Statistical Acquisition of Terminology Dictionary*

Huang Xuan-jing , Wu Li-de , Wang Wen-xin
Dept. of Computer Science, Fudan University , 200433 Shanghai
960048@ms.fudan.sh.cn, ldwu@fudan.ihep.ac.cn

Abstract: Terminologies are specialized words and compound words used in a particular domain, such as computer science. Since they are very common in scientific articles, the ability to automatic identification of terminology could greatly assist any domain related natural language processing applications. Unfortunately, the collection of terminology information is very difficult and requires much tedious and time consuming manual work. In this paper, a semi-automatic approach is developed to extract technical words and phrases from on-line corpora. This approach can significantly reduce the manual effort in the generation of terminology dictionary. First, those domain specific words which have no entries in the universal dictionary are identified. Second, terminology words are extracted from these new words as well as the universal dictionary. Then compound words are extracted from the combination of terminology words and other words. The final computer terminology dictionary contains 1,034 words and 3,471 compound words. Experiment shows that 89.5 percent of all the occurrences of computer terminology can be identified with this terminology dictionary.

keyword: ***Chi-square Test, Automatic Indexing, Mutual Information***

1. Introduction

Terminologies are specialized words and compound words used in a particular domain, such as computer science. They are extensively used in scientific articles. Previous research had shown that about 25% of the words in science abstract were technical words【6】. Therefore, the ability to automatic identification of terminology could greatly aid any domain related natural language processing applications, such as automatic indexing, information retrieval and document categorization. For example, automatic indexing is the foundation of many other relevant tasks. It needs to automatically identify those words which most appropriately reflect a text's theme. Since terminologies are highly relevant to the text's domain, they are proved to be much valuable index words. Even in more universal applications such as semantic analysis and translation, terminologies also play important roles, and therefore require special treatment.

Unfortunately, the identification of terminology is a hard work. Most terminologies don't have entries in universal dictionaries. In addition, terminology dictionaries are highly variable in the coverage. For example, computer science dictionaries' coverage of computer science terminology ranged from 24% to 66%【6】.

* This paper is supported by Chinese Natural Science Foundation and high technology 863 project.

With regard to Chinese, the identification procedure is even more difficult. First, there are scarcely any available machine readable Chinese dictionaries for specialized domains. Therefore, the generation of terminology dictionary would inevitably require a great deal of tedious and time consuming manual work. Second, in most Indo-European languages, even a word couldn't be found in the dictionary, it still could be separated by the spaces between it and neighboring words; however, Chinese is written in character sequences, with no delimiters between successive words. Hence the first step of Chinese information processing is necessarily to segment the character sequences into word sequences. The main knowledge base of segmentation is the dictionary. However, most of the terminologies couldn't be found in the dictionary. Therefore, before further processing, those domain specific words which are unavailable in the dictionary should be extracted and added to it. This procedure is called new word extraction.

Due to the availability of large scale on-line real text, corpus based natural language research has become one of the focuses of computational linguistics. Among all the corpus based researches, some of them are quite similar to the work reported here, including sublanguage vocabulary identification【6】, automatic suggestion of significant terminology【15】, identification and translation of technical terminology【3】, automatic extraction of terminology【4】. For example, Haas introduced a method for automatic identification of sublanguage vocabulary words. First, words that could be easily identified as belonging to the vocabulary of the given domain were extracted, then the rest of the vocabulary were extracted using these seed words.

Another relevant research is statistical collocation extraction. In fact, terminology phrase belongs to one certain kind of collocation — fixed collocation. whether two or more words can compose a collocation is measured by the correlation coefficient of these words【11】. If these words' correlation coefficient is large enough, they may probably make up a collocation. There are many statistical methods to calculate words' correlation coefficient, including co-occurrence frequency【10】, mutual information【1】,generalized likelihood estimation【5】, chi-square test【2】【7】, Dice coefficient【11】, etc.

There are also many valuable works in China, especially about the distinctive new word extraction of Chinese text. Wang Kai-zhu presented a statistical method to extract possible words from texts. Weights of possible words were calculated using their frequency and length information【13】. Zhang Shu-wu also presented a strategy which made use of co-occurrence frequencies to collect new words【14】. Pascale Fung extended a tool originally designed for extracting English compounds -- CXtract to collect new words in order to improve the segmentation precision【9】.

