Extracting Terminologically Relevant Collocations in the Translation of Chinese Monograph*

Byeong-Kwu Kang, Bao-Bao Chang, Yi-Rong Chen, and Shi-Wen Yu

The Institute of Computational Linguistics, Peking University, Beijing, 100871, China
{kbg43, chbb, chenyr, yusw}@pku.edu.cn

Abstract. This paper suggests a methodology which is aimed to extract the terminologically relevant collocations for translation purposes. Our basic idea is to use a hybrid method which combines the statistical method and linguistic rules. The extraction system used in our work operated at three steps: (1) Tokenization and POS tagging of the corpus; (2) Extraction of multi-word units using statistical measure; (3) Linguistic filtering to make use of syntactic patterns and stop-word list. As a result, hybrid method using linguistic filters proved to be a suitable method for selecting terminological collocations, it has considerably improved the precision of the extraction which is much higher than that of purely statistical method. In our test, hybrid method combining "Log-likelihood ratio" and "linguistic rules" had the best performance in the extraction. We believe that terminological collocations and phrases extracted in this way, could be used effectively either to supplement existing terminological collections or to be used in addition to traditional reference works.

1 Introduction

Communication between different individuals and nations is not always easy, especially when more than one language is involved. This kind of communication can include translation problems, which can be solved by the translators who bridge the gap between two different languages.

Through the past decade, China and Korea have been undergoing large economic, cultural exchange, which invariably affects all aspects of communication, particularly translation. New international contacts, foreign investments as well as cross-cultural communication have caused an enormous increase in the volume of translations produced and required. But by now, most of all this translation work has been conducted by translators alone, which bears the burden of an enormous translation task to them.

In order to accomplish these tasks with maximum efficiency and quality, a new translation method supported by computer technology has been suggested. MAHT, also known as computer-assisted translation involves some interaction between translator and the computer. It seems to be more suited for the needs of many

* This work has been supported by The National Basic Research Program of China(973 program, No. 2004CB318102) and the 863 program (No. 2001AA114210, 2002AA117010).

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 1017 – 1028, 2005.

organizations which have to handle the translation of the documents. Computer-assisted translation systems are based on “translation memory” and “terminology databases”. With translation memory tools, translators have immediate access to previous translations of the text, which they can then accept or modify.

Terminology management systems also can prove very useful in supporting translator’s work [2, 11]. Most translators use some sort of glossary or terminology database, especially in the translation of the technical documents or academic monograph. Many translation bureaux have the collection of the terminology data bases. But time pressure and costs make it difficult to get glossary building task done fully manually. Thus there is a pressing need for the tool which is computationally supported. For Chinese, other than for English, terminology management tools are not so sophisticated that they could provide wide enough coverage to be directly usable for the translators.

We are contemplating, in this article, situations where computational support is sought to extract the term candidate, construct or enhance such terminology databases. Our work will be more focused on the problem of terminologically relevant collocation extraction.

In order to extract multiword terms from the domain corpus, three main strategies have been proposed in the literature. First, linguistic rule-based systems propose to extract relevant terms by making use of parts of speech, lexicons, syntax or other linguistic structure [2, 4]. This methodology is language dependent rather than language independent, and the system requires highly specialized linguistic techniques to identify the possible candidate terms. Second, purely statistical systems extract discriminating multiword terms from the text corpora by means of association measures [5, 6, 7]. As they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and extract relevant units independently from the domain and the language of the input text. Finally, hybrid methodologies define co-occurrences of interest in terms of syntactical patterns and statistical regularities [1, 3, 9].

There is no question that the term extraction work comes into play when the tools are parameterized in such a way as to provide as much relevant material (maximizing recall and precision), and as little “noise” as possible. As seen in the literature, neither purely rule-based approach nor statistic based approach could bring an encouraging result alone[3, 4]. The main problem is the "noise". So we need to find a combined technique for reducing this “noise”. In this paper, we have taken a hybrid approach which combines the linguistic rules and statistical method. First, we applied a linguistic filter which selects candidates from the corpus. Second, the statistical method was used to extract the word class combinations. And then, the results of several experiments were evaluated and compared with each other.

