Web-Based Terminology Translation Mining

Gaolin Fang, Hao Yu, and Fumihito Nishino

Fujitsu Research and Development Center, Co., LTD. Beijing 100016, China
{glfang, yu, nishino}@frdc.fujitsu.com

Abstract. Mining terminology translation from a large amount of Web data can be applied in many fields such as reading/writing assistant, machine translation and cross-language information retrieval. How to find more comprehensive results from the Web and obtain the boundary of candidate translations, and how to remove irrelevant noises and rank the remained candidates are the challenging issues. In this paper, after reviewing and analyzing all possible methods of acquiring translations, a feasible statistics-based method is proposed to mine terminology translation from the Web. In the proposed method, on the basis of an analysis of different forms of term translation distributions, character-based string frequency estimation is presented to construct term translation candidates for exploring more translations and their boundaries, and then sort-based subset deletion and mutual information methods are respectively proposed to deal with subset redundancy information and prefix/suffix redundancy information formed in the process of estimation. Extensive experiments on two test sets of 401 and 3511 English terms validate that our system has better performance.

1 Introduction

The goal of Web-based terminology translation mining is to mine the translations of terminologies 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. Mont Blanc→万宝龙，白朗峰). Translators and technical researchers cannot yet obtain an accurate translation after many lookup efforts when they encounter terminology or proper noun during translating or writing foreign language. According to Web statistics by Google, 76.59% of Web pages are English. In China, statistical results by China Internet Network Information Center in July 2004 show that the number of Internet users has reached 94 million, and nearly 87.4% of users have educational backgrounds beyond high school. These users can smoothly read general English pages, but some terminologies in the Web hamper them to exactly understand the whole content. Some skilled users perhaps resort to a Web search engine, but they cannot obtain effective information from a large amount of retrieved irrelevant pages and redundancy information. Thus, it is necessary to provide a system to automatically mine translation knowledge of terms or proper nouns using abundant Web information so as to help users accurately read or write foreign language.

The system of Web-based terminology translation mining has many applications. 1) Reading/writing assistant, as one part of computer-assisted language learning (CALL) used in the E-learning. During reading or writing, users often meet terms

whose translations cannot be found in the dictionary, but this system can help them mine native and accurate translations from the Web. 2) The tool for constructing bilingual dictionary. The system can not only provide translation candidates for compiling bilingual lexicon, but also evaluate or rescore the candidate list of the dictionary. The constructed dictionary can be further applied in cross-language information retrieval (CLIR) and machine translation. 3) As one of the typical application paradigms of the combination of CLIR and Web mining.

There are some issues that need to be solved using Web information to mine terminology translation: 1) How to find more comprehensive results, i.e. mining all possible forms of annotation pairs in the Web. 2) How to obtain the boundary of candidate translations, especially for the language without the boundary mark such as Chinese and Japanese. Because we don't know the translation is at left or right, and what is between the pair, and where is the candidate endpoint? 3) How to remove the noises formed in the statistics and rank the remained candidates.

On the basis of reviewing all possible methods of acquiring translations, a feasible statistics-based method is proposed to mine terminology translation from the Web. In the proposed method, after analyzing different forms of term translation distributions, character-based string frequency estimation is employed to construct term candidate translations for exploring more translations and their boundaries, and then the candidate noises formed in the process of statistics are defined as two categories: subset redundancy information and prefix/suffix redundancy information. Sort-based subset deletion and mutual information methods are respectively proposed to deal with two redundancy information. Experiments on two test sets of 401 and 3511 English terms show that our system has better performance. In all reported literatures, our experiment is the first time for the extensive research on Web-based terminology translation mining on the largest scale.

2   Related Work

Automatic acquisition of bilingual word pairs or translations has been extensively researched in the literature. The methods of acquiring translations are usually summarized as four categories: 1) acquiring translation from parallel corpora, 2) acquiring translation from a combination of translations of constituent words, 3) acquiring translation from bilingual annotation in the Web, and 4) acquiring translation from non-parallel corpora.

1) Acquiring translation from parallel corpora.