Due to the distinct characteristic of Chinese, there is still no systematic approach to generate practical and relatively complete Chinese terminology dictionaries from on-line corpora. In this paper, a semi-automatic approach is developed to extract technical words and phrases from corpora. This approach integrates such methods as new word collecting, terminology word extraction and terminology phrase generation. It can significantly reduce the manual effort in the generation of terminology dictionary. First, those domain specific words which can't be found in the universal dictionary are identified. Second, terminology words are extracted from these new

words as well as the universal dictionary. Then compound words which are combined by terminology words and other words are generated.

The following sections are organized as such: Section 2 introduces the identification of domain specific words; Section 3 describes how to extract terminology words from the universal dictionary; Section 4 presents the method for terminology phrase extraction; Section 5 provides detailed experimental results; The final section is the concluding remarks.

2. New Word Extraction

A Chinese word is usually composed of no more than 4 Chinese characters. Most of the words are uni-grams, bi-grams, tri-grams and 4-grams. Uni-grams only consist of one character, and most of them are common words and then can be found in universal dictionaries. The number of n-grams with n>4 is very small, and the occurrence of most of them is rare. Among the 9000 most frequently used words, far below 1% of them are longer than 4 characters【9】. In addition, most of these words are idioms or terminologies, then can be extracted in the phrase generation phase. Therefore, in this section, only bi-grams, tri-grams and 4-grams are taken into consideration.

Now consider two neighboring characters A and B. We call these two characters as a bi-gram candidate. They belong to either the same word, or two neighboring words. We can intuitively suppose that the two characters are more correlate to each other when they belong to the same word. Therefore, we may choose a statistic to measure the correlation coefficient of neighboring characters, then use this statistic to judge the probability that they belong to the same word.

The correlation coefficient could be measured by several methods, such as co-occurrence frequency, mutual information, generalized likelihood estimation, chi-square test, Dice coefficient. Among them, chi-square test needs special attention. First, it is closely related to the binomial distribution model of text. Second, the computation is quite simple. Experiment in section 5 also showed that it could lead to better performance. Following is the detailed description of this method.

Compare each bi-gram (A, B) candidate to every two neighboring characters (C_i, C_{i-1}) in the text sequence $C = (C_1C_2\cdots C_iC_{i-1}\cdots C_n)$, where n is the size of the text, and record the comparison results. Thus there are four types of results altogether:

Result 1: $C_i=A$ and $C_{i-1}=B$, which is noted as (A, B);

Result 2: $C_i=A$ and $C_{i-1}\neq B$, which is noted as $(A, \overline{B})$;

Result 3: $C_i\neq A$ and $C_{i-1}=B$, which is noted as $(\overline{A}, B)$;

Result 4: $C_i\neq A$ and $C_{i-1}\neq B$, which is noted as $(\overline{A}, \overline{B})$.

Let n be the count of (C_i, C_{i-1}), $n_{11}, n_{12}, n_{21}, n_{22}$ be the count of (A, B), $(A, \overline{B})$, $(\overline{A}, B)$, $(\overline{A}, \overline{B})$ respectively. Obviously, $n = n_{11} + n_{12} + n_{21} + n_{22}$.

Let $n_{i\cdot} = n_{i1} + n_{i2}$, $n_{\cdot j} = n_{1j} + n_{2j}$, $(i = 1, 2; j = 1, 2)$.

Then a contingency table is established as such:

Table 1: Contingency Table of Characters A and B

	B	$\overline{B}$	Σ
A	n_{11}	n_{12}	$n_{1\cdot}$
$\overline{A}$	n_{21}	n_{22}	$n_{2\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	n

If the characters A and B occur independently, then we would expect $P(AB)=P(A)\times P(B)$, where $P(AB)$ is the probability of A and B occurring next to each other; $P(A)$ is the probability of A, $P(B)$ is the probability of B. To test the null hypothesis $P(AB)=P(A)\times P(B)$, we compute the chi-square statistic:

$$\chi^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij}\times\frac{n_{i\cdot}\times n_{\cdot j}}{n}\right)^2}{n_{i\cdot}\times n_{\cdot j}};$$

The above equation can be simplified as: $\chi^2 = \frac{n\left(n_{11}\times n_{22}-n_{12}\times n_{21}\right)^2}{n_{1\cdot}\times n_{2\cdot}\times n_{\cdot 1}\times n_{\cdot 2}}$.