2 Methodology Overview

The basic idea in our work is that the extraction tool operates on pre-processed corpus which contains the results of tokenizing word and word class annotation (POS-tagging). Figure1 contains an annotated sentence from one of the Chinese academic monograph[18].

```
<s id=2>
随着/p 社会/n 生活/n 的/u 日益/d 信息化/v ，/w 人们/n 越来越/d 强烈/a 地/u 希望/v 用/p 自然/n 语言/n 同/p 计算机/n 交流/v 信息/n 。/w
```

Fig. 1. Sample annotated text (tagged by the Peking University Tagger)

And the extraction routine used in our work operated at three steps: (1)Tokenization and POS Tagging; (2)Extraction of the candidates from the corpus; (3)Linguistic filtering(making use of syntactic patterns and stop-word list). The schema in Figure2 summarizes the three steps of pre-processing and extracting the term candidate. The extraction is automatic once the appropriate templates are designed.

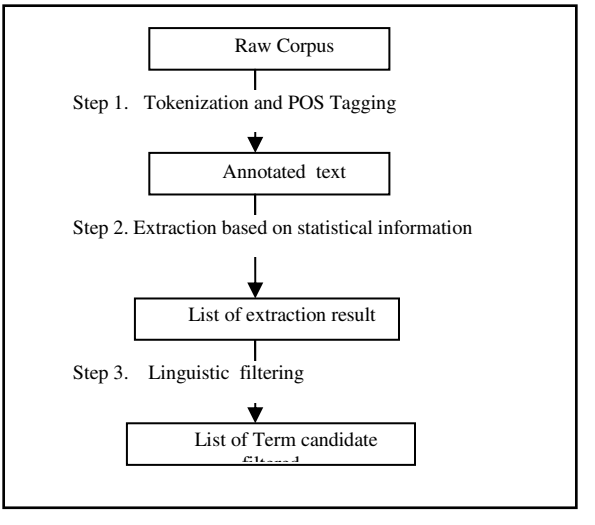

Fig. 2. Simplified schema of term extraction from a corpus

3 Statistical Method

Statistical methods in computational linguistics generally share the fundamental approach to language viewed as a string of characters, tokens or other units, where patterns are discovered on the basis of their recurrence and co-occurrence. Accordingly, when we approach the extraction of multi-word terms from a statistical point of view, we initially retrieve the word sequences which are not only frequent in their occurrence but also collocating each other.

Before a statistical methodology could be developed, some characteristics of terms in Chinese had to be established. In Chinese, the length of terms can vary from single word to multi-words(n-gram), with the majority of entries being less than 4-word items, usually two word items(bi-gram) (See in 4.3). The number of n-grams with n>4

is very small, and the occurrence of which is also rare. Therefore, the problems of bi-grams, tri-grams and 4-grams are primarily taken into considerations in our work.

Now let us consider the correlation between two neighboring words A and B. Assuming that these two words are terminologically relevant units, we can intuitively expect that they occur more often than random chance. From a statistical point of view, this probability can be measured by several statistical methods, such as "co-occurrence frequency", "Mutual Information", "Dice coefficient", "Chi-square test", "log-likelihood", etc[1, 6, 15].

Table 1 lists several statistical measures which have been widely used in extracting collocations. In table 1: *XY* represents any two word item; $\overline{X}$ stands for all words except *X*; *N* is the size of corpus; f_X and P_X are frequency and probability of *X* respectively; f_{XY} and P_{XY} are frequency and probability of *XY* respectively。And assuming that two words *X* and *Y* are independent of each other, the formulas are represented as follows:

Table 1. Statistical methods used in multi word extraction

Method	Formula
Frequency(Freq)	f_{XY}
Mutual Information (MI)	$\log_2 \frac{P_{XY}}{P_X P_Y}$
Dice Formula (Dice)	$\frac{2f_{XY}}{f_X + f_Y}$
Log-likelihood(Log-L)	$-2\log \frac{(P_X P_Y P_{\bar{X}} P_{\bar{Y}})^{f_Y}}{(P_{XY} P_{\overline{XY}})^{f_{XY}} (P_{X\bar{Y}} P_{\overline{X\bar{Y}}})^{f_{\bar{X}Y}}}$
Chi-squared(Chi)	$\frac{N(f_{XY} f_{\bar{X}\bar{Y}} - f_{X\bar{Y}} f_{\bar{X}Y})^2}{(f_{XY}+f_{X\bar{Y}})(f_{XY}+f_{\bar{X}Y})(f_{\bar{X}\bar{Y}}+f_{X\bar{Y}})(f_{\bar{X}\bar{Y}}+f_{\bar{X}Y})}$

For the purposes of this work, we used these five statistics to measure the correlation of neighboring words. The statistical criterion of judgments is the value of measures which can judge the probability whether they belong to the rigid collocations or not. From a statistical point of view, we can say that if the value of measure is high, the two word combination is more likely to be a rigid collocation. And *XY* could be accepted as a collocation if its statistical value is larger than a given threshold. 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 were processed in the same way. To compute the correlation coefficient of all tri-grams, we just considered a tri-gram as the

combination of one bi-gram and one word, and then calculated their correlation coefficient. Similarly, a 4-gram was considered either as the combination of a tri-gram and a word, or the combination of two bi-grams [12].

As mentioned before, our methodology was tested on pre-processed corpus which contained the result of word class annotation. The extraction test was delivered on word sequence (POS tags) combinations. And the test corpus was a Chinese academic monograph [18]. The size of this corpus is 0.2 million Chinese characters, including about 5,000 sentences. In our test, the extraction of multi-word units was based on 65,663 candidate bi-grams. Among these candidates, when their correlation coefficients were higher than a given threshold, they were considered as multi-word unit, and then sorted in descending order. The results of experiment are shown in Figure3.

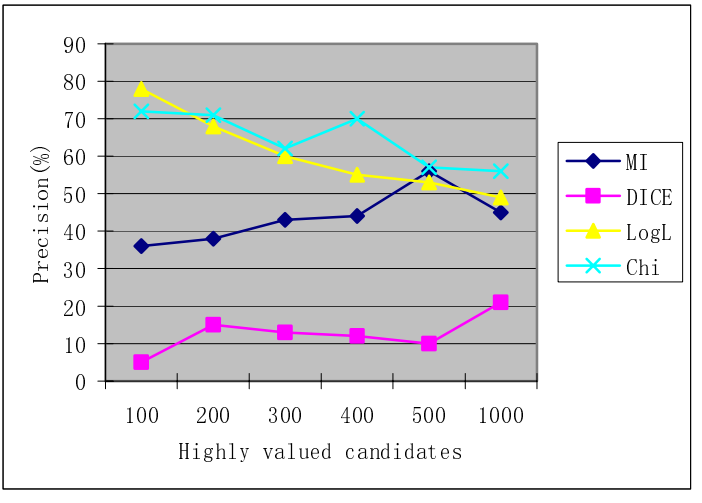

Fig. 3. Comparison of Extraction Performance between different statistical measures

Table 2. The sample result sorted by Chi Square value

1stWord	2ndWord	Chi	LogL	DICE	MI
信息	词典	7822.14	1278.48	517.581	5.28183
显现	出来	4233.43	42.8348	2520	10.4636
集体	量词	3085.64	160.647	4560	7.59925
字段	填	1461.41	424.818	767.36	3.90891
概括	地	809.168	38.2226	844	7.66964
趋向	动词	752.637	124.353	787.243	5.16173
增加	了	619.694	111.527	1600	5.02194
转换	为	582.425	52.0341	516.444	6.40501
不同	的	549.119	286.884	17037.1	2.66906
状态	词	336.283	58.0757	2166.67	5.13281
查	词典	296.196	52.8541	544.348	4.96744
也	是	228.596	122.119	523.597	2.2667

An examination of the results first showed a significant difference in precision. Checked by hand, the precisions of Chi-square value and Log-likelihood ratio were relatively high. In contrast, the precisions of Mutual information and Dice formula were not so ideal.

Considering the size of the corpus and the terminological richness of the texts, this result is not very encouraging. Regardless of any statistical measure, the precision and coverage of the extraction are not so high that could be directly used in the application system.

