Chinese-English Term Translation Mining Based on Semantic Prediction

Gaolin Fang, Hao Yu, and Fumihito Nishino
Fujitsu Research and Development Center, Co., LTD. Beijing 100016, China
{glfang, yu, nishino}@cn.fujitsu.com

Abstract

Using abundant Web resources to mine Chinese term translations can be applied in many fields such as reading/writing assistant, machine translation and cross-language information retrieval. In mining English translations of Chinese terms, how to obtain effective Web pages and evaluate translation candidates are two challenging issues. In this paper, the approach based on semantic prediction is first proposed to obtain effective Web pages. The proposed method predicts possible English meanings according to each constituent unit of Chinese term, and expands these English items using semantically relevant knowledge for searching. The refined related terms are extracted from top retrieved documents through feedback learning to construct a new query expansion for acquiring more effective Web pages. For obtaining a correct translation list, a translation evaluation method in the weighted sum of multi-features is presented to rank these candidates estimated from effective Web pages. Experimental results demonstrate that the proposed method has good performance in Chinese-English term translation acquisition, and achieves 82.9% accuracy.

1 Introduction

The goal of Web-based Chinese-English (C-E) term translation mining is to acquire translations of terms or proper nouns which cannot be looked up in the dictionary from the Web using a statistical method, and then construct an application system for reading/writing assistant (e.g., 三国演义→The Romance of Three Kingdoms). During translating or writing foreign language articles, people usually encounter terms, but they cannot obtain native translations after many lookup efforts. Some skilled users perhaps resort to a Web search engine, but a large amount of retrieved irrelevant pages and redundant information hamper them to acquire effective information. Thus, it is necessary to provide a system to automatically mine translation knowledge of terms using abundant Web information so as to help users accurately read or write foreign language articles.

The system of Web-based term translation mining has many applications. 1) Reading/writing assistant. 2) The construction tool of bilingual or multilingual dictionary for machine translation. The system can not only provide translation candidates for compiling a lexicon, but also rescore the candidate list of the dictionary. We can also use English as a medium language to build a lexicon translation bridge between two languages with few bilingual annotations (e.g., Japanese and Chinese). 3) Provide the translations of unknown queries in cross-language information retrieval (CLIR). 4) As one of the typical application paradigms of the combination of CLIR and Web mining.

Automatic acquisition of bilingual translations has been extensively researched in the literature. The methods of acquiring translations are usually summarized as the following six categories. 1) Acquiring translations from parallel corpora. To reduce the workload of manual annotations, researchers have proposed different methods to automatically collect parallel corpora of different language versions from the Web (Kilgarriff, 2003). 2) Acquiring translations from non-parallel corpora (Fung, 1997; Rapp, 1999). It is based on the clue that the context of source term is very similar to that of target translation in a large amount of corpora. 3) Acquiring translations from a combination of translations of constituent words (Li et al., 2003). 4) Acquiring translations using cognate matching (Gey, 2004)

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 199–206,
Sydney, July 2006.

or transliteration (Seo et al., 2004). This method is very suitable for the translation between two languages with some intrinsic relationships, e.g., acquiring translations from Japanese to Chinese or from Korean to English. 5) Acquiring translations using anchor text information (Lu et al., 2004). 6) Acquiring translations from the Web. When people use Asia language (Chinese, Japanese, and Korean) to write, they often annotate associated English meanings after terms. With the development of Web and the open of accessible electronic documents, digital library, and scientific articles, these resources will become more and more abundant. Thus, acquiring term translations from the Web is a feasible and effective way. Nagata et al. (2001) proposed an empirical function of the byte distance between Japanese and English terms as an evaluation criterion to extract translations of Japanese words, and the results could be used as a Japanese-English dictionary.

Cheng et al. (2004) utilized the Web as the corpus source to translate English unknown queries for CLIR. They proposed context-vector and chi-square methods to determine Chinese translations for unknown query terms via mining of top 100 search-result pages from Web search engines.

Zhang and Vines (2004) proposed using a Web search engine to obtain translations of Chinese out-of-vocabulary terms from the Web to improve CLIR performance. The method used Chinese as query items, and retrieved previous 100 document snippets by Google, and then estimated possible translations using co-occurrence information.

