value) will be selected as the new domain-specific terms.

***Definition 5: The Significance Estimation Function***

$S(Y) = (f_i(Y)/f(T_i))/\ (f_g(Y)/f(T_g))$, where Y is a candidate term, $f_i\ (Y)$ is the frequency of $Y$ in collection $T_i$, $f(T_i)$ is total number of strings in collection $T_i$, $f_g(Y)$ is the frequency of $Y$ in the general collection, and $f(T_g)$ is total number of strings in the general collection.

The above estimation compares the relative frequency in the text collection of interest with the relative frequency in a reference collection. The necessary parameters are all easy to be computed with the PAT-tree indices. Among them, $f_i(Y)$ and $f(T_i)$ can be obtained directly in PAT tree $I_i$. As to $f_g(Y)$ and $f(T_g)$ can be obtained by summing up all of the domain-specific PAT trees in the system.

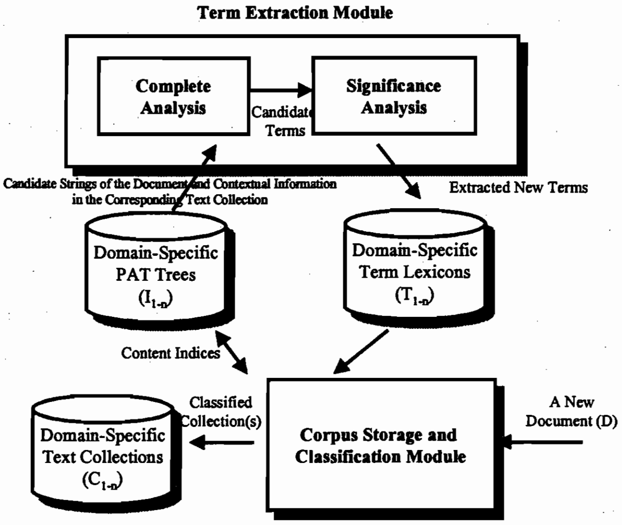
Term Extraction Module
Complete Analysis
Candidate Terms
Significance Analysis
Candidate Strings of the Document and Contextual Information in the Corresponding Text Collection
Extracted New Terms
Domain-Specific PAT Trees ($I_{1-n}$)
Domain-Specific Term Lexicons ($T_{1-n}$)
Content Indices
Classified Collection(s)
A New Document (D)
Domain-Specific Text Collections ($C_{1-n}$)
Corpus Storage and Classification Module

Fig. 2 An abstract diagram showing the proposed method for incremental term extraction.

### 4-2 Experimental Results

An experiment was performed to realize the effectiveness of the proposed approach for incremental term extraction. The experiment used a Chinese online-news database from Central News Agency (CNA) in Taiwan as the testing platform. At first, a total of 1,872 political news abstracts published in July 1997 were tested. The testing database contained 5-months manually-classified documents and one-month automatically-classified documents at that stage. In this experiment, the 1,872 news documents were added in sequence for both corpus classification and term extraction. Only the new terms extracted from the politics collection were counted. Tables 1 and 2 show the obtained results. It has to point out that before the processing of term extraction the political text collection has contained 6-month of news documents, in which the sixth-month documents were automatically classified and have only 45.1% precision and 99.4% top2 recall. Meanwhile, the corresponding term lexicon is empty in the initial stage.

Table 1 shows the obtained recall and precision rates with different threshold values in the significance analysis. The correct domain-specific terms of the testing 1,872 documents were extracted manually in advance. The terms extracted with different threshold values were compared with the correct set. It can be found the best performance in terms of both recall and precision rates was that using the threshold value 2. In that case, 1,135 correct terms can be extracted and the obtained precision and recall rates were 0.78 and 0.44 respectively. Such a performance is satisfied in many applications. It is worthy to note that 258 of the extracted terms were not included in the KUH dictionary, the largest Chinese dictionary we can find, which contains more than 160,000 word entries, and is believed covers many of terminological vocabulary used in news papers.