More over, as shown in table 2, the purely statistical system extracts all multi-word units regardless of their types, so that we can also find sequences like "增加 [zengjia](add) 了 [le](auxiliary word)", "不同 [butong](different) 的 [de](auxiliary word)", "也[ye](also)是[shi](be)", "转换[zhuanhuan](change)为[wei](become)", etc., for which we have no use in terminology. Clearly the output must be thoroughly filtered before the result can be used in any productive way.

On the whole, the somewhat disappointing outcome of the statistical method provoked us to rethink the methodology and tried to include more linguistic information in the extraction of terminology.

4 Hybrid Method Combining Statistical Method and Linguistic Rules

To improve the precision and recall of the extraction system, it was decided to use two criteria determining whether a sequence was terminologically relevant or not. The first was to use the frequent syntactic patterns of terms. The idea underlying this method is that multi-word terms are constructed according to more or less fixed syntactic patterns, and if such patterns are identified for each language, it is possible to extract them from a POS tagged corpus. The second was to use a stop-word filter that a term can never begin or end in a stop-word. This would filter out things not relevant with the domain-specific collocation or term.

4.1 Syntactic Patterns of Terms in Chinese

Before a methodology for extracting the terminologically relevant word units could be developed, some characteristics of terms in Chinese had to be established. We were especially interested in the following: How many words do terms usually have in Chinese? What is the structure of multi-word units in terms of syntax and morphology? What kind of terms can be successfully retrieved by computational methods?

To find answers to the above questions, an existing terminology database could be used as a sample. Because the source text to be tested in our work is related with computational or linguistic domain, we selected the terminology database of computational linguistics which was constructed by Peking University. This term bank currently contains over 6,500 entries in English and Chinese.

An analysis of 6,500 term entries in Chinese showed that the length of terms can vary from 1 to over 6 words, with the majority of entries being two-word items, usually a "noun+noun" sequence. The second most frequent type is a single-word term. As less than 5% of all entries exceed 4 words and single word terms can be

identified with the use of monolingual or bilingual dictionary[1], we decided that automatic extraction should be limited to sequences of 2-4 words.

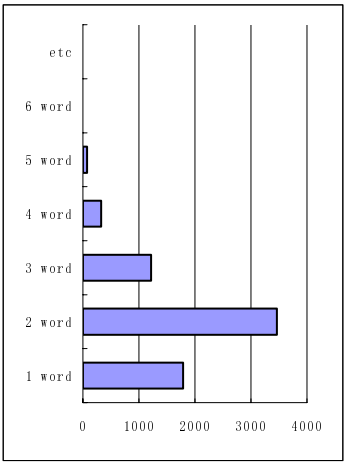

Fig. 4. Length of Chinese terms

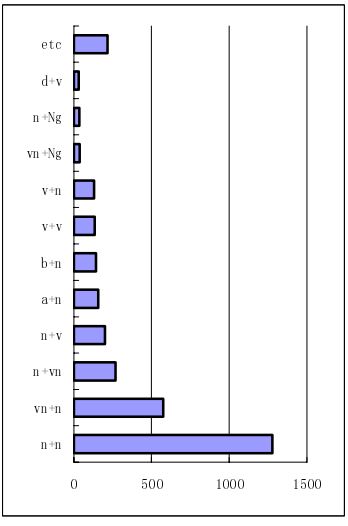

Fig. 5. Syntactic patterns of two word terms

As the next step we manually analyzed the syntactic patterns of Chinese terms and ordered them according to frequency. These patterns were needed for the second part of the experiment, the "linguistically motivated" filtering. According to the analysis of the existing terms, multi-word terms have some kinds of fixed syntactic patterns. In many cases, these syntactic patterns are based on the combinations of different two word classes, such as "noun+noun", "gerend verb+noun", "adjective+noun", "noun+suffix" etc. We found that there were about 30 syntactic patterns which covered almost 95% in the two word combinations. Therefore, we decided that these patterns could be used filtering in the extraction. In figure 6, certain types of word combinations are more typical for technical vocabulary than for general language.