From the review above, we know that previous related researches didn't concern the issue how to obtain effective Web pages with bilingual annotations, and they mainly utilized the frequency feature as the clue to mine the translation. In fact, previous 100 Web results seldom contain effective English equivalents. Apart from the frequency information, there are some other features such as distribution, length ratio, distance, keywords, key symbols and boundary information which have very important impacts on term translation mining. In this paper, the approach based on semantic prediction is proposed to obtain effective Web pages; for acquiring a correct translation list, the evaluation strategy in the weighted sum of multi-features is employed to rank the candidates.

The remainder of this paper is organized as follows. In Section 2, we give an overview of the system. Section 3 proposes effective Web page collection. In Section 4, we introduce translation candidate construction and noise solution. Section 5 presents candidate evaluation based on multi-features. Section 6 shows experimental results. The conclusion is drawn in the last section.

2 System Overview

The C-E term translation mining system based on semantic prediction is illustrated in Figure 1.

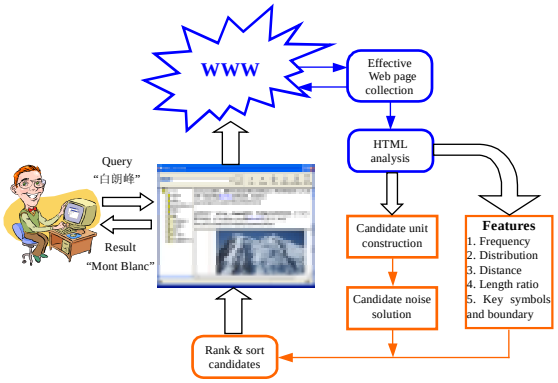

Figure 1. The Chinese-English term translation mining system based on semantic prediction

The system consists of two parts: Web page handling and term translation mining. Web page handling includes effective Web page collection and HTML analysis. The function of effective Web page collection is to collect these Web pages with bilingual annotations using semantic prediction, and then these pages are inputted into HTML analysis module, where possible features and text information are extracted. Term translation mining includes candidate unit construction, candidate noise solution, and rank&sort candidates. Translation candidates are formed through candidate unit construction module, and then we analyze their noises and propose the corresponding methods to handle them. At last, the approach using multi-features is employed to rank these candidates.

Correctly exploring all kinds of bilingual annotation forms on the Web can make a mining system extract comprehensive translation results. After analyzing a large amount of Web page examples, translation distribution forms is summarized as six categories in Figure 2: 1) Direct annotation (a). some have nothing (a1), and some have symbol marks (a2, a3) between the pair; 2) Separate annotation. There are English letters (b1) or some Chinese words (b2, b3) between the pair; 3) Subset form (c); 4) Table form (d); 5) List form (e); and 6) Explanation form (f).

Figure 2. The examples of translation distribution forms

3 Effective Web page collection

For mining the English translations of Chinese terms and proper names, we must obtain effective Web pages, that is, collecting these Web pages that contain not only Chinese characters but also the corresponding English equivalents. However, in a general Web search engine, when you input a Chinese technical term, the number of retrieved relevant Web pages is very large. It is infeasible to download all the Web pages because of a huge time-consuming process. If only the 100 abstracts of Web pages are used for the translation estimation just as in the previous work, effective English equivalent words are seldom contained for most Chinese terms in our experiments, for example: “三国演义, 三好学生, 百慕大三角, 车牌号”. In this paper, a feasible method based on semantic prediction is proposed to automatically acquire effective Web pages. In the proposed method, possible English meanings of every constituent unit of a Chinese term are predicted and further expanded by using semantically relevant knowledge, and these expansion units with the original query are inputted to search bilingual Web pages. In the retrieved top-20 Web pages, feedback learning is employed to extract more semantically-relevant terms by frequency and average length. The refined expansion terms, together with the original query, are once more sent to retrieve effective relevant Web pages.