We define the correlation coefficient of characters A and B to be the value of chi-square test. Those bi-gram candidates with correlation coefficient smaller than a pre-defined threshold are considered to occur randomly and should be discarded. Others are sorted according to their correlation coefficient in descending order.

Tri-gram and 4-gram candidates are processed in the same way. To compute the correlation coefficient of all tri-grams, we shouldn't set the null hypothesis to $P(ABC)=P(A)\times P(B)\times P(C)$, otherwise we would be faced with the critical problem of data sparseness and then get unreliable and vulnerable results. In alternate, we just look a tri-gram as the combination of a bi-gram and a character, then calculate their correlation coefficient. Similarly, a 4-gram can be looked either as the combination of a tri-gram and a character, or two bi-grams.

The rest of bi-gram, tri-gram, 4-gram candidates constitute 3 separate tables. In these tables, many candidates are available in the universal dictionary, others are potential words. These potential words are carefully examined by skillful computer professionals, and many of them are accepted and then appended to the dictionary in order to improve segmentation precision. These words are called new words. Human intervention is still inevitable, since statistical methods not only generate useful, but also noisy words. Thresholds can be applied to limit this effect, but can't eliminate it.

Terminology Word Extraction

rminology words are divided into two subsets and treated respectively. Most of them have s in the universal dictionary. These words should be extracted from the new word tables. number of new words is limited, and most of new words are domain specific words inologies and proper names, this work is also done manually.

minologies are available in the universal dictionary. They are either frequently used

words, such as "计算机 (computer)" and "网络 (network)", or have meanings outside of science areas, such as "代理 (agent)" and " 过程 (procedure)". These words are also extracted in statistical method.

If a word is a terminology, then it probably occurs more often in related domain corpus than normal. Let $P_c(W)$ be the frequency of word W in domain corpus, $P_n(W)$ be the normal frequency of W. If $P_c(W)>>P_n(W)$, W is extracted and further examined by professionals, otherwise it is discarded. In the following experiment, this formula is replaced with $P_c(W) > T_2 \cdot P_n(W)$, where T_2 is a threshold. Similar method could be found in Zhou95【15】.

To gather all word frequency information in a specific domain, the domain corpus should be first segmented with the augmented dictionary. The normal frequency could be obtained either from a balanced on-line frequency dictionary or a universal corpus. Since on-line frequency dictionary is not available for us, another universal corpus is used. For those words which appear in the domain corpus, but don't appear in the universal corpus, P_n is approximately replaced with the average frequency of all words.

4. Terminology Phrase Generation

Terminology phrases are word pairs composed of terminology words and other words. Current research only concerns word pairs. Terminology phrases are generated in three steps.

At the first step, all the candidate phrases are extracted. The whole corpus is segmented with the augmented dictionary in advance. A small window is put over each terminology word appearing in the text sequence. Candidate terminology phrases are those word pairs which are composed of one terminology word and another word inside this terminology's border window. Those word pairs with too low frequencies are filtered out.

Whether a word pair is a phrase is measured by its weight. At the next step, most of candidates are also filtered out if their weights are too small. A word pair's weight is mainly decided by its correlation coefficient. In addition, two heuristic rules are adopted to modify the weights:

Rule 1: If a word pair is composed of two terminology words, its weight is strengthened.

Rule 2: If a word pair contains function words, it is also filtered out. A stop word table is introduced for this reason. This table contains more than 1000 Chinese function words, such as "的 (of)" and "是 (be)".

At the last step, all the remaining word pairs are manually examined. Those accepted phrases as well as terminologies words compose the final terminology dictionary.