Except the above effectiveness issues, there are other issues such as the average number of document inputs to find a new term, the average frequency as the new terms to be extracted, and how often the extracted terms can be used, etc. For this reason, Table 2 shows the detailed results with the threshold value larger than 2. It is noted that in the table term length" is the number of characters of extracted terms. Since terms with different lengths behave differently (for example three-character terms are very often personal names, and four-character or longer terms are very often compound words), the results are shown with the term length as a special parameter. From this table, it can be observed that on average every 2.41 document inputs can find new terms. Also, each extracted new terms occur 28.95 times in the one-month testing documents and was extracted at the 9.25 time on average. This indicates most of the extracted terms are not late to be extracted and many real-time reactions can be performed.

| S(Y) | Total Extracted Terms(A) | No. of Correct Terms Extracted(B) | No. of Correct Terms Outside Dictionary(C) | Precision (B/A) | Recall |
|---|---|---|---|---|---|
| >1.5 | 2,291 | 1,374 | 297 | 0.60 | 0.53 |
| >2 | 1,455 | 1,135 | 258 | 0.78 | 0.44 |
| >2.5 | 723 | 593 | 172 | 0.82 | 0.23 |
| >3 | 214 | 184 | 66 | 0.86 | 0.07 |

Table 1. The testing results for incremental term extraction with different threshold values in the significance analysis which were obtained from a total of 1,872 political news abstracts published in July, 1997.

The proposed approach has been tested extensively and found very efficient in extracting terms from online text collections. For example, as shown in Table 3 there were more than ten thousand political terms can be extracted from a total of 13,849 political news abstracts published from Aug. to Dec. in 1997. The obtained results were found similar to that extracted from one-month news abstracts. With the increase of the news documents, the frequency values of the extracted terms are obviously increased but the frequency the terms can be extracted are similar.

| Term length (character N-gram) | Number of extracted new terms | Number of documents with new terms extracted | Average number of document inputs can find new terms (A) | Average frequency of the extracted new terms | Average frequency as the term can be extracted |
|---|---|---|---|---|---|
| 2 | 776 | 515 | 3.93 | 34.22 | 9.37 |
| 3 | 416 | 325 | 6.04 | 24.60 | 9.09 |
| 4 | 171 | 157 | 12.16 | 19.22 | 8.97 |
| 5 | 51 | 49 | 37.28 | 20.35 | 9.18 |
| 6 | 17 | 17 | 109.81 | 27.00 | 8.65 |
| 7 | 15 | 15 | 123.60 | 27.40 | 11.20 |
| 8 | 6 | 6 | 274.67 | 13.00 | 9.83 |
| 9 | 3 | 3 | 205.67 | 18.00 | 11.33 |
| Total N-grams | 1,455 | 814 | 2.41 | 28.95 | 9.25 |

Table 2. The detailed results for incremental term extraction with the threshold value larger than 2 in the significance analysis, which were obtained from a total of 1,872 political news abstracts published in July 1997.

| Term length (character N-gram) | Number of extracted new terms | Number of documents with new terms extracted | Average number of document inputs can find new terms (A) | Average frequency of the extracted new terms | Average frequency as the term can be extracted |
|---|---|---|---|---|---|
| 2 | 3,376 | 2,502 | 5.75 | 72.12 | 11.41 |
| 3 | 4,274 | 3,056 | 4.69 | 31.15 | 9.51 |
| 4 | 2,408 | 2,021 | 7.10 | 22.23 | 9.17 |
| 5 | 694 | 642 | 21.89 | 25.02 | 9.40 |
| 6 | 303 | 295 | 47.20 | 26.25 | 10.29 |
| 7 | 145 | 145 | 95.17 | 33.23 | 14.59 |
| 8 | 87 | 87 | 156.90 | 25.08 | 11.86 |
| 9 | 52 | 51 | 265.65 | 24.33 | 13.54 |
| Total N-grams | 11,339 | 6,242 | 2.28 | 40.90 | 10.12 |

Table 3: The detailed results for incremental term extraction with the threshold value larger than 2 in the significance analysis, which were obtained from a total of 13,849 political news abstracts published from Aug. to Dec. in 1997.

However, there exist some difficulties to be discussed. Taking all the terms with different lengths into account, it can be found that the precision rate for three-character terms was relatively low because many frequently used single-character words and two-character words are easily combined to produce three-character terms which are not necessarily key elements for most IR and NLP applications. Close examination of the extracted terms indicates that most of them are domain-specific such as proper nouns and topic terms, which are often very important in IR applications. This phenomenon is especially significant for terms with three or more characters. While it is important to indicate that the proposed approach is weak in extracting low-frequency terms, because the extracted terms should at least occur 9.25 times and 10.12 times as in Tables 2 and 3 respectively. To deal with the extraction of low-frequency but domain-specific terms, we are considering the combination of linguistic analysis methods as that developed in the named entity extraction module.

