# Web Translation Mining Based on Suffix Arrays

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

**Abstract**

*Mining translations from abundant Web data can be applied in many fields such as computer assisted learning, machine translation and cross-language information retrieval. How to mine possible translations from the Web and obtain the boundary of candidates, and how to remove irrelevant noises and rank the candidates are the challenging issues. In this paper, after reviewing and analyzing all possible methods of acquiring translations, a statistics method based on suffix arrays is proposed to mine term translations from the Web. The proposed method can not only mine different forms of Web translation distributions but also effectively obtain the correct boundary of translations, and then sort-based subset deletion and mutual information methods are respectively proposed to deal with subset redundancy information and affix redundancy information formed in the process of estimation. Experiments on two test sets of 401 English-Chinese terms and 100 English-Japanese terms validate that our system has good performance.*

**Keywords**

*Web mining, term translation mining, suffix array, mutual information, translation acquisition, reading/writing assistant*

## 1 Introduction

The goal of Web translation mining is to mine the 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. Mont Blanc→万宝龙 , 白朗峰; モンブラン). Translators and technical researchers cannot 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, 77% of Web pages are English. Usually, people can smoothly read general English pages, but some terminologies on the Web hamper them to exactly understand the whole content. Some skilled users may 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 translation mining has many applications. 1) Reading/writing assistant, as one part of computer-assisted language learning (CALL). 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 construction tool of bilingual or multilingual dictionary for machine translation. The system can not only provide translation candidates for compiling a lexicon, but also evaluate or rescore the candidate list of the dictionary. We can also use English as a medium language to build a translation bridge between two languages with few bilingual annotations such as 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.

There are some issues that need to be solved using Web information to mine term translations: 1) How to find more comprehensive results, i.e. mining all possible forms of annotation pairs on 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 statistics method based on suffix arrays is proposed to mine term translations from the Web. The proposed method can not only mine different forms of term translation distributions but also effectively obtain their translation boundaries. And then, the candidate noises formed in the process of statistics are defined as two categories: subset redundancy information and affix redundancy information. Sort-based subset deletion and mutual information methods are respectively proposed to deal with two kinds of redundancy information. Experiments on the Chinese and Japanese language show that our method is a general method to mine the Web translations for similar languages.

The remainder of this paper is organized as follows. In Section 2, we review the related work. Section 3 gives an overview of the system framework. In Section 4, we introduce translation statistics based on suffix arrays, which consists of translation candidate construction using suffix arrays and translation noise solution. Section 5 shows experimental results. The conclusion is drawn in the last section.

## 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 on 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 (Somers 2001; Veronis 2000) 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 (1999) 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 (2003) 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 (Li et al. 2003).

Navigli et al. (2003) 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 on the Web

Nagata et al. (2001) 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 (Fang et al. 2006). 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 or Japanese equivalents using English terms, so we have to cope with how to obtain the correct boundaries of Chinese translations. Therefore, the issue and the proposed method in this paper are distinctly different with Nagata’s.

In our early work (Fang et al. 2005), the method of string frequency estimation is proposed to construct English term translation candidates, and obtains good performance in

English-Chinese term translation mining. However, the method is related with the language, and is difficult to be extended to similar languages.

### 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 (1995) 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 (1996) 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 (1999). 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 (1995) 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 (Fung 1997). 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 (Fung et al. 1998). 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. (1999) 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 term 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 term translation acquisition, which is also validated by the following experiments.

## 3 The framework of the Web translation mining system

The Web-based term translation mining system is depicted in Figure 1 as follows:

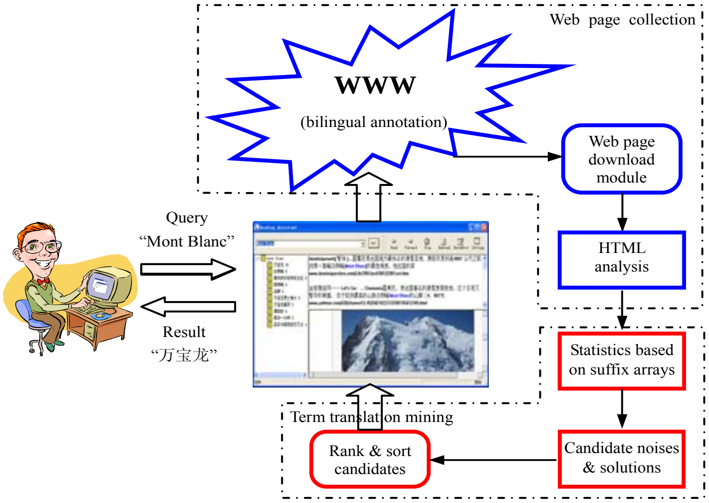
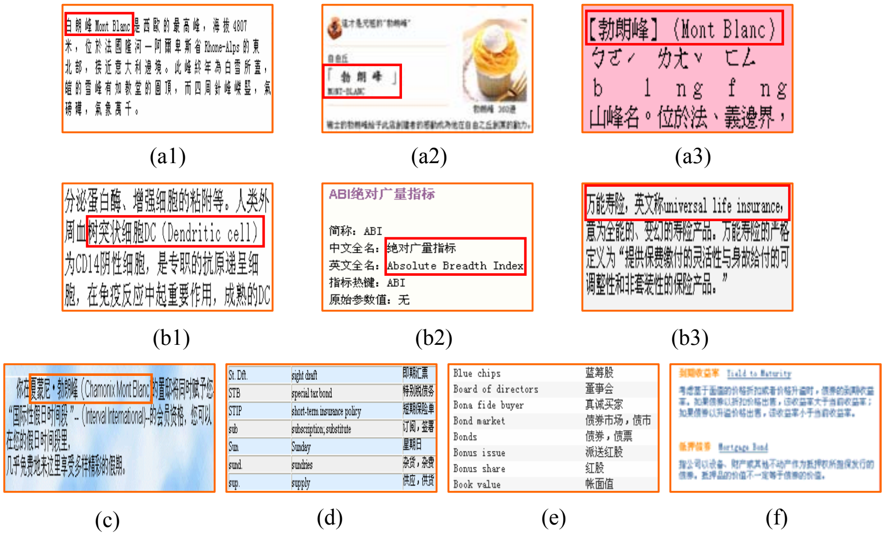


**Figure 1.** The Web-based term translation mining system

The system consists of two parts: Web page collection and term 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.

Term translation mining includes translation statistics based on suffix arrays, candidate noises and their solutions, and rank & sort candidates. Translation candidates are constructed through suffix array statistics 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.

On 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 (Figure 2): 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 on 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.



白朗峰Mont Blanc是西歐的最高峰，海拔4807
米，位於法國隆河一阿爾卑斯省Rhone-Alps的東
北部，接近意大利邊境。此峰終年為白雪所蓋，
皚的雪峰有如教堂的圓頂，而四周針峰嶙豎，氣
磅礴，氣象萬千。

(a1)

「勃朗峰」
MONT-BLANC

(a2)

【勃朗峰】（Mont Blanc）
ㄅㄛˊ ㄌㄤˇ ㄈㄥ
b l ng f ng
山峰名。位於法、義邊界，

(a3)

分泌蛋白酶、增强细胞的粘附等。人类外
周血树突状细胞DC（Dendritic cell）
为CD14阴性细胞，是专职的抗原递呈细
胞，在免疫反应中起重要作用，成熟的DC

(b1)

ABI绝对广量指标

简称：ABI
中文全名：绝对广量指标
英文全名：Absolute Breadth Index
指标热键：ABI
原始参数值：无

(b2)

万能寿险，英文称universal life insurance，
意为全能的、变幻的寿险产品。万能寿险的严格
定义为“提供保费缴付的灵活性与身故给付的可
调整性和非套装性的保险产品。”

(b3)

你在夏蒙尼·勃朗峰（Chamonix Mont Blanc）的置邸将同时赋予您
“国际性假日时间段”--（Interval International）-的会员资格，您可以
在您的假日时间段里，
几乎免费地来这里享受多样精彩的假期。

(c)

| | | |
|---|---|---|
| St. Dft. | sight draft | 即期汇票 |
| STB | special tax bond | 特别税债务 |
| STIP | short-term insurance policy | 短期保险单 |
| sub | subscription; substitute | 订阅，签署 |
| Sun | Sunday | 星期日 |
| sund. | sundries | 杂货，杂费 |
| sup. | supply | 供应，供货 |

(d)

| | |
|---|---|
| Blue chips | 蓝筹股 |
| Board of directors | 董事会 |
| Bona fide buyer | 真诚买家 |
| Bond market | 债券市场，债市 |
| Bonds | 债券，债票 |
| Bonus issue | 派送红股 |
| Bonus share | 红股 |
| Book value | 帐面值 |

(e)

到期收益率 Yield to Maturity
[illegible]

抵押债券 Mortgage Bond
[illegible]

(f)

**Figure 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 Translation statistics based on suffix arrays

### 4.1 Translation candidate construction using suffix arrays

All kinds of possible translation forms of terms on the Web can be effectively and comprehensively mined through translation candidate construction using suffix arrays. Suffix array, as a compact representation of suffix trees (The major advantage of suffix arrays over suffix trees is space, especially more significant for larger alphabets such as Chinese and Japanese characters), is an efficient data structure to compute the frequency and location of substring in a long sequence. The basic idea of suffix arrays is as follows. First, one suffix array is created for the estimated text data. The array will store the pointers to the text suffixes, where each suffix is a string starting at a certain pointer position in the text and ending at the end of the text. And then, the arrays are sorted in lexicographical order, where the frequency and location of substring can be effectively extracted.