5. Implementation and Results

Two corpora were chosen for this research. One is a Computer World corpus (CW). It is composed of all articles of the newspaper "Computer World (计算机世界)" from 1990 to 1994. The 100M bytes corpus contains more than 40M Chinese characters. The other is a universal corpus — XinHua news (新华社电讯稿) corpus (XN). It contains more than 8,000

news articles with 10M bytes of text.

CW corpus contains many computer terminologies, most of which just appeared in last two decades. Therefore, only a small number of them have entries in universal dictionaries. XN corpus also contains many new words, but the number is much smaller.

To collect new words, each article was scanned and all the bi-gram, tri-gram and 4-gram candidates with frequency greater than threshold T_1 were extracted (for CW corpus, T_1=4, for XN corpus, T_1=2). In addition, some shorter candidates were actually parts of longer ones, and couldn't exist independently. For example, every time "算机" was seen in the text, it followed "计"; every time "阿富" was seen, it was followed by "汗". So "算机" and "阿富" are only parts of longer candidates "计算机 (computer)" and "阿富汗 (Afghanistan)". Thus they should be removed from candidate tables.

The remaining candidates were sorted by their correlation coefficient in descending order. Those candidates on the top of the table have higher probability to be real words. To evaluate the computing methods, we may consider the distribution in the candidate table of those words available in the dictionary. These words are called as available words. Let D be the sorted candidate table, DS be a sub-table of D starting from the beginning of D. Two evaluation standards precision and recall were defined as follows:

Precision of DS = Number of available words in DS / Number of candidates in DS;

Recall of DS = Number of available words in DS / Number of all available words in D.

Obviously, since many new words have no entries in the dictionary, the real precision and recall should be somewhat higher. Figure 1 is the Recall-Precision curves of the bi-gram candidate table of CW corpus. Figure 2 is those of XN corpus.

Figure 1: Computer World Corpus

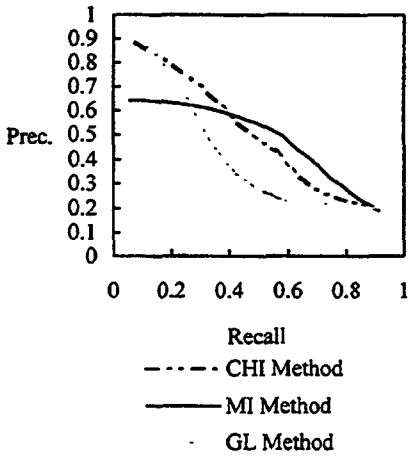
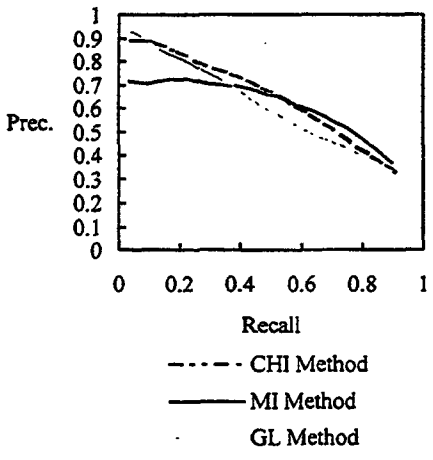


Figure 2: XinHua News Corpus



Three computing method were used: mutual information (MI) 【1】, generalized likelihood estimation (GL) 【5】 and chi-square test (CHI). From these figures we can see that the performance of GL method is the worst. When recall is not much high (less than 40-50%) , which means that only those top candidates are considered, CHI method is the best. When recall

becomes higher, MI is better than others. Since only top of the table should be further examined manually, CHI method was chosen.

Figure 3 demonstrates the Recall-Precision curves of two corpora using CHI method. Although XN corpus is only one tenth of CW in size, it gains better results. This result can be attributed to the fact that XN corpus contains less new words.