Acquiring bilingual lexicon or translations from parallel corpora (including sentence alignment and paragraph alignment) is to utilize statistics information such as co-occurrence, position, and length between source word and translation equivalence in parallel texts as an evaluation criterion to obtain one-to-one map word pairs. Many previous researches focused on extracting bilingual lexicon from parallel corpora, and readers can refer to the reviews [1], [2] for the details. However, due to the restriction of current available parallel corpora of different languages, together with the fact that corpus annotation requires a lot of manpower and resources, researchers have attempted to extract translations from non-parallel corpus or Web data. As opposed to extracting from parallel corpora, there are no corresponding units in non-parallel

corpora so that statistics information such as co-occurrence, position and length become unreliable. New statistical clues have to be proposed to build the relationship for acquiring translation pairs from non-parallel corpora, which is more difficult to handle than in parallel corpora.

2) Acquiring translation from a combination of translations of constituent words.
Grefenstette [3] employed an example-based approach to obtain compound word translations. His method first combined possible translations of each constituent, and then searched them in WWW, where the retrieved number was viewed as an evaluation criterion. Experiments on a set of 724 German words and a set of 1140 Spanish terms showed that the accuracies of English translations were about 87% and 86%, respectively.

Cao and Li [4] proposed a dictionary-based translation combination method to collect translation candidates of English base noun phrases, and then employed a naive Bayesian classifier and TF-IDF vector constructed with EM algorithm as evaluation criterions for translation selection. In an experiment with 1000 English base noun phrases, the coverage of acquiring translations was 91.4%, and the accuracy of top 3 choices was 79.8%. The system was further improved in the literature [5].

Navigli et al. [6] proposed an ontology learning method for acquiring terminology translations from English to Italian. His method was based on bilingual lexicon and semantic relation between the constituents of source language derived from ontology learning, where disambiguated terms dramatically reduced the number of alternative translations and their combinations. This system can automatically extract the translations of 405 complex terms in the tourism domain.

Using the translation combination of each constituent to acquire the translation of a multiword term is very suitable for translation acquisitions of base noun phrases. However, terminologies and technical terms often consist of unknown words, and their translations are seldom the combination of each constituent. Thus, the result of direct combination is not very desirable for terminology translation acquisition.

3) Acquiring translation from bilingual annotation in the Web.
Nagata et al. [7] proposed an empirical function of the byte distance between Japanese and English terms as an evaluation criterion to extract the translation of Japanese word, and their results could be used as a Japanese-English dictionary. Preliminary experiments on the 50 word pairs showed that an accuracy of top 50 candidates reached 56%. The reasons for such experimental results have two aspects: first, the system didn't further deal with candidate noises for mining useful knowledge; second, this system only handled top 100 Web pages retrieved from search engine. In fact, previous 100 Web pages seldom contain effective bilingual annotation information only directly using keyword search rather than imposing other restrictions. Thus, this problem should be further researched for practical applications. Since his research focused on finding English translation given a Japanese term, the segmentation of Japanese could be avoided. However, our problem is to find Chinese equivalent using English term, so we have to cope with how to obtain the correct boundary of Chinese translations. Therefore, the issue and the proposed method in this paper are distinctly different with Nagata's.

4) Acquiring translation from non-parallel corpora.

Acquiring translation from non-parallel corpora is based on the clue that the context of the source term is very similar to that of the target translation in a large amount of corpora. In 1995, Rapp [8] assumed that there is a correlation between the patterns of word co-occurrence in non-parallel texts of different languages, and then proposed a matrix permutation method to match these patterns. However, computational limitation hampered further extension of this method. In 1996, Tannaka and Iwasaki [9] demonstrated how to extract lexical translation candidates from non-aligned corpora using the similar idea. In 1999, this method was developed and improved by Rapp [10]. Rather than computing the co-occurrence relation matrix between one word and all words, the matrix between one word and a small base lexicon are estimated. Experiments on 100 German words indicated that an accuracy of top 1 English translation was 72%, and top 10 was 89%. This system was only suitable for the situation of one word to one word, and didn't further research on the translation acquisition from multiword to multiword.

In 1995, Fung [11] proposed a "context heterogeneity" method to compute the measure similarity between word and its translation for finding translation candidates. In the experiment with 58 English words, an accuracy of 50% is obtained in the top 10 Chinese word candidates. Based on this work, Fung presented the word relation matrix to find the translation pair in 1997 [12]. This method respectively computed the correlation vectors between source word and seed word, target word and seed word. In 19 Japanese term test set, the accuracy of English translations reached 30%. In 1998, the method was improved to extend to non-parallel, comparable texts for translation acquisition [13]. This system use TF/IDF as the feature, and different measure functions as the similarity computation between the candidate pair. However, the system was restricted to the assumption that there are no missing translations and all translations are included in the candidate word list.