More than three word combinations also can be divided into two small parts whose syntactic structures are the same as those of two word terms. For example: "(n+n)+n", "(vn+n)+n", "(v+n)+(n+vn)", "(a+n)+(vn+n)", etc. Therefore when we extracted three-word or four-word units, we didn't set another syntactic rule for them. We just considered tri-gram as the combination of one bi-gram and one word. Similarly, 4-gram was considered as the combination of different two bi-grams.

Although we admit that these syntactic patterns are typical for certain type of technical prose only, we don't think that they could filter out all the irrelevant units. If

[1] To extract a glossary of terms from a corpus, we must first identify single-word terms. But it might be slightly confusing for the computer to identify the single word terms alone. So we would like to set aside this problem for the sake of achieving efficiency. But we believe that the translator might not be troubled with single terms if he has some kind of dictionary in the translation of the source text.

we extract all combinations of a certain POS-shape, additional filters are needed afterwards, to identify those combinations which are terminologically relevant.

```
Char*Patterns={"n+n","vn+n","n+vn","n+v","a+n","b+n","v+v","v+n"
,"vn+Ng","n+Ng","d+v","m+n","h+n","f+v","a+v","f+n","j+n","a+Ng
","vn+k","b+vn","b+Ng","Ag+n","v+Ng","a+nz","vn+v","nz+n","b+k
","v+k","j+n","nz+v",null};
```

Fig. 6. The syntactic patterns for filtering[2]

4.2 Stop-Word Filter in Chinese

When we examine multi word units regardless of their type, we can easily find some words which have no use in terminology. These irrelevant or meaningless data is a noise for extracting desired data. To resolve this problem, we can make use of the stop word list to be filtered. In the system, it would filter out things irrelevant with the domain-specific collocation or term. But how can we make the set of stop words? Indeed, the stop word list is rather flexible than firmly fixed in their usage. Whenever the words are frequent and meaningless in text, they can be stop words in a given task.

For practical purposes, we used the word frequency data of the large technical domain corpora which was constructed by Beijing Language and Cultural University. In this data, we randomly selected the 2,000 words most highly frequent in their usage. And then we examined whether the frequent words were terminologically relevant or not. The analysis of the word data showed that 77.6% were domain dependent which could be the part of term, and 22.4% were general words. It means that terminologically irrelevant words amounted to about 450 words of the highly frequent 2000 words in technical corpora. The results are shown in Table 3.

Table 3. The results of analysis on the high frequency words

Frequency	Terminologically Relevant words	Terminologically Irrelevant words	Example
1-100	44(44%)	56(56%)	的(aux), 是(be), 和(and), 在(at), 中(middle), etc.
101-200	58(58%)	42(42%)	提供(provide),给(to),当(serve as),具有(possess), etc.
201-500	229(76.3%)	71(23.7%)	好(good),为了(for),某(some), 只(only),其它(other), etc.
501-1000	408(81.6%)	92(18.4%)	相当(quite), 看(see), 引起(arose), 指出(indicate),etc.
1001-2000	813(81.3%)	187(18.7%)	出发(leave),从事(engage),甚至(even),不必(need not),etc
Total	1552(77.6%)	448(22.4%)	

[2] These POS patterns are based on the tag sets of Peking University.

According to these analyzed data, we made the set of stop words which amounted to about 450 words. And we used them for filtering out the frequent, meaningless words in a given text before the output can be used in any productive way.

5 Experiments

The hybrid methods combining statistical measure and linguistic rules were tested on pre-processed corpus. Based on the statistical method, the extraction test was limited to the boundary of the frequent syntactic patterns first, and then filtered out by the stop word list. Three different statistical measures were used to enhance the precision of the extraction, such as Log-likelihood ratio, Chi-square test and Mutual information. Because of the poor performance in our first test, Dice formula was not used in hybrid method any more. Therefore, we have delivered three different experiments using like "LogL + Liguistic Filter", "Chi + Liguistic Filter", "MI + Liguistic Filter" methods.

In Figure 7, we present the comparative results of precision rate among these different experiments. In order to measure the precision rate of the result, we used the grammatical criterion: A multi word n-gram could be considered as accurate result if it is grammatically appropriate. By grammatical appropriation, we refer to compound noun phrase or compound verb phrase, since with majority of multi-word terms have these structures.