Let's get started with the suffix array construction. Suppose the sample text ``后缀数组是一个有效的统计数组". Index points are assigned to the sample text character by character. In our system, using Chinese or Japanese character as the basic unit of statistics can not only obtain the correct boundary of the translation candidate, but also conveniently mine these candidates that usually consist of unknown words or unknown compound words. The suffix array S stores position information, which represents the string from this point to the end of text. For example: S[0] denotes the string "后缀数组是一个有效的统计数组", and S[1] denotes the string "缀数组是一个有效的统计数组". The detail is illustrated in Figure 3.

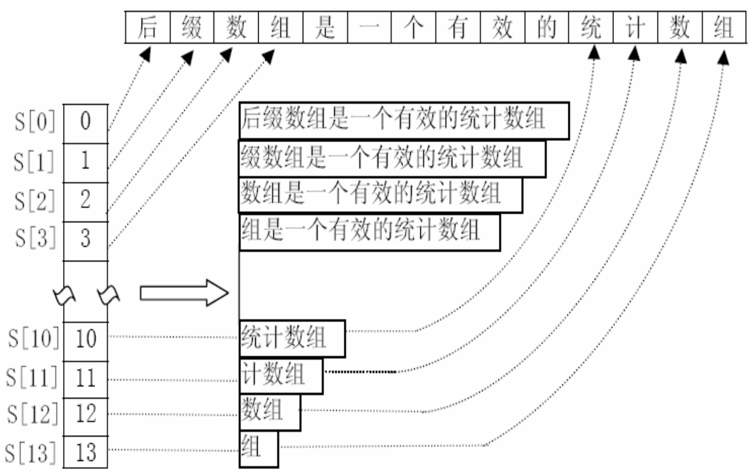


**Figure 3.** The construction of the suffix array

After the process above, the array needs to be sorted in lexicographical order for estimation. Suffix sorting differs from ordinary string sorting because the elements to sort are overlapping strings. Manber and Myers (1993) proposed an elegant radix-sorting based algorithm which takes at most O(nlogn) time. Based on Manber-Myers algorithm, Larsson and Sadakane (1999) proposed a fast and memory efficient algorithm for sorting the suffixes of string. Here, Larsson-Sadakane algorithm is employed in our system. In the sorted arrays, the frequencies and substring boundaries can be easily extracted. For example, for estimating the frequency of "数组", we find that the adjacent storages S[12] and S[2] in the sorted suffix arrays are to use "数组" as prefix, so its frequency is 2. For the frequency of "组", S[13] and S[3] are to use "组" as prefix, and the frequency is also 2.

Statistics is performed on different Web pages, whose contents consist of many segments or sentences. In our system, for utilizing the suffix array, the following method is proposed. The segments from different Web pages are connected with '$', in the segment, all punctuations and boundary information are replaced with '#'. After this, all the segments become an integrated text, so the suffix array can be easily applied to mine the longest and possible terms.

The extraction process of the segment from Web pages is described as follows. Web pages are transformed into text through a HTML analysis module. The term position is located as the center point through fuzzy keyword search. On Web pages, terms 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”, and “MontBlanc”. For finding different forms of keywords on the Web, the fuzzy keyword search 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.

In the 100-byte windows with keyword as the center, blank, boundary and punctuation information of the text are replaced with ‘#’. At the same time, some rules are made to calibrate different forms of translations such as “Edinburgh International Festival（エジンバラ国際フェスティバル, エディンバラ・インターナショナル・フェスティバル）” for estimation.

### 4.2 Translation noises and their solutions

All possible forms of translations can be mined after translation candidate construction using suffix arrays. However, there are many irrelevant items and redundant 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) Affix 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 affix redundancy information is so complex that we need an evaluation method to decide to retain or drop this candidate.