Shahzad et al. [14] first extracted the sentence corpora that are likely to contain the target translation using bilingual dictionary and transformation table. And then, the heuristics method was employed to obtain the correct candidate by analyzing the relations of source compound nouns and using partial context information. Experiments on the 10 compound nouns showed that the average accuracy and recall were respectively 34% and 60%.

As shown from the current situation of translation acquisition from non-parallel corpora, all experiments above are basically performed on small-scaled word set, and their results are very inspiring but difficult to put into practical use. Furthermore, most experimental methods are only suitable for one word translation, i.e. the word number ratio of translation pair is on a basis of 1:1. Thus, there are many issues to be further researched before it is used to explore new translation in the application area.

From the review above, we know that Method 1 requires a large number of parallel corpora, and Method 2 and Method 4 have some limitations when they are applied to acquire the terminology translation, and Method 3 makes the best of mass Web resources and is a feasible approach. When people use Asia language such as Chinese, Japanese, and Korean to write, especially scientific article or technical paper, they often annotate the associated English meaning after the terminology. 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, Method 3 is a feasible way to solve the terminology translation acquisition, which is also validated by the following experiments.

3   The Framework of the Terminology Translation Mining System

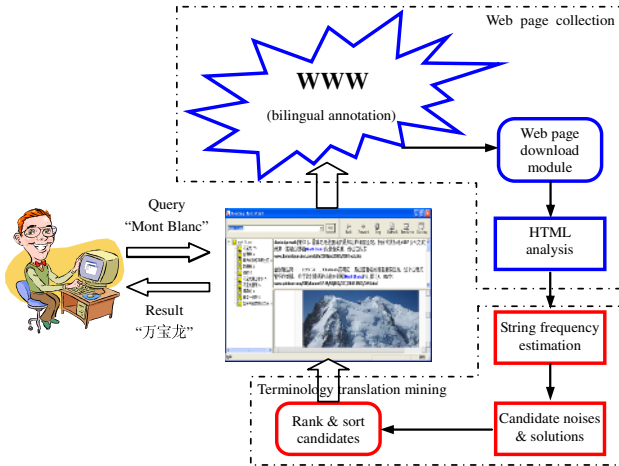The Web-based terminology translation mining system is depicted in Fig. 1 as follows:

Fig. 1. The Web-based terminology translation mining system

The system consists of two parts: Web page collection and terminology translation mining. Web page collection includes download module and HTML analysis module. The function of download module is to collect these Web pages with terms' associated bilingual annotations, and then the pages are inputted into HTML analysis module. In HTML analysis, Web pages are built as a tree structure from which possible features for the bilingual pair and text information in the HTML page are simultaneously extracted.

Terminology translation mining includes string frequency estimation, candidate noises and their solutions, and rank & sort candidates. Translation candidates are constructed through string frequency estimation module, and then we analyze their noises and propose the corresponding methods to handle them. At last, the approach combining the possible features such as frequency, distribution, length proportion, distance, keywords and key symbols is employed to rank these candidates.

In Web pages, there are a variety of bilingual annotation forms. Correctly exploring all kinds of forms can make the mining system extract the comprehensive translation results. After analyzing a large amount of Web page examples, we summarize translation distribution forms as the following six categories: 1) Direct annotation 2) Separate annotation 3) Subset form 4) Table form 5) List form 6) Explanation form. Direct annotation is the most widely used form in the Web, where English meaning often