There are more than 400,000 bi-gram candidates in CW corpus. Among them, 17,779 are available words. Only 61,584 candidates have frequencies greater than T_1(T_1=4), including 7,089 available words. These candidates compose the bi-gram candidate table. New words are extracted from the top 16% of this table. Among these 9,856 high-rank candidates, 4,041 are available in the dictionary, which amount to 57% of all the available words in the whole table. The remaining 5,815 were potential new words and then further examined by computer professionals. Finally, 1,699 were accepted. Similar results were obtained from tri-gram and 4-gram candidates. A little more differently, the proportion of available words in tri-gram and 4-gram candidate tables is much smaller than in bi-gram table. Therefore, new words were only extracted from the top 4% tri-grams and the top 2% 4-grams. The quantities of accepted tri-grams and 4-grams is also smaller than that of bi-grams. Table 2 presents the vocabulary distribution of CW corpus. Among the whole vocabulary, more than 10% are extracted new words. Later the recall and precision were recalculated using the augmented dictionary. Figure 4 demonstrates the Recall-Precision curves of Computer World corpus using original dictionary and augmented dictionary respectively. We can find that the precision is significantly improved after new words were appended.

Figure 3: Comparison between XN and CW Corpus

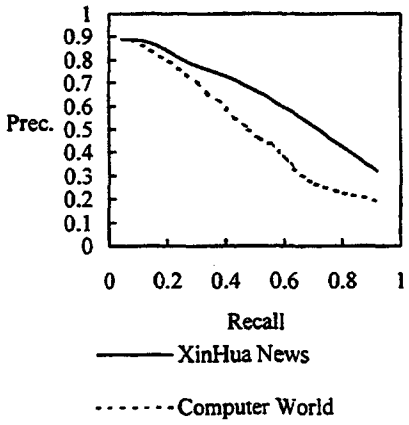


Figure 4: Comparison between Original and Augemented Dictionary

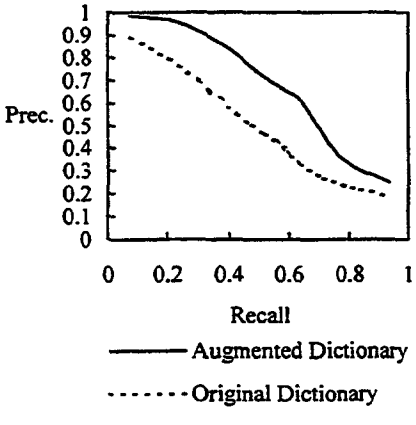


Table 2: the Vocabulary Distribution of Computer World Corpus

	Uni-gram	bi-gram	tri-gram	4-gram	Total
Available Words	3298	17779	1830	2370	25277
New Words		1699	1122	49	2870
Sum	3298	19478	2952	2419	28147

To extract terminology words from new words, all new words were manually examined and put to any of three categories: terminology words, proper names and other domain specific words, or to say, those words which are related to this domain to some degree, but cannot be considered as terminology of this domain, for example: 闭路电视 (cable TV) and computer domain. Table 3 shows the distribution of new words. Table 4 presents some example words with highest correlation coefficient. From table 3 and table 4 we can see, about one fourth of new words are terminology words; another one fourth are proper names; the rest are other domain specific words. Those words with highest correlation coefficient are almost terminology words and proper names. In addition, many tri-grams are proper names, because most of Chinese names are composed of 3 characters. Since Chinese name recognition is also an complex problem in Chinese real text processing, this method can also be utilized to recognize names.

Table 3: the Distribution of New Words

	terminology	proper names	others	total
bi-gram	389	215	1095	1699
tri-gram	302	503	317	1122
4-gram	20	8	21	49
all	711	726	1433	2870

Table 4 : Examples of New Words

	Examples
bi-gram	病毒 (virus) 蜂窝 (honeycomb) 瓶颈 (bottleneck) 共享 (share) 东芝 (Toshiba) 媒体(media) 便携 (portable) 扇区 (sector) 接口 (interface)
tri-gram	澧河泾 (place) 屠友灿 (name) 胡苏泰 (name) 俄勒冈 (Oregon) 砷化镓(chemical compound) 工作站 (work station) 数据库 (database)
4-gram	巴塞罗那 (Barcelona) 雀尼威尔 (Honeywell) 马尔可夫 (Markov) 及物动词 (Vt.) 自底向上 (bottom up) 闭路电视 (cable TV)

To extract terminologies from the original universal dictionary, the frequency of each of the 25,277 words in CW corpus was compared to the frequency in XN corpus. The threshold of T_2 was set to 3. only 1,938 words' frequencies in CW corpus were three times higher than in XN and then satisfied this threshold limitation. These words were further categorized manually. The categorization results are demonstrated in table 5.