## 5. Low-frequency Named Entity Extraction

For those low-frequency terms, it is hard to judge if these terms own complete word boundaries based on statistical information, because the possible patterns in their context are very limited to be investigated. To deal with the extraction of low-frequency but significant keyterms , we present another method to perform semantic completeness analysis. As found in our experiments, the presented method can be used to handle the extraction of low-frequency named entities.

Names are some symbols that represent some characters or organizations and are conventionally used to identify the named entity. Named entity extraction (NE) is to identify all named locations, named persons, named organizations, dates, monetary amounts, and percentages in text. There are several features with the named entity. First, each type of named entity owns separate rule sets. In Chinese NE, family names predict personal names quite well. Former researches dealt with NE problem with rule-based heuristic approaches. Second, each named entity plays some specific roles. This situation can be revealed by neighboring contextual conventions of the named entity. Taking Chinese news articles for example, most of the personal names appear with a title in the context to indicate the identification or profession of the person. Such context not only gives information about the character that the named entity represents, but also helps to identify more named entities.

Since named entities usually have templates in the context and can be modeled by other named entities in the same category, a preliminary method and initial experiment were therefore developed. The basic idea of the method is described as follows:

**Context Dependency Estimation:**

Assume a named entity x belong to class K. Context of all named entities in K can be therefore used, if L(x) similar to L(K) or R(x) similar to R(K), where L(.) is the left context and R(.) is the associated right context respectively. A three-step process of context semantic learning is designed as below.

Step 1: Given an initial named entity set K and corpus D, the first step is to generate L(K) and R(K) based on K and D.
Step 2: For each "possible" new named entity x, add x into K if P(x) > Th or (Tl < P(x) <Th and L(x) c L(K) or R(x) c R(K)), where P(x) is a trained Markovian probability function, Tl, Th are two predefined threshold values, L(x) c L(K) means that L(x) belongs to L(x), and R(x) c R(K) that belongs to R(x).
Step 3: Extend L(K), R(K) by L(x), R(x) and repeat Step 2 until the K set cannot be increased obviously.

We have done a small scale of experiments on Chinese personal named entity extraction based on the above method. The testing data size is 1.65MB of news documents, and the initial Markovian probability of personal names is based on order-one Markov model and Sinica corpus. The obtained recall and precision rates with the change of threshold values have been obtained and shown in Table 4. It is can be easily to see that based on context information the extraction accuracy can be improved.

| | Probability based (baseline) | With Context Estimation (weight 0.06) | With Context Estimation (weight 0.04) | With Context Estimation (weight 0.02) | With Context Estimation (weight 0.005) | With Context Estimation (weight 0.003) | With Context Estimation (weight 0.001) |
|---|---|---|---|---|---|---|---|
| Corrected names extracted | 7,768 | 8,608 | 8,608 | 8,632 | 8,778 | 9,067 | 9,073 |
| Error names Extracted | 1,123 | 1,188 | 1,193 | 1,524 | 2,423 | 3,351 | 5,074 |
| Recall rate | 0.848 | 0.917 | **0.940** | 0.942 | 0.958 | 0.990 | 0.990 |
| Precision rate | 0.873 | 0.876 | **0.878** | 0.849 | 0.783 | 0.730 | 0.641 |
| 9157 names to be extracted from 1,876 news abstracts | | | | | | | |

Table 4: The obtained results for personal named entity extraction.

## 6. Similar Term Extraction

Automatic construction of a thesaurus from online text resources is important but a challenging research topic. A thesaurus is a set of items ( phrases or words ) plus a set of relations between these items [Jing'94] . Some researchers have used head-modifier relationships or descriptions of entities to determine similar words[Strzalkowski'95][Radev'98][Lin'98]. Others make use of lexical occurrence information to build related words[Jing'94][Crouch'92][Schutze'97]. Our research towards this topic is just in the beginning. The first step we would like to try is to extract similar terms from the set of domain-specific terms extracted based on the above term extraction methods.