follows after Chinese terminology, and some have symbol marks such as bracket parentheses and bracket, and some have nothing, e.g. "白朗峰Mont Blanc". Separate annotation is referred to as the case that there are some Chinese words or English letters between the translation pair, e.g. "万能寿险,英文称universal life insurance". Subset form is that the extracted translation pair is a subset of existing bilingual pair, for example, during searching the term "Mont Blanc", the term pair "夏蒙尼·勃朗峰 (Chamonix Mont Blanc)" also provides the valid information. Table or list form is the Web page in the form of table or list. Explanation form is the explanation and illustration for technical terms.
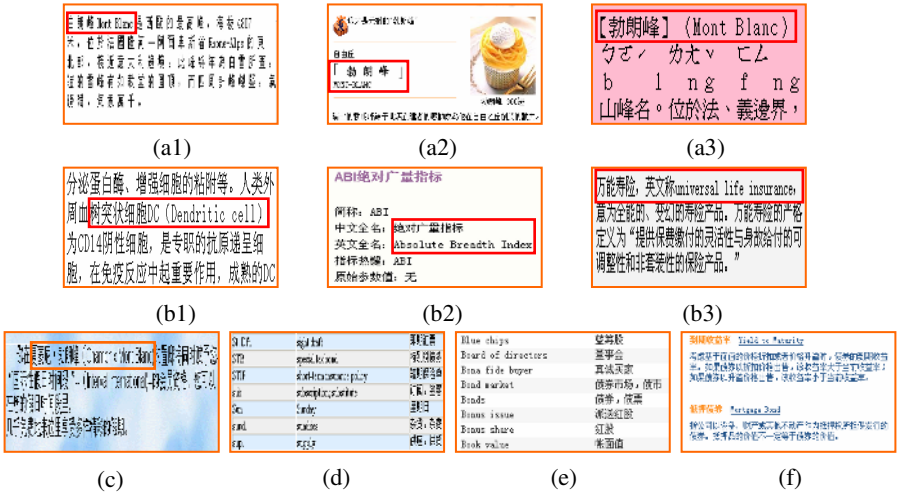

Fig. 2. The examples of translation distribution forms, (a) Direct annotation, some has no mark (a1), and some have some symbol marks (a2, a3) (b) Separate annotation, there are English letters (b1) or some Chinese words (b2, b3) between the translation pair (c) Subset form (d) Table form (e) List form (f) Explanation form

4   Statistics Based Translation Finding

4.1   Character-Based String Frequency Estimation

All kinds of possible translation forms of terminologies in the Web can be effectively and comprehensively mined through character-based string frequency estimation. The proposed method with Chinese character as the basic unit of statistics can not only obtain the correct boundary of the translation candidate, but also conveniently explore these Chinese candidate terminologies that usually consist of unknown words or unknown compound words.

String frequency information is one of the important clues during extracting candidate translations. Its estimation method has a direct influence on the system performance efficiency. The method combing hash index and binary search is employed to

construct the index for all translation candidates. The definition of hash function is calculated according to 6763 Chinese characters in GB2312 system with a one-to-one map. Hash function is formulized as:

$$Y = \begin{cases} 94(c_0 - 176) + (c_1 - 161) & 215 \geq c_0 \geq 176 \\ 94(c_0 - 176) + (c_1 - 161) - 5 & c_0 > 215 \\ 6763 & otherwise \end{cases}, \qquad (1)$$

where c_0, c_1 are respectively the unsigned encoding values of the first, second bytes of first Chinese character of candidate items. All strings are partitioned into different blocks in terms of the first Chinese character with the hash function above, where the strings with the same first character are sorted by lexicographic order, and the strings with non-Chinese character as the first position are indexed to the value of 6763. Here, GB2312 is employed as our statistics standard. Other encoding system is converted to the corresponding characters in GB2312, and the characters will be omitted if there is no counterpart. The reasons for this strategy are as follows: 1) terminology seldom consists of rare words out of GB2312, 2) the index space is dramatically reduced using GB2312 rather than the Unicode encoding so as to quicken the estimation speed.

The terminology to be looked up is inputted into search engine, and the relevant Web pages with this term's associated bilingual annotation are collected. Web pages are transformed into text through HTML analysis module. The term position is located as the center point through keyword search, and then string frequency and distribution estimation is performed in a window of 100 bytes. In Web pages, terminologies are often written as different forms because of the effect of noise. For example, the term "Mont Blanc" may be written as "MONT BLANC", "Mont-Blanc", "Mont ??Blanc", and "MontBlanc". For finding different forms of keywords in the Web, the fuzzy string matching approach is proposed. This method takes 26 English letters in the keyword as effective matching symbols, while ignoring the blank space and other symbols. In the matched text, only these English letters are viewed as effective items for comparison. Using this method can effectively locate different forms of terms and therefore obtain comprehensive translation candidates.

The process of string frequency estimation is described as follows. In the windows with keyword as the center, each character is built as a beginning index, and then the string candidates are constructed with the increase of the string in the form of one Chinese character unit. Since terminology translation usually consists of unknown words or compound words, character is employed as the basic unit of statistics rather than word so as to explore these unknown term translations as more as possible. 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 the database at this position. 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 constructed to accelerate the statistics process.