As a result, hybrid method using linguistic filters proved to be a suitable method for selecting terminological collocations, and it has considerably improved the precision of the extraction. The precision was much higher than that of purely statistical method, retrieving appropriate result almost 10%-20% higher than in the first experiment. In our test, hybrid method combining "Log-likelihood ratio" and "linguistic rules" had the best performance in the extraction. The precision was higher than 90%. According to their performance, the results of different experiments can be arranged like:

LogL+Filter > Chi+Filter > MI+Filter > LogL > Chi > MI > Dice

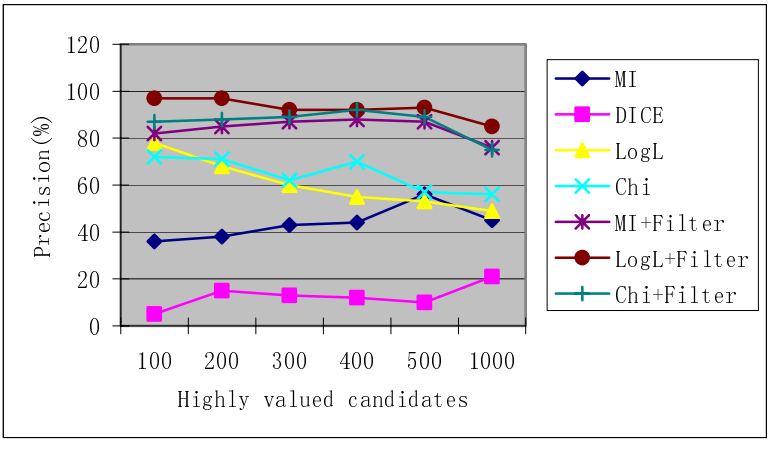

Fig. 7. Comparison of Extraction Performance between statistical measures and hybrid measure

In the analysis of the extraction data, we examined the precision of every 100 multi-word candidates which sorted in descending order. Considering the size of corpus, we compared the results within the highly valued 1000 candidates. A sample of the highly valued output is seen in Table 4.

Table 4. The sample result sorted by Log-likelihood ratio

1stWord	2ndWord	LogL+Filter	CHI+Filter	MI+Filter
语法	信息	1026.38	3748.65	4.20189
信息	处理	1020.43	5102.98	4.93017
信息	词典	981.323	3651.52	4.23672
自然	语言	899.731	7805.59	6.16647
汉语	语法	734.213	2284.06	3.76964
计算	语言学	718.016	14931.3	7.80401
语言学	研究所	557.888	13569.4	8.11656
语法	功能	537.196	2361.49	4.60008
本	字段	500.011	12919.7	8.26776
前接	成分	363.259	3535.04	6.19858
电子	词典	355.499	2053.22	5.29117
单	音节	345.551	6733.13	7.55539
趋向	补语	339.45	6092.73	7.41208
语言	信息	329.536	1061.09	3.76944
专有	项目	316.792	8130.74	8.08733

As seen in Table 4, although not all these units would be considered terms in the traditional sense of the word, most of them either contain terms or include terminologically relevant collocations. Besides, our extraction started from these two word items, expanded to extract multi-word units like three word or four word units. Finally we could extract multi word units such as the following sample:

Table 5. The sample of multi-word terms

	Terminologically relevant units
Two word units	语法功能(grammatical function), 趋向补语(directional complement), 规格说明书(specification), 容器量词(container classifier), 使用频度(usage frequency), etc.
Three word units	语法信息词典(grammatical knowledge-base), 中文信息处理(Chinese Information Processing), 语音识别系统(speech recognition system), etc.
Four word units	机器翻译系统设计(MT system design), 语言信息处理技术(language information processing technology), 上下文无关语法(context free grammar), etc.

On the whole, as we think that the performance of the extraction was quite good, this method could be applicable in the translation system.