Since it can extract a number of domain-specific erms which were excluded in general dictionary right now, it seems to be possible to deal with the similarity and association among these extracted terms. According to the demand of different computer processing, we simply divide the similar terms into three categories:

(1) Abbreviation : (中央研究院，中研院)

(2) Named entity with associated title or description : (李登輝總統，總統李登輝，李登輝), or (網球名將張德培，張德培)

(3) Terms different in content but similar in concept : 資訊，電腦，計算機

The first two types of similar terms are that similar in content, i.e., sharing common composed character strings among similar terms. The third type of similar terms has no obvious common sub-strings. Since it is more difficult to extract the third type of similar terms, in the beginning stage we just investigate the extraction of the first two types.

## 6-1 The Proposed Method

### (1) Similarity Measurement

The proposed method is based on an assumption that similar terms frequently co-occur in the same documents. There are several ways to measure the correlation of two terms. The Dice coefficient as defined below was found more effective and therefore adopted:

$$\textbf{Dice}\ (k1,k2) = 2f_{k1k2}/(f_{k1}+f_{k2}),$$

where $f_{k1}$, $f_{k2}$ and $f_{k1k2}$ are the numbers of document occurring ***k1*,*k2*** and both ***k1*** and ***k2*** together , respectively.

### (2) The Extraction Algorithm

The extraction algorithm used is very simple as shown below:

```
1. Term Extraction:
   1.1 Use the above PAT-Tree-Based and named entity extraction methods to extract keyterms.
2. Estimation of Similar Terms:
   2.1 choose any two keyterms k1, k2
   2.2 if k1 is substring of k2, then compute Dice(k1,k2)
   2.3 if Dice(k1,k2) > t1, where t1 is the threshold,
         then k1 and k2 is a pair of similar terms
```

## 6-2 Experimental Results

The first experiment is to test the accuracy of the similarity estimation. A total of 466KB CNA news articles related to the judiciary and transport subject domains were tested. Some of the experimental results are shown in Fig.1, where the horizontal axis indicates variation of different t1 values, and vertical axis indicates the corresponding ratios of recall and precision with the change of different t1 values. The results show that the precision can be high, if t1 is set at 0.5.

The second experiment is to test the accuracy using different sets and sizes of news articles. The test news were grouped manually into four sets, namely CNA11:congress/politics 1996, CNA12:congress/politics 1998, CNA21:judiciary/transport 1996, CNA22: judiciary/transport 1998. The results are shown in Table 5. It can be found that the average precision rate of 73.75% (112.5/153.25) can be achieved. Appendix 1 and 2 show some samples of extracted similar terms.

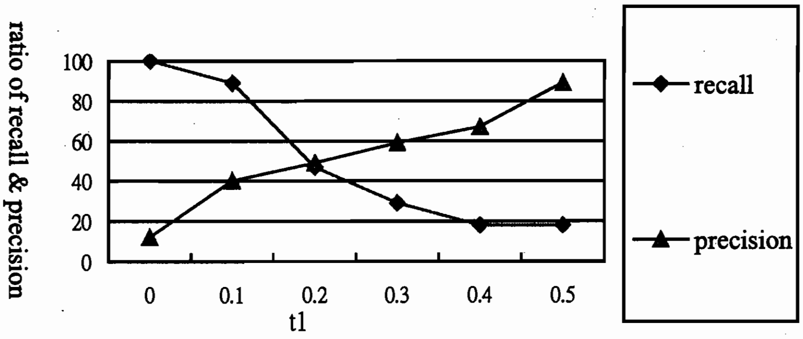

Fig.3 The ratios of recall and precision with different t1

| | Text Size(MB) | Total Number of Similar Term Pairs | Number of Correct Extracted Similar Term Pairs | Obtained Precision |
|---|---|---|---|---|
| CNA11 | 11.34 | 329 | 238 | 72% |
| CNA12 | 3.84 | 153 | 115 | 75% |
| CNA21 | 5.21 | 66 | 55 | 83% |
| CNA22 | 2.13 | 65 | 42 | 65% |
| average | 5.63 | 153.25 | 112.5 | 73.75% |

Table 5. Obtained results of the Similar Term Extraction Experiment.