4.2   Translation Noises and Their Solutions

All possible forms of terminology translations can be comprehensively mined after character-based string frequency estimation. However, there are many irrelevant items and redundancy noises formed in the process of mining. These noises are defined as the following two categories.

1) Subset redundancy information. The characteristic of this kind information is that this item is a subset of one item, but its frequency is lower than that item. For example: "Mont Blanc万宝龙(38) 万宝(27) 宝龙(11)", where "万宝", "宝龙" belong to subset redundancy information. They should be removed.

2) Prefix/suffix redundancy information. The characteristic of this kind information is that this item is the prefix or suffix of one item, but its frequency is greater than that item. For example: 1. "Mont Blanc 朗峰(16) 白朗峰(9) 勃朗峰(8)", 2. "Credit Rating 信用(12) 信用等级(10)", 3. "Knowledge Portal 知识门户(33) 企业知识门户(30)". In Example 1, the item "朗峰" is suffix redundancy information and should be removed. In Example 2, the item "信用" is prefix redundancy information and should also be removed. In Example 3, the term "知识门户" is in accord with the definition of suffix redundancy information, but this term is a correct candidate. Thus, the problem of prefix/suffix redundancy information is so complex that we need an evaluation method to decide to retain or drop this candidate.

```
1.    Sort by entropy value
2.    Sort by boundary[*] for the same entropy
3.    Sort by length and lexical sort for the same entropy and boundary
4.    int nNum = 0;   //record the number of remained candidates
5.    for(int i=0; i<m_nDataNum; i++)  {
6.         int nIsSubString = FALSE;
7.         if(nNum == 0)   //for the first item to be remained
8.              Judge whether to remain this item using boundary and length proportion
                 information;
9.         else  {
10.             for(int j=0; j< nNum; j++)  {
11.                  Judge if the ith candidate is a subset of the jth, and doesn't emerge in
                      the isolated form, if yes
12.                  {  nIsSubString = TRUE;   break;   }
13.             }
14.        }
15.        if(!nIsSubString)  {
16.             Move the ith candidate information to nNum position, and save;
17.             The saved number nNum++;
18.        }
19.   }
20.   m_nDataNum = nNum; //Save the total number.
[*]Note: refer to the case that the string has the distinct left and right boundary in the Web
```

Fig. 3. The description of the sort-based subset deletion algorithm

4.2.1   Sort-Based Subset Deletion Method

Aiming at subset redundancy information, we propose sort-based subset deletion method to handle it. Because subset redundancy information is an intermediate of estimating terminology translations, its information is basically contained by the

longer string candidate with higher frequency. Therefore, this problem can be well solved by first sorting and then judging if this item is a subset of the preceding candidates. The detailed algorithm is described in Fig. 3.

4.2.2 Mutual Information Based Method

Prefix/suffix redundancy information is very complicated to deal with. In some cases, previous candidate is a correct translation and should be retained, while in other cases, it is a noise and should be deleted. In this paper, mutual information based method is proposed to decide if the candidate should be retained or deleted.

The concept of information entropy is first proposed by Shannon in 1948. Entropy is a measure of uncertainty of a random variable, and defined as:

$$H(X) = -\sum_{i=1}^{k} p(x_i) \log_2 p(x_i),$$
(2)

where $p(x_i)$ is a probability function of a random variable $X=x_i$.

Mutual information is a concept of information theory, and is a measure of the amount of information that one random variable contains about another variable. The mutual information of two events X and Y is defined as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y),$$
(3)

where $H(X)$ and $H(Y)$ are respectively the entropies of the random variables of X and Y, and $H(X,Y)$ is the co-occurrence entropy of X and Y.

Mutual information reflects a closeness degree of the combination of X and Y. If there is no interesting relationship between X and Y, I(X,Y)=0, that is, X and Y are independent each other. If there is a genuine association between X and Y, the co-occurrence of XY will be bigger than the random individual occurrence chance of X or Y, and consequently I>>0. In this case, the possibility as a fixed compound phrase of XY becomes very big. Small mutual information hints that the combination of X and Y is very loose, and therefore there is a great possibility of a boundary between two words X, Y.