#### 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 term 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 Figure 4. In our algorithm, Line 8 can be further explained: if there are right and left boundaries for one candidate, then this candidate is an integrated unit which isn’t the subset, and will be retained. Let’s give an example how to utilize length information: if there is two-word English term, the Chinese character number of its translation will be the range 3-7 according to our estimation in a large number translation pairs. Thus, if the candidate is out of this range, then it will be dropped.

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

**Figure 4.** The description of the sort-based subset deletion algorithm

### 4.2.2 Mutual information based method

Affix 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:

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

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:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)\,, \tag{2}$$

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 essential 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. On 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 on 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:

$$H(X) = -\sum_{i=1}^{k} \frac{n_i}{N} \log_2 \frac{n_i}{N} = -\frac{1}{N} \sum_{i=1}^{k} n_i \log_2 n_i + \log_2 N \tag{3}$$

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

Entropy can not only reflect the frequency information N, but also the distribution information on 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 affix 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 affix 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 100 English-Japanese term pairs in the financial domain. Each term often consists of 1-6 English words, and the associated translation contains 2-8 Chinese or Japanese characters. In the test set of 401 terms, there are more than one Chinese translation for one English term, and only one Japanese translation for 100 English-Japanese term pairs. The top n accuracy is defined as the percentage of terms whose top n translations include correct translation in the term pairs.

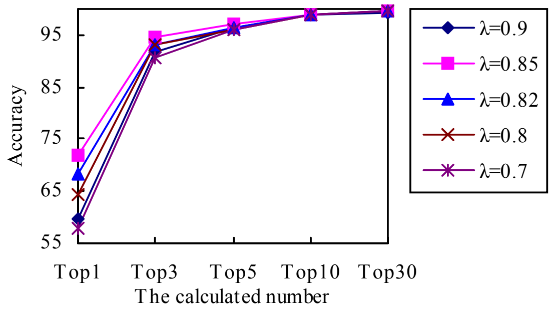


**Figure 5.** The relationship between the parameter λ and the accuracy

For testing in what condition, mutual information based method is the best to deal with the affix 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 Figure 5. From the figure, we know that λ=0.85 is the best parameter.

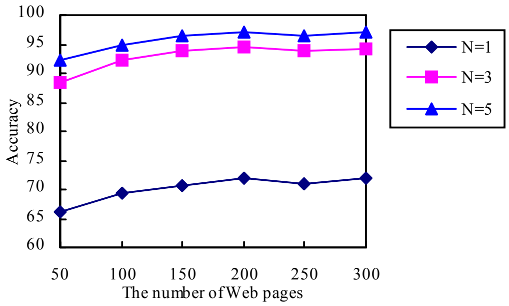


**Figure 6.** The relationship between the number of Web pages and the accuracy

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 Figure 6, 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, we perform the experiments with simple frequency method, sort-based subset deletion method (SBSD), and sort-based subset deletion and mutual information methods (SBSD+MI) on the test set of 401 English-Chinese terms, respectively. Experiments show that simple frequency method is very bad because of many statistics noises. After using SBSD, many redundant subset noises are deleted. The performance is distinctly improved, and increases about 30% in top 3 candidates. To reduce the effect of affix noises, mutual information based method, together with SBSD, is used in our system.

The Chinese translation accuracy of top 1 is 71.8%, and top 3 is 94.5%, and top 5 is 97% (see Table 1).

| Candidates | Top30 | Top10 | Top5 | Top3 | Top1 |
|---|---|---|---|---|---|
| Frequency method | 70.1% | 55.1% | 47.9% | 41.9% | 34.7% |
| SBSD | 90.3% | 85% | 79.1% | 71.8% | 55.9% |
| SBSD+MI | 99.5% | 99% | 97% | 94.5% | 71.8% |

**Table 1. Experimental results on a test set of 401 English-Chinese terms**

Using the previous trained parameters, we perform term translation mining experiments on the test set of 100 English-Japanese terms. Experimental results are listed in Table 2. From this table, the accuracy of top 3 is 65%. There is only one corresponding Japanese translation for English term in our English-Japanese test set. Some possible correct translation didn't count as correct one because it isn't the exact one in the test set. If other form of translation is also viewed as correct one, the accuracy will reach 76% in the top 3 candidates. 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.

| Candidates | Top30 | Top10 | Top5 | Top3 | Top1 |
|---|---|---|---|---|---|
| Accuracy | 90% | 84% | 71% | 65% | 52% |

**Table 2. Experimental results on a test set of 100 English-Japanese terms**

Some examples of acquiring English-Chinese and English-Japanese translations are provided in Table 3.