## 7. Conclusion

In this paper an ongoing research called Live Dictionary Construction has been introduced. Such research effort has been integrated with a number of techniques. This paper focuses on the introduction of incremental term extraction and similar term extraction. Preliminary experimental results show that th it is very promising build a dynamic dictionary with IR systems.

## References

1. [Chang'99] Yuan-Lu Chang , Chun-Liang Chen and Lee-Feng Chien (1999). An Integrative Approach for Chinese Named Entity Extraction, Paper in preparation.
2. [Chen'98] Chen, Chun-Liang (1998). PAT-tree-based Natural Language Processing and Applications under Internet Environment. Master Thesis, Dept. of CS&IE, National Taiwan University,.
3. [Chien'97] Chien, Lee-Feng (1997) *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval*, Proceedings of ACM SIGIR'97, Philadelphia, USA, pp. 50-58.
4. [Chien'99] Chien, Lee-Feng (1999) *PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval,*, to appear on Information Processing and Management , Elsevier Press.
5. [Crouch'92] Crouch C. J. and Yang, B. (1992). Experiments in automatic statistical thesaurus construction, Proceedings of SIGIR'92.
6. [Gonnet'92] Gonnet, G. H., Baeza-yates, R. et al. (1992) *New Indices for Text: Pat Trees and Pat Arrays*. Information Retrieval Data Structures & Algorithms, pp. 66-82, Prentice Hall.
7. [Jing'94] Yufeng Jing and W. Bruce Croft (1994). An Association Thesaurus for Information Retrieval", UMass Technical Report 94-17. 1994
8. [Lewis'96] Lewis, David D. and Sparck Jones, Karen (1996) *Natural Language Processing for Information Retrieval*, Communications of the ACM, Vol. 39, No. 1, Jan. 1996, pp. 92-101.
9. [Lin'98] Dekang Lin (1998) Automatic Retrieval and Clustering of Similar Words, COLING'98.
10. [Radev'98] Dragomir R. Radev (1998) Learning Correlation between Linguistic Indicators and Semantic Constraints: Reuse of context-Dependent Descriptions of Entities, COLING'98.
11. [Schutze'97] Hinrich Schutze and Jan O. Pedersen, (1997) A Coocurrence-based Thesaurus and Two Applications to Information Retrieval", Information Processing & Management, Vol. 33, No. 3, pp. 307-318,1997.
12. [Schutze'98] Schutze, Hinrich (1998) *The Hypertext Concordance: A Better Back-of-the-Book Index,* Proceedings of the First Workshop on Computational Terminology (Computerm'98), pp. 101-104.
13. [Smadja'93] Smadja, F., (1993) *Retrieving Collocations from Text: Xtract*, Computational Linguistics, 19 (1), pp. 143-177.
14. [Wan'97] Wan, T. L., Evens, M. et al. (1997) *Experiments with Automatic Indexing and a Relational Thesaurus in a Chinese Information Retrieval System*, Journal of the American Society for Information Science,48(12), pp. 1068-1096.
15. [Wu'95] Wu, Z., Tseng, G. (1995) *ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval*. Journal of the American Society for Information Science, 46 (2), pp. 83-96.
16. [Zernik'91] Zernik, Uri (1991) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, Publishers.

### Appendix 1. Some Samples with Similar Terms from CNA news

@19980113 12:07:30 **蕭萬長**說明選擇訪問菲律賓的原因
（中央社記者梁君棣台北十三日電）**行政院長蕭萬長**今 天說明他這次訪問菲律賓的理由，主要是拜訪總部設在菲律賓的**亞洲開發銀行**，**亞銀**有五十六個會員國，中華民國也是**亞銀**的會員國之一，到**亞銀**訪問，可以進一步了解整個局勢。

==============================================================

@19980113 11:53:44 **蕭萬長**結束訪菲返國舉行記者說明
（中央社記者梁君棣台北十三日電）**行政院長蕭萬長**今天早上從菲律……..
第一、**亞銀**對當前亞洲金融風暴有見解，與我國的了解相當接近，差不多…。
第二、**亞銀**對我經濟發展表示肯定，並對這次我國處理金融風暴的政策…。
第三、**亞銀**對由行政院經濟建設委員會主任委員江丙坤所率訪問團的行動…。
**蕭萬長**說，在與相關人士討論時曾提到中華民國如何參加區域金融合作機制，亞銀表示將支持我國參與。