String frequency estimation is performed on different Web pages. In each Web page there is more than one occurrence for a candidate translation. Mapping this estimation process to the entropy calculation, we define $p(x_i) = n_i / N$, where n_i denotes the number of a translation candidate in one Web page, and N represents the total number of this candidate. We define k as the number of the estimated Web pages. The calculation of entropy is rewritten as:

$$H(X) = -\sum_{i=1}^{k} \frac{n_i}{N} \log_2 \frac{n_i}{N} = -\frac{1}{N} \sum_{i=1}^{k} n_i \log_2 n_i + \log_2 N.$$
(4)

Through this formula, the candidate entropy can be computed directly rather than after counting all Web data. Therefore, it can reduce the time of statistics.

Entropy can not only reflect the frequency information N, but also the distribution information in different Webs. The higher the frequency is, and the larger the entropy is. If the distribution is more uniform, this entropy value will become bigger. This is also in accord with our intuition.

Given two candidate patterns of t_1, t_2 in the set of translation candidates, $C(t_1) > C(t_2)$, where C denotes the frequency of estimation. For suffix redundancy information, $t_1 = suff(t_2)$; for prefix redundancy information, $t_1 = pref(t_2)$. According to the definition of mutual information, $I(t_2) = H(t_1) + H(t_2 - t_1) - H(t_2)$.

The mutual information based method for prefix/suffix redundancy information is described as follows. First, judge if the condition of $\sum_i C(t_1 t_i)/C(t_1) \geq 0.95$ or $\sum_i C(t_i t_1)/C(t_1) \geq 0.95$ is satisfied, where the candidates $t_1 t_i$ represent the items that do not contained each other in the windows of 10 candidates after the candidate t_1. If the condition is met, then delete t_1. In an example of "Dendritic Cell 细胞(62) 树突状细胞(40) 树突细胞(15) 树枝状细胞(4)", because (40+15+4)/62=0.952>0.95, the candidate "细胞" is deleted. If prefix/suffix redundancy information don't satisfy the condition above, then judge the condition of $\lambda I(t_1) < I(t_2)$, if yes, then delete t_1, otherwise retain it. The value of λ is determined by the experiments, and the following experimental results demonstrate that λ=0.85 is the best parameter.

5   Experiments

Our experimental database consists of two sets of 401 English-Chinese term pairs and 3511 English-Chinese term pairs in the financial domain. There is no intersection between two sets. Each terminology often consists of 1-6 English words, and the associated translation contains 2-8 Chinese characters. In the test set of 401 terms, there are more than one Chinese translation for one English term, and only one Chinese translation for 3511 term pairs. The top n accuracy is defined as the percentage of terms whose top n translations include correct translation in the term pairs.
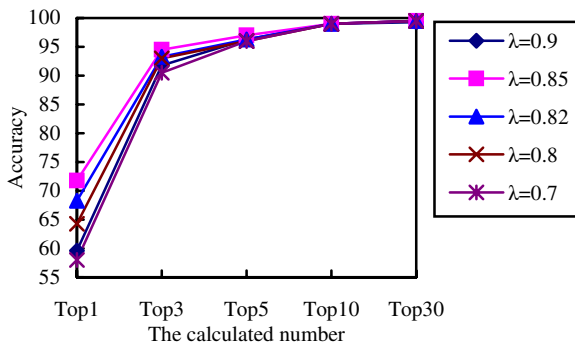

Fig. 4. The relationship between the parameter λ and the accuracy

For testing in what condition, mutual information based method is the best to deal with the prefix/suffix redundancy information. The parameter of λ is respectively set

to 0.7, 0.8, 0.82, 0.85, and 0.9 in the experiment on the test set of 401 terms. Experimental results are shown in Fig. 4. From the figure, we know that λ=0.85 is the best parameter.
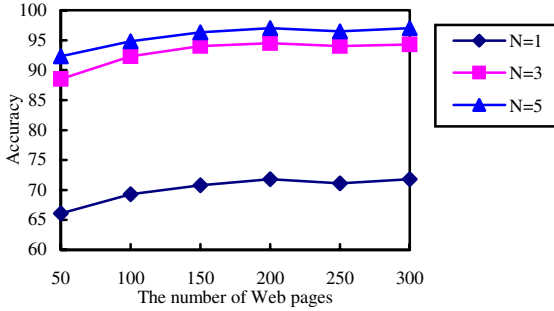

Fig. 5. The relationship between the number of Web pages and the accuracy

A second experiment is to analyze the number of Web pages influencing the term translation accuracy. The experiments are respectively performed on 50, 100, 150, 200, 250, and 300 Web pages retrieved from the Web. Experimental results are illustrated in Fig. 5, where N=1, 3, 5 represent the results of top 1, top 3, and top 5. As seen from the figure, the result of using 200 Web pages is best. When the Web pages increase more than 200 Web pages, the performance isn't improved distinctly, while the computation cost grows. In the case of 200 Web pages, the Chinese translation accuracy of top 1 is 71.8%, and top 3 is 94.5%, and top 5 is 97% on the test set of 401 English terms (see Table 1).