| **English terms** | **Chinese translations** | **Japanese translations** |
|---|---|---|
| Mont Blanc | 万宝龙，白朗峰，勃朗峰 | モンブラン |
| Agritourism | 农业旅游 | アグリツーリズム |
| Biodiesel | 生物柴油 | バイオディーゼル |
| Mouse potato | 电脑迷，网虫 | マウスポテト |
| Blue Chip | 蓝筹股, 绩优股 | ブルーチップ |
| Dendritic cell | 树突状细胞 | 樹状細胞 |

**Table 3. Some examples of English-Chinese and English-Japanese translation mining**

## 6 Conclusions

In this paper, after reviewing and analyzing all possible methods of acquiring translations, a statistics-based method based on suffix arrays is proposed to mine term translation from the Web. In the proposed method, character-based suffix array estimation is first presented to construct possible term translation candidates, and then sort-based subset deletion and mutual information methods are respectively proposed to deal with two redundancy information: subset redundancy and affix redundancy in the process of estimation.

Experiments on two vocabularies of 401 English-Chinese and 100 English-Japanese terms show that our system has good performance, about 94.5% and 65% in the top 3 candidates. The contributions of this paper focus on the following two aspects: 1) The method for construction possible term translation using character-based suffix array estimation is proposed, and 2) sort-based subset deletion and mutual information methods are respectively presented to deal with two kinds of redundancy information.

## 7 References

Cao, Y. and Li, H., 2002, Base Noun Phrase Translation Using Web Data and the EM Algorithm. Proc. 19th Int'l Conf. Computational Linguistics, pp. 127-133.

Fang, G.L., Yu, H., and Nishino, F., 2005, Web-based Terminology Translation Mining, The Second International Joint Conference on Natural Language Processing (IJCNLP-05), pp. 1004-1016.

Fang, G.L., Yu, H., and Nishino, F., 2006, Chinese-English Term Translation Mining Based on Semantic Prediction, COLING/ACL 2006, pp. 199-206.

Fung, P., 1995, Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus, Proc. Third Annual Workshop on Very Large Corpora, pp. 173-183.

Fung, P., 1997, Finding Terminology Translations from Nonparallel Corpora. Proc. Fifth Annual Workshop on Very Large Corpora (WVLC'97), pp. 192-202.

Fung P. and Yee, L.P., 1998, 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, pp. 414-420.

Grefenstette, G., 1999, The WWW as a Resource for Example-Based MT Tasks, Proc. ASLIB Translating and the Computer 21 Conference.

Larsson, N. and Sadakane, K., 1999, Faster Suffix Sorting. Technical Report LU--CS--TR:99--214, Lund University, pp. 1-20.

Li, H., Cao, Y., and Li, C., 2003, Using Bilingual Web Data to Mine and Rank Translations, IEEE Intelligent Systems, vol. 4, pp. 54-59.

Manber, U. and Myers, E.W., 1993, Suffix Arrays: A New Method for Online String Searches. SIAM Journal on Computing, vol. 22, no.5, pp. 935-948.

Nagata, M., Saito, T., and Suzuki, K., 2001, Using the Web as a Bilingual Dictionary, Proc. ACL 2001 Workshop Data-Driven Methods in Machine Translation, pp. 95–102.

Navigli, R., Velardi, P., and Gangemi, A., 2003, Ontology Learning and Its Application to Automated Terminology Translation, IEEE Intelligent Systems, vol. 1, pp. 22-31.

Rapp, R., 1995, Identifying Word Translations in Nonparallel Texts. Proc. 33th Annual Meeting of the Association for Computational Linguistics, pp. 320-322.

Rapp, R., 1999, Automatic Identification of Word Translations from Unrelated English and German Corpora, Proc. 37th Annual Meeting Assoc. Computational Linguistics, pp. 519-526.

Shahzad, I., Ohtake, K., Masuyama, S., and Yamamoto, K., 1999, Identifying Translations of Compound Nouns Using Non-Aligned Corpora, Proc. Workshop on Multilingual Information Processing and Asian Language Processing, pp. 108-113.

Somers, H., 2001, Bilingual Parallel Corpora and Language Engineering, Proc. Anglo-Indian Workshop "Language Engineering for South-Asian languages".

Tanaka, K. and Iwasaki, H., 1996, Extraction of Lexical Translation from Non-Aligned Corpora, Proc. 16th Int'l Conf. Computational Linguistics, pp. 580-585.

Véronis, J., 2000, Parallel Text Processing - Alignment and Use of Translation Corpora. The Netherlands: Kluwer Academic Publishers.