Table 5: Manual Examination results of the Universal Dictionary

	terminology	others	total
bi-gram	287	1427	1714
tri-gram	33	155	188
4-gram	4	32	36
all	323	1615	1938

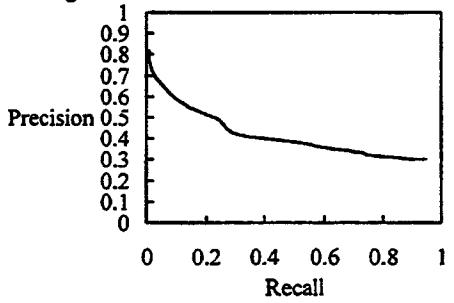
We can find terminologies extracted from the universal dictionary are much fewer than those extracted from new words: of the 1,938 words, only 323 were accepted finally. In addition, to make sure only a small portion of terminology words had been missed, 1,000 words were randomly selected from the rest 23,329 words and only 4 were found to be terminologies. This helped to explained that most of the terminology words in the universal dictionary had been extracted.

Terminology phrases were later extracted from the combination of 1,034 terminology words and their neighboring words within a distance of ±3. There are altogether 35,178 phrase candidates with frequency greater than a threshold T_3 (here T_3=3) . Random sampling showed that 30% of them are acceptable terminology phrases. These candidates' weights were computed in the method introduced in section 4. Then they were sorted in descending weight order. Figure 5 shows the approximate recall-precision curve of terminology phrase extraction. The reason for approximate evaluation was that it was impossible to manually examine all 35,178 terminology phrases, therefore only randomly selected 3,000 candidates were examined. From figure 5, we can find that the performance of phrase extraction wasn't as good as that of word extraction. This phenomenon can be explained by the fact that some highly associated candidates still couldn't compose terminology phrases. Most of these pseudo phrases can be divided into two classes:

Class 1: The two words compose a Verb-Object, Subject-Verb, or other phrases. For example, "左键 (left mouse key) 拖动 (drag)".

Class 2: The two words are two highly associate words, but have no direct syntactic relations. For example, "全拼 简拼" (two Chinese character input methods). In fact, similar phenomena can also be found in English 【8】.Therefore, the precision will surely be improved when syntactic information is used to further filter candidates.

Figure 5: Recall-Precision Curve of Phrase Table



Terminology phrases were extracted from the top 20% (with precision of about 50%) terminology phrase candidates. these candidates were examined manually and 3,471 were accepted. These 3,471 phrases as well as the 1,034 words compose our computer terminology dictionary. Table 6 presents some example terminologies with high rank.

100 pieces of article of 72K bytes were randomly selected to test the coverage of this terminology dictionary. A simple automatic pattern matching program was used to identify terminologies and 1,174 occurrences of terminologies were spotted. This identification procedure was also done by several graduate students major in computer science. The automatic recognition

results were compared to the union set of three experimenters. 89.5% of all terminologies found by experimenters were also found by the program. And 73.9% of all the program output was judged to be correct. The relatively lower precision can be attributed to the fact that some terminologies, especially those available in the original dictionary, have meaning outside computer domain. In large scale natural language processing applications where context information and local parsing are available, the precision would be increased certainly.

Table 6: the Distribution of Terminology

	Number	Example
available words	323	软件（software）并行（concurrent）程序（program）计算机（computer）二进制（binary）机器翻译（machine translation）
bi-gram	389	病毒（virus） 瓶颈（bottleneck） 共享（share） 媒体(media） 便携（portable） 扇区（sector） 接口（interface） 视频（video）
tri-gram	302	工作站（work station） 数据库（database） 多媒体（multimedia） 局域网（LAN） 驱动器（driver） 分布式（distributed） 扫描仪（scanner）
4-gram	20	马尔可夫（Markov）自底向上（bottom up） 闭路电视（cable TV） 机器人学（Robot science） 非格式化（unformat）
phrase	3471	贝叶斯信念（Bayes belief） 主题词典（Thesaurus）解压缩(decompression） 谓词演算（Predicate calculation） 调制解调器（MODEM）

6. Conclusion

This research presents a chi-square method based approach to semi-automatically generate terminology dictionaries. This approach integrates such methods as new word collecting, terminology word extraction and terminology phrase generation. It significantly reduces most of the hard work which should be done manually, and reduce the effort and time which are needed to transport a natural language processing work from one domain to another. Using this terminology dictionary, encouraging results has been achieved about the coverage of terminologies.