Table 1. Experimental results on a test set of 401 terms

Candidates	Top30	Top10	Top5	Top3	Top1
Accuracy	99.5%	99%	97%	94.5%	71.8%

Using the previous trained parameters, we perform term translation mining experiments in the test set of 3511 terms. Experimental results are listed in Table 2. From this table, the accuracy of top 3 is 83.6%. Experiments also validate that the accuracy of top 30 is nearly equal to the coverage of translations (the percentage of term translations found by our system). This is because there is no change on the accuracy when increasing the candidate number after top 30.

Table 2. Experimental results on a test set of 3511 terms

Candidates	Top30	Top10	Top5	Top3	Top1
Accuracy	95.4%	93.8%	89.1%	83.6%	56.4%

6   Conclusions

In this paper, after reviewing and analyzing all possible methods of acquiring translations, a statistics-based method is proposed to mine terminology translation from the Web. In the proposed method, character-based string frequency estimation is first presented to construct term translation candidates, and then sort-based subset deletion and mutual information methods are respectively proposed to deal with two redundancy information formed in the process of estimation. Experiments on two vocabularies of 401 and 3511 English terms show that our system has better performance, about 94.5% and 83.6% in the top 3 Chinese candidates. The contributions of this paper focus on the following two aspects: 1) On the basis of reviewing all possible methods of acquiring translations and analyzing different forms of term translation distribution, a statistics-based method is proposed to mine terminology translation from the Web. 2) The candidate noises are defined as two categories: subset redundancy information and prefix/suffix redundancy information. Sort-based subset deletion and mutual information methods are respectively proposed to deal with two redundancy information.

References

1. Somers, H.: Bilingual Parallel Corpora and Language Engineering. Proc. Anglo-Indian Workshop "Language Engineering for South-Asian languages", (2001)
2. Véronis, J.: Parallel Text Processing - Alignment and Use of Translation Corpora. The Netherlands: Kluwer Academic Publishers, (2000)
3. Grefenstette, G.: The WWW as a Resource for Example-Based MT Tasks. Proc. ASLIB Translating and the Computer 21 Conference, (1999)
4. Cao, Y., Li, H.: Base Noun Phrase Translation Using Web Data and the EM Algorithm. Proc. 19th Int'l Conf. Computational Linguistics, (2002) 127-133
5. Li, H., Cao, Y., Li, C.: Using Bilingual Web Data to Mine and Rank Translations. IEEE Intelligent Systems. 4 (2003) 54-59
6. Navigli, R., Velardi, P., Gangemi, A.: Ontology Learning and Its Application to Automated Terminology Translation. IEEE Intelligent Systems. 1 (2003) 22-31
7. Nagata, M., Saito, T., Suzuki, K.: Using the Web as a Bilingual Dictionary. Proc. ACL 2001 Workshop Data-Driven Methods in Machine Translation, (2001) 95–102
8. Rapp, R.: Identifying Word Translations in Nonparallel Texts. Proc. 33th Annual Meeting of the Association for Computational Linguistics, (1995) 320-322
9. Tanaka, K., Iwasaki, H.: Extraction of Lexical Translation from Non-Aligned Corpora, Proc. 16th Int'l Conf. Computational Linguistics, (1996) 580-585
10. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. Proc. 37th Annual Meeting Assoc. Computational Linguistics, (1999) 519-526
11. Fung, P.: Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. Proc. Third Annual Workshop on Very Large Corpora, (1995) 173-183

12. Fung, P.: Finding Terminology Translations from Nonparallel Corpora. Proc. Fifth Annual Workshop on Very Large Corpora (WVLC'97), (1997) 192-202
13. Fung P., Yee, L.P.: An IR Approach for Translation New Words from Nonparallel, Comparable Texts. Proc. 17th Int'l Conf. Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, (1998) 414-420
14. Shahzad, I., Ohtake, K., Masuyama, S., Yamamoto, K.: Identifying Translations of Compound Nouns Using Non-Aligned Corpora. Proc. Workshop on Multilingual Information Processing and Asian Language Processing, (1999) 108-113