3.1 Term expansion

Term expansion is to use predictive semantically-relevant terms of target language as the expansion of queries, and therefore resolve the issue that top retrieved Web pages seldom contain effective English annotations. Our idea is based on the assumption that the meanings of Chinese technical terms aren’t exactly known just through their constituent characters and words, but the closely related semantics and vocabulary information may be inferred and predicted. For example, the corresponding unit translations of a term “三国演义” are respectively: three(三), country, nation(国), act, practice(演), and meaning, justice(义). As seen from these English translations, we have a general impression of “things about three countries”. After expanding, the query item for the example above becomes "三国演义"+ (three | country | nation | act | practice | meaning | justice). The whole procedure consists of three steps: unit segmentation, item translation knowledge base construction, and expansion knowledge base evaluation.

Unit segmentation. Getting the constituent units of a technical term is a segmentation procedure. Because most Chinese terms consist of out-of-vocabulary words or meaningless characters, the performance using general word segmentation programs is not very desirable. In this paper, a segmentation method is employed to handle term segmentation so that possible meaningful constituent units are found. In the inner structure of proper nouns or terms, the rightmost unit usually contains a headword to reflect the major meaning of the term. Sometimes, the modifier starts from the leftmost point of a term to form a multi-character unit. As a result, forward maximum matching and backward maximum matching are respectively conducted on the term, and all the overlapped segmented units are added to candidate items. For example, for the term “abcd”, forward segmented units are “ab cd”, backward are “a bcd”, so “ab cd a bcd” will be viewed as our segmented items.

Item translation knowledge base construction. Because the segmented units of a technical term or proper name often consist of abbreviation items with shorter length, limited translations provided by general dictionaries often cannot satisfy the demand of translation prediction. Here, a semantic expansion based method is proposed to construct item translation knowledge base. In this method, we only keep these nouns or adjective items consisting of 1-3 characters in the dictionary. If an item length is greater than two characters and contains any item in the knowledge base, its translation will be added as translation candidates of this item. For example, the Chinese term “流通股” can be segmented into the units “流通” and “股”, where “股” has only two English meanings “section, thigh” in the dictionary. However, we can derive its meaning us-

ing the longer word including this item such as “股东, 股票”. Thus, their respective translations “stock, stockholder” are added into the knowledge base list of “股” (see Figure 3).

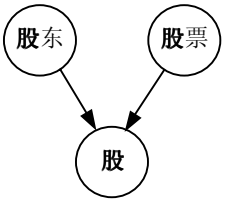

Figure 3. An expansion example in the dictionary knowledge base

Expansion knowledge base evaluation. To avoid over-expanding of translations for one item, using the retrieved number from the Web as our scoring criterion is employed to remove irrelevant expansion items and rank those possible candidates. For example, “股” and its expansion translation “stock” are combined as a new query “股 stock –股票”. It is sent to a general search engine like Google to obtain the count number, where only the co-occurrence of “股” and “stock” excluding the word “股票” is counted. The retrieved number is about 316000. If the occurrence number of an item is lower than a certain threshold (100), the evaluated translation will not be added to the item in the knowledge base. Those expanded candidates for the item in the dictionary are sorted through their retrieved number.

3.2 Feedback learning

Though pseudo-relevance feedback (PRF) has been successfully used in the information retrieval (IR), whether PRF in single-language IR or pre-translation PRF and post-translation PRF in CLIR, the feedback results are from source language to source language or target language to target language, that is, the language of feedback units is same as the retrieval language. Our novel is that the input language (Chinese) is different from the feedback target language (English), that is, realizing the feedback from source language to target language, and this feedback technique is also first applied to the term mining field.

After the expansion of semantic prediction, the predicted meaning of an item has some deviations with its actual sense, so the retrieved documents are perhaps not our expected results. In this paper, a PRF technique is employed to acquire more accurate, semantically relevant terms. At first, we collect top-20 documents from search results after term expansion, and then select target language units from these documents, get language units from these documents, which are highly related with the original query in source language. However, how to effectively select these units is a challenging issue. In the literature, researchers have proposed different methods such as Rocchio’s method or Robertson’s probabilistic method to solve this problem. After some experimental comparisons, a simple evaluation method using term frequency and average length is presented in this paper. The evaluation method is defined as follows:

$$w(t) = f(t) + \frac{1}{\Delta(t)+1}, \text{ where } \Delta(t) = \frac{\sum_{i=1}^{N} D_i(s,t)}{N} \quad (1)$$