==============================================================

@19980103 12:41:06 多數**陸委會**諮詢委員指人民幣短期內會撐下去
（中央社記者許雅靜台北三日電）**行政院大陸委員會**今天召開諮詢委員會議，**陸委會**企劃處處長詹志宏表示，多數諮詢委員對大陸今年經濟不樂觀，但認爲短期內人民幣仍會撐下去，整體而言，人民幣如果貶值，對我經濟影響不大。

==========================================================

@19980116 15:34:26 **海基會**願協助河北震災救援我希望**海協會**回應
（中央社記者曾淳良台北十六日電）大陸河北張家口一月十日發生強烈地震，至少造成五十人死亡，一萬多人受傷，整個救災工作仍在極惡劣環境下進行，我方基於人道考量，主動表示願配合災區協助工作，陸委會希望 大陸海協會作出回應。

==========================================================

@19980119 18:34:45 **陸委會**指張京育未安排與汪道涵在東京會面
（中央社記者曾淳良台北十九日電）**行政院大陸委員會**主任委員張京育正在東京訪問，**大陸海峽兩岸關係協會**會長汪道涵目前也在東京，**陸委會**副主委兼發言人許柯生今天表示，張京育此行主要是參加學術研討會，並未安排與汪道涵會面。

## Appendix 2. Samples of Extracted Similar Terms

- Correct abbreviations:

世界自由民主聯盟=>世盟 [0.695652]
北大西洋公約組織=>北約 [0.516129]
金門防衛司令部=>金防部 [0.642857]
中央選舉委員會=>中選會 [0.647059]
中山科學研究院=>中科院 [0.600000]
公共工程委員會=>工程會 [0.600000]
亞洲開發銀行=>亞銀 [0.666667]
中央研究院=>中研院 [0.555556]
國家安全局=>國安局 [0.631579]
李登輝總統=>李總統 [0.692105]
民主進步黨=>民進黨 (273)[0.546000]
違章建築=>違建 [0.545455]
台灣銀行=>台銀 [0.583333]
空軍總部=>空總 [0.577778]
社會福利=>社福 [0.547368]
歐洲聯盟=>歐盟 [0.784615]

- Correct description of entities

李元簇夫人徐曼雲=>徐曼雲 [0.666667]
台北市議員林瑞圖=>林瑞圖 [0.560000]
教宗若望保祿二世=>教宗 [0.523077]
俄羅斯總統葉爾勤=>葉爾勤 [0.533333]
古巴總統卡斯楚=>卡斯楚 [0.500000]
參謀總長羅本立=>羅本立 [0.695652]
立法委員王志雄=>王志雄 [0.590909]
交通部長蔡兆陽=>蔡兆陽 [0.660870]
高雄市長吳敦義=>吳敦義 [0.595745]
法務部長廖正豪=>廖正豪 [0.732394]
台灣省長宋楚瑜=>宋楚瑜 [0.617143]
台北市長陳水扁=>陳水扁 [0.566038]
印尼總統蘇哈托=>蘇哈托 [0.537313]
內政部長黃主文=>黃主文 [0.676471]
行政院長蕭萬長=>蕭萬長 [0.691814]
參謀總長唐飛=>唐飛 [0.500000]
先總統蔣公=>蔣公 [0.526316]
大使謝棟樑=>謝棟樑 [0.777778]
團長何明德=>何明德 [0.842105]
中非共和國=>中非 [0.705882]
二二八事件=>二二八 [0.666667]
副總統連戰=>連戰 [0.674121]
政變傳聞=>政變 [0.526316]
排雷工程=>排雷 [0.500000]
拜耳公司=>拜耳 [0.740741]
樞機主教=>主教 [0.510638]
赴法國=>赴法 [0.600000]
警義消=>義消 [0.666667]
智障者=>智障 [0.500000]
縣農會=>農會 [0.600000]
花蓮縣=>花蓮 [0.533333]

- Incorrect pairs:

性侵害防治委員會=>性侵害 [0.600000]
監察院糾正=>糾正 [0.625000]
海洋政策=>海洋 [0.533333]
拜耳案=>拜耳 [0.640000]