This research has practical importance in many domain related natural language applications. It can improve indexing results. It can help to decide texts' category. It also can help to rank documents with user queries. In fact, this approach will soon be embedded into an integrated Chinese information processing system -- FDASCT【12】.

Our future work mainly includes the utilization of deeper text processing techniques such as part of speech tagging and partial syntactic analysis in phrase generation. Word pairs would be discarded if there are no consistent syntactic relations between constituent words. And those non-noun phrases would also be discarded since terminologies are always nouns. Thus manual effort can be further reduced.

Reference

【1】 Church K.W, Hanks P., *Word Association Norms, Mutual Information, and Lexicography*, Computational Linguistic , 16:1, 1990 , 22 — 29

【2】 Church K.W, Gale W.A. et. al, *Using Statistics in Lexical Analysis*, Lexical Acquisition: Using On-line Resources to Build a Lexicon, edited by Uri Zernik, Lawrence Erlbaum, Hillsdale, New Jersey, 115-165

【3】 Dagan I. and Church K.W, *Termight: Identifying and translating technical terminology*, ANLP, 34--40, 1994

【4】 Daille B., *Study and implementation of combined techniques for automatic extraction of terminology*, 29--36, The Balancing Act, Combining Symbolic and Statistical Approaches to Language -- Proceedings of the Workshop, 1994

【5】 Dunning T., *Accurate Methods for the Statistics of Surprise and Coincidence*, Computational Linguistic 19:1, 1993, 61 — 74

【6】 Haas, Stephanie, He Shaoyi. *Toward the automatic Identification of Sublanguage Vocabulary*, Information Processing & Management, 29:6, 1993, 721-732

【7】 Huang Xuan- jing, Wu Li-de, Wang Wen-xin, Ye Dan-jin, *基于机器学习的无需人工编制词典的切词系统* (*A Machine Learning Based System Without Maual Dictionary*) , 模式识别与人工智能 (Pattern Recognition and Artificial Intelligence), 1996. 12 , 9:4 , 297 — 303

【8】 Justeson J. and Katz S., *Technical terminology: some linguistic properties and an algorithm for identification in text*, Natural Language Engineering, 1995, 1:1, 9 — 28

【9】 Pascale Fung, Dekai Wu, *Statistical Augmentation of a Chinese machine-readable Dictionary*, Technical Report HKUST-CS94-31, November 1994

【10】 Smadja, Frank, *Retrieve collocations from text: Xtract*, Computational Linguistic 19:1. 1993, 143 — 177

【11】 Smadja, Frank, et al., *Translating collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistic 22:1, 1996, 1 — 38

【12】 Wu Li-de ,Wei Xiong-guan. Huang Xuan-jing, et al, *Fudan Abstract System of Chinese Text*, 1996. 6, Communications of COLIPS

【13】 Wang Kai-zhu, et al, *无词典自动分词的研究 (Study of Nondictionay Chinese Segmentation)*, 计算语言学进展与应用 (Advances and Applications on Computational Linguistics), Tsinghua University Press, 1995 , 359

【14】 Zhang Shu-wu, et al, *汉语语音识别用电子词典的自动建立方法研究 (An Automatic Buildng Method of Electronic Dictionary Used for Chinese Speech Recognition)*, 计算语言学进展与应用 (Advances and Applications on Computational Linguistics), Tsinghua University Press, 1995 , 219 — 224

【15】 Zhou J. and Dapkus P., *Automatic suggestion of significant terms for a predefined topic*, Proceedings of Third the Workshop on Very Large Corpora, 131--147, 1995