$\Delta(t)$ represents the average length between the source word s and the target candidate t. If the greater that the average length is, the relevance degree between source terms and candidates will become lower. The purpose of adding $\Delta(t)$ to 1 is to avoid the divide overflow in the case that the average length is equal to zero. $D_i(s,t)$ denotes the byte distance between source words and target candidates, and N represents the total number of candidate occurrences in the estimated Web pages. This evaluation method is very suitable for the discrimination of these words with lower, but same term frequencies. In the ranked candidates after PRF feedback, top-5 candidates are selected as our refined expansion items. In the previous example, the refined expansion items are: Kingdoms, Three, Romance, Chinese, Traditional. These refined expansion terms, together with the original query, "三国演义"+(Kingdoms | Three | Romance | Chinese | Traditional) are once more sent to retrieve relevant results, which are viewed as effective Web pages used in the process of the following estimation.

4 Translation candidate construction and noise solution

The goal of translation candidate construction is to construct and mine all kinds of possible translation forms of terms from the Web, and effectively estimate their feature information such as frequency and distribution. In the transferred text, we locate the position of a query keyword, and then obtain a 100-byte window with keyword as the center. In this window, each English word is built as a beginning index, and then string candidates are constructed with the increase of string in the form of one English word unit. String candidates are indexed in the database with hash and binary search method. If there exists the same item as the inputted candidate, its frequency is increased by 1, otherwise, this candidate is added

to this position of the database. After handling one Web page, the distribution information is also estimated at the same time. In the programming implementation, the table of stop words and some heuristic rules of the beginning and end with respect to the keyword position are employed to accelerate the statistics process.

The aim of noise solution is to remove these irrelevant items and redundant information formed in the process of mining. These noises are defined as the following two categories.

1) Subset redundancy. The characteristic is that this item is a subset of one item, but its frequency is lower than that item. For example, "车牌号：License plate number (6), License plate (5)", where the candidate "License plate" belongs to subset redundancy. They should be removed.

2) Affix redundancy. The characteristic is that this item is the prefix or suffix of one item, but its frequency is greater than that item. For example, 1. "三国演义: Three Kingdoms (30), Romance of the Three Kingdoms (22), The Romance of Three Kingdoms (7)", 2. "蓝筹股: Blue Chip (35), Blue Chip Economic Indicators (10)". In Example 1, the item "Three Kingdoms" is suffix redundancy and should be removed. In Example 2, the term "Blue Chip" is in accord with the definition of prefix redundancy information, but this term is a correct translation candidate. Thus, the problem of affix redundancy information is so complex that we need an evaluation method to decide to retain or drop the candidate.

To deal with subset redundancy and affix redundancy information, sort-based subset deletion and mutual information methods are respectively proposed. More details refer to our previous paper (Fang et al., 2005).

5 Candidate evaluation based on multi-features

5.1 Possible features for translation pairs

Through analyzing mass Web pages, we obtain the following possible features that have important influences on term translation mining. They include: 1) candidate frequency and its distribution in different Web pages, 2) length ratio between source terms and target candidates (S-T), 3) distance between S-T, and 4) keywords, key symbols and boundary information between S-T.

1) Candidate frequency and its distribution

Translation candidate frequency is the most important feature and is the basis of decision-making. Only the terms whose frequencies are greater than a certain threshold are further considered as candidates in our system. Distribution feature reflects the occurrence information of one candidate in different Webs. If the distribution is very uniform, this candidate will more possibly become as the translation equivalent with a greater weight. This is also in accord with our intuition. For example, the translation candidates of the term "认股期权" include "put option" and "short put", and their frequencies are both 5. However, their distributions are "1, 1, 1, 1, 1" and "2, 2, 1". The distribution of "put option" is more uniform, so it will become as a translation candidate of "认股期权" with a greater weight.

2) Length ratio between S-T

The length ratio between S-T should satisfy certain constraints. Only the word number of a candidate falls within a certain range, the possibility of becoming a translation is great.

To estimate the length ratio relation between S-T, we conduct the statistics on the database with 5800 term translation pairs. For example, when Chinese term has three characters, i.e. W=3, the probability of English translations with two words is largest, about P(E=2 |W =3)= 78%, and there is nearly no occurrence out of the range of 1-4. Thus, different weights can be impacted on different candidates by using statistical distribution information of length ratio. The weight contributing to the evaluation function is set according to these estimated probabilities in the experiments.