6 Conclusions and Future Work

The paper presents a methodology for the extraction of terminological collocations from academic documents for translation purposes. It shows that statistical methods are useful because they can automatically extract all the possible multi word units according to the correlation coefficient. But the purely statistical system extracts all multi-word units regardless of their types, so that we also find sequences which are meaningless in terminology. Clearly the output must be thoroughly filtered before the result can be used in any productive way. To improve the precision of the extraction system, we decided to use linguistic rules determining whether a sequence was terminologically relevant or not. The frequent syntactic patterns of terminology and the stop-word list were used to filter out the irrelevant candidates. As a consequence, hybrid method using linguistic filters proved to be a suitable method for selecting terminological collocations, and it has considerably improved the precision of the extraction. The precision was much higher than that of purely statistical method.

We believe that terminological collocations and phrases extracted in this way, could be used effectively either to supplement existing terminological collections or to be used in addition to traditional reference works.

In future we envisage the development of techniques for the alignment of exact translation equivalents of multi-word terms in Chinese and Korean, and one way of doing so is by finding correspondences between syntactic patterns in both languages. Translation memory systems already store translations in a format similar to a parallel corpus, and terminology tools already involve functions such as “auto-translate” that statistically calculate the most probable translation equivalent. By refining these functions and making them language specific, we could soon be facing a new generation of tools for translators. It remains to be seen, however, whether they can really be implemented into translation environments on broad scale.

References

1. Chang Bao-Bao, Extraction of Translation Equivalent Pairs from Chinese-English Parallel Corpus, Terminology Standardization and Information Technology, pp24-29, 2002.
2. Bourigault, D. Lexter, A Natural Language Processing Tool for Terminology Extraction. In Proceedings of 7th EURALEX International Congress, 1996.
3. Daille, B. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In The balancing act combining symbolic and statistical approaches to language. MIT Press, 1995.
4. Ulrich Heid, A linguistic bootstrapping approach to the extraction of term candidates from German text, http://www.ims.uni-stuttgart.de/~uli/papers.html, 2000 .
5. Sayori Shimohata, Toshiyuki Sugio, JunjiI Nagata, Retrieving Domain-Specific Collocations By Co-Occurrences and Word Order Constraints, Computational Intelligence, Vol 15, pp92-100, 1999.
6. Shengfen Luo, Maosong Sun Nation,Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures, 2003
7. Smadja, F. Retrieving Collocations From Text: XTRACT. In Computational Linguistics, 19(1) (pp 143--177).1993.

8. David Vogel, Using Generic Corpora to Learn Domain-Specific Terminology, Workshop on Link Analysis for Detecting Complex Behavior, 2003
9. Dias, G. & Guilloré, S. & Lopes, J.G.P. Multiword Lexical Units Extraction. In Proceedings of the International Symposium on Machine Translation and Computer Language Information Processing. Beijing, China. 1999.
10. Feng Zhi-Wei, An Introduction to Modern Terminology, Yuwen press, China, 1997.
11. Gaël Dias etc, Combining Linguistics with Statistics for Multiword Term Extraction, In Proc. of Recherche d'Informations Assistee par Ordinateur, 2000.
12. Huang Xuan-jing & Wu Li-de & Wang Wen-xin, Statistical Acquisition of Terminology Dictionary, the Fifth Workshop on Very Large Corpora, 1997
13. Jiangsheng Yu，Automatic Detection of Collocation，http://icl.pku.edu.cn/yujs/，2003
14. Jong-Hoon Oh, Jae-Ho Kim, Key-Sun Choi, Automatic Term Recognition Through EM Algorithm, http://nlplab.kaist.ac.kr/, 2003
15. Patrick Schone and Daniel Jurafsky, Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?, In proceedings of EMNLP, 2001.
16. Philip Resnik, I. Dan Melamed, Semi-Automatic Acquisition of Domain-Specific Translation Lexicons, Proceedings of the fifth conference on Applied natural language processing, pp 340-347, 1997.
17. Sui Zhi-Fang, Terminology Standardization using the NLP Technology, Issues in Chinese Information Processing,pp341-352, 2003.
18. Yu Shi-wen, *A Complete Specification on The Grammatical Knowledge-base of Contemporary Chinese,* Qinghua Univ. Press, 2003