3) Distance between S-T

Intuitively, if the distance between S-T is longer, the probability of being a translation pair will become smaller. Using this knowledge we can alleviate the effect of some noises through impacting different weights when we collect possible correct candidates far from the source term.

To estimate the distance between S-T, experiments are carried on 5800*200 pages with 5800 term pairs, and statistical results are depicted as the histogram of distances in Figure 4.

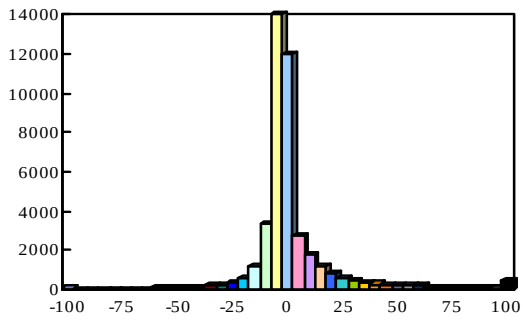

Figure 4. The histogram of distances between S-T

In the figure, negative value represents that English translation located in front of the Chinese term, and positive value represents English translation is behind the Chinese term. As shown from the figure, we know that most candidates are distributed in the range of -60-60 bytes, and few occurrences are out of this range. The numbers of translations appearing in front of the term and after the term are nearly equal. The curve looks like Gaussian probability distribution, so Gaussian models are proposed to model it. By the curve fitting, the parameters of Gaussian models are obtained, i.e. u=1 and sigma=2. Thus, the contribution probability of distance to the ranking function is formulized as $p_D(i,j) = \frac{1}{2\sqrt{2\pi}} e^{-(D(i,j)-1)^2/8}$, where D(*i*,*j*) represents the byte distance between the source term *i* and the candidate *j*.

4) Keywords, key symbols and boundary information between S-T

Some Chinese keywords or capital English abbreviation letters between S-T can provide an important clue for the acquisition of possible correct translations. These Chinese keywords include the words such as "中文叫，中文译为，中文名称，中文名称为，中文称为，或称为，又称为，英文叫，英文名为，英文称为，英文全称". The punctuations between S-T can also provide very strong constraints, for example, when the marks "（）( ) [ ]" exist, the probability of being a translation pair will greatly increase. Thus, correctly judging these cases can not only make translation finding results more comprehensive, but also increase the possibility that this candidate is as one of correct translations. Boundary information refers to the fact that the context of candidates on the Web has distinct mark information, for example, the position of transition from continuous Chinese to English, the place with bracket ellipsis and independent units in the HTML text.

5.2 Candidate evaluation method

After translation noise handling, we evaluate candidate translations so that possible candidates get higher scores. The method in the weighted sum of multi-features including: candidate frequency, distribution, length ratio, distance, keywords, key symbols and boundary information between S-T, is proposed to rank the candidates. The evaluation method is formulized as follows:

$$Score(t) = p_L(s,t)\sum_{i=1}^{N}[\lambda_1 \sum_j (p_D(i,j) + \delta(i,j)w) + \lambda_2 \max_j (p_D(i,j) + \delta(i,j)w)], \ \lambda_1 + \lambda_2 = 1 \quad (2)$$

In the equation, *Score(t)* is proportional to $p_L(s,t)$, *N* and $p_D(i,j)$. If the bigger these component values are, the more they contribute to the whole evaluation formula, and correspondingly the candidate has higher score. The length ratio relation $p_L(s,t)$ reflects the proportion relation between S-T as a whole, so its weight will be impacted on the *Score(t)* in the macro-view. The weights are trained through a large amount of technical terms and proper nouns, where each relation corresponds to one probability. *N* denotes the total number of Web pages that contain candidates, and partly reflects the distribution information of candidates in different Web pages. If the greater *N* is, the greater *Score(t)* will become. The distance relation $p_D(i,j)$ is defined as the distance contribution probability of the *j*th source-candidate pair on the *i*th Web pages, which is impacted on every word pair emerged on the Web in the point of micro-view. Its calculation formula is defined in Section 5.1. The weights of λ_1 and λ_2 represent the proportion of term frequency and term distribution, and λ_1 denotes the weight of the total number of one candidate occurrences, and λ_2 represents the weight of counting the nearest distance occurrence for each Web page. $\delta(i,j)w$ is the contribution probability of keywords, key symbols and boundary information. If there are predefined keywords, key symbols, and boundary information between S-T, i.e., $\delta(i,j) = 1$, then the evaluation formula will give a reward *w*, otherwise, $\delta(i,j) = 0$ indicate that there is no impact on the whole equation.

6 Experiments

Our experimental data consist of two sets: 400 C-E term pairs and 3511 C-E term pairs in the financial domain. There is no intersection between the two sets. Each term often consists of 2-8 Chinese characters, and the associated translation contains 2-5 English words. In the test set of 400 terms, there are more than one English translation for every Chinese term, and only one English translation for 3511 term pairs. In the test sets, Chinese terms are inputted to our system on batch, and their corresponding translations are viewed as a criterion to evaluate these mined candidates. The top n accuracy is defined as the

percentage of terms whose top n translations include correct translations in the term pairs. A series of experiments are conducted on the two test sets.

Experiments on the number of feedback pages: To obtain the best parameter of feedback Web pages that influence the whole system accuracy, we perform the experiments on the test set of 400 terms. The number of feedback Web pages is respectively set to 0, 10, 20, 30, and 40. N=1, 3, 5 represent the accuracies of top 1, 3, and 5. From the feedback pages, previous 5 semantically-relevant terms are extracted to construct a new query expansion for retrieving more effective Web pages. Translation candidates are mined from these effective pages, whose accuracy curves are depicted in Figure 5.

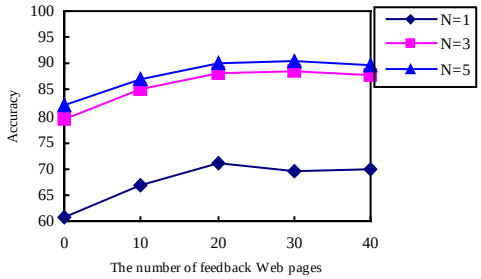

Figure 5. The number of feedback Web pages

As seen from the figure above, when the number of feedback Web pages is 20, the accuracy reaches the best. Thus, the feedback parameter in our experiments is set to 20.

Experiments on the parameter λ_1: In the candidate evaluation method using multi-features, the parameter of λ_1 need be chosen through the experiments. To obtain the best parameter, the experiments are set as follows. The accuracy of top 5 candidates is viewed as a performance criterion. The parameters are respectively set from 0 to 1 with the increase of 0.1 step. The results are listed in Figure 6. As seen from the figure, λ_1=0.4 is best parameter, and therefore λ_2 =0.6. In the following experiments, the parameters are all set to this value.

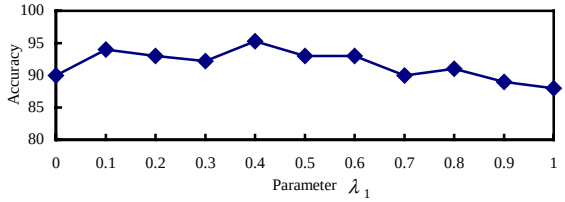

Figure 6. The relation between the parameter λ_1 and the accuracy

Experiments on the test set of 400 terms using different methods: The methods respectively without prediction(NP), with prediction(P), with prediction and feedback(PF) only using term frequency (TM), and with prediction and feedback using multi-features(PF+MF) are employed on the test set of 400 terms. The results are listed in Table 1. As seen from this table, if there is no semantic prediction, the obtained translations from Web pages are about 48% in the top 30 candidates. This is because general search engines will retrieve more relevant Chinese Web pages rather than those effective pages including English meanings. Thus, the semantic prediction method is employed. Experiments demonstrate the method with semantic prediction distinctly improves the accuracy, about 36.8%. To further improve the performance, the feedback learning technique is proposed, and it increases the average accuracy of 6.5%. Though TM is very effective in mining the term translation, the multi-feature method fully utilizes the context of candidates, and therefore obtains more accurate results, about 92.8% in the top 5 candidates.

Table 1. The term translation results using different methods

	Top30	**Top10**	**Top5**	**Top3**	**Top1**
NP	48.0	47.5	46.0	44.0	28.0
P	84.8	83.3	82.3	79.3	60.8
PF+TM	91.3	90.8	90.3	88.3	71.0
PF+MF	95.0	94.5	92.8	91.5	78.8

Experiments on a large vocabulary: To validate our system performance, experiments are carried on a large vocabulary of 3511 terms using different methods. One method is to use term frequency (TM) as an evaluation criterion, and the other method is to use multi-features (MF) as an evaluation criterion. Experimental results are shown as follows.

Table 2. The term translation results on a large vocabulary

	Top30	**Top10**	**Top5**	**Top3**	**Top1**
TM	82.5	81.2	78.3	73.5	49.4
MF	89.1	88.4	86.0	82.9	58.2

From Table 2, we know the accuracy with top 5 candidates is about 86.0%. The method using multi-features is better than that of using term frequency, and improves an average accuracy of 7.94%

Some examples of acquiring English translations of Chinese terms are provided in Table 3.

Only top 3 English translations are listed for each Chinese term.

Table 3. Some C-E mining examples

Chinese terms	The list of English translations (Top 3)
三国演义	The Three Kingdoms The Romance of the Three Kingdoms The Romance of Three Kingdoms
三好学生	Merit student "Three Goods" student Excellent League member
蓝筹股	Blue Chip Blue Chips Blue chip stocks
白朗峰	Mont Blanc Mont-Blanc Chamonix Mont-Blanc
百慕大三角	Burmuda Triangle Bermuda Triangle The Bermuda Triangle
车牌号	License plate number Vehicle plate number Vehicle identification no

7 Conclusions

In this paper, the method based on semantic prediction is first proposed to acquire effective Web pages. The proposed method predicts possible meanings according to each constituent unit of Chinese term, and expands these items for searching using semantically relevant knowledge, and then the refined related terms are extracted from top retrieved documents through feedback learning to construct a new query expansion for acquiring more effective Web pages. For obtaining a correct translation list, the translation evaluation method using multi-features is presented to rank these candidates. Experimental results show that this method has good performance in Chinese-English translation acquisition, about 82.9% accuracy in the top 3 candidates.

References

P.J. Cheng, J.W. Teng, R.C. Chen, et al. 2004. Translating unknown queries with web corpora for cross-language information retrieval, Proc. ACM SIGIR, pp. 146-153.

G.L. Fang, H. Yu, and F. Nishino. 2005. Web-Based Terminology Translation Mining, Proc. IJCNLP, pp. 1004-1016.

P. Fung. 1997. Finding terminology translations from nonparallel corpora, Proc. Fifth Annual Workshop on Very Large Corpora (WVLC'97), pp. 192-202.

F.C. Gey. 2004. Chinese and Korean topic search of Japanese news collections, In Working Notes of the Fourth NTCIR Workshop Meeting, Cross-Lingual Information Retrieval Task, pp. 214-218.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the Web as corpus, Computational Linguistics, 29(3): 333-348.

H. Li, Y. Cao, and C. Li. 2003.Using bilingual web data to mine and rank translations, IEEE Intelligent Systems, 18(4): 54-59.

W.H. Lu, L.F. Chien, and H.J. Lee. 2004. Anchor text mining for translation of Web queries: A transitive translation approach, ACM Trans. Information System, 22(2): 242-269.

M. Nagata, T. Saito, and K. Suzuki. 2001. Using the web as a bilingual dictionary, Proc. ACL 2001 Workshop Data-Driven Methods in Machine Translation, pp. 95-102.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora, Proc. 37th Annual Meeting Assoc. Computational Linguistics, pp. 519-526.

H.C. Seo, S.B. Kim, H.G. Lim and H.C. Rim. 2004. KUNLP system for NTCIR-4 Korean-English cross language information retrieval, In Working Notes of the Fourth NTCIR Workshop Meeting, Cross-Lingual Information Retrieval Task, pp. 103-109.

Y. Zhang and P. Vines. 2004. Using the web for automated translation extraction in cross-language information retrieval, Proc. ACM SIGIR, pp. 162-169.