A Bottom-up Term Extraction Approach for Web-based Translation in Chinese-English IR Systems

Chengye Lu, Yue Xu and Shlomo Geva
School of Software Engineering and Data Communications
Queensland University of Technology
Brisbane, QLD 4001, Australia
(c.lu,yue.xu,s.geva)@qut.edu.au

ABSTRACT

The extraction of Multiword Lexical Units (MLUs) in lexica is important to language related methods such as Natural Language Processing (NLP) and machine translation. As one word in one language may be translated into an MLU in another language, the extraction of MLUs plays an important role in Cross-Language Information Retrieval (CLIR), especially in finding the translation for words that are not in a dictionary.

Web mining has been used for translating the query terms that are missing from dictionaries. MLU extraction is one of the key parts in search engine based translation. The MLU extraction result will finally affect the transition quality.

Most statistical approaches to MLU extraction rely on large statistical information from huge corpora. In the case of search engine based translation, those approaches do not perform well because the size of corpus returned from a search engine is usually small. In this paper, we present a new string measurement and new Chinese MLU extraction process that works well on small corpora.

Keywords

Cross-language Information retrieval, CLIR, query translation, web mining, OOV problem, term extraction

1. INTRODUCTION

As more and more documents written in various languages become available on the Internet, increasingly users wish to explore documents that were written in their native language rather than English. Cross-language information retrieval (CLIR) systems enable users to retrieve documents written in more than one language through a single query. Obviously translation is needed in the CLIR process. The common approach is to translate the query into the document language using a dictionary. Dictionary based translation has been adopted in cross-language information retrieval because bilingual dictionaries are widely available, dictionary based approaches are easy to implement, and the efficiency of word translation with a dictionary is high. However, because of the vocabulary limitation of dictionaries, very often the translations of some words in a query can't be found in a dictionary. This problem is called the Out of Vocabulary (OOV) problem.

The OOV problem usually happens when translating Multiword Lexical Units (MLUs) such as proper names, phrases or newly created words. As the length of input queries is usually short, query expansion does not provide enough information to recover the missing words. Furthermore, very often the OOV terms are key terms in a query. In particular, the OOV terms such as proper names or newly created technical terms carry the most important information in a query. For example, a query "SARS, CHINA" may be entered by a user in order to find information about SARS in China. However SARS is a newly created term and may not be included in a dictionary which was published only a few years ago. If the word SARS is left out of the translated query or translated incorrectly, it is most likely that the user will practically be unable to find any relevant documents at all.

Web mining has been used for OOV term translation [4; 5; 7; 9; 13]. It is based on the observation that there exist web pages which contain more than one language. Investigation has found that, when a new English term such as a new technical term or a proper name is introduced into another language (target language), the translation of this term and the original English term very often appear together in documents written in the target language in an attempt to avoid misunderstanding. Mining this kind of web pages can help discover the translation of the new terms. Some earlier research already addressed the problem of how those kinds of documents can be extracted by using web search engines such as Google and Yahoo. Popular search engines allow us to search English terms for pages in a certain language, e.g., Chinese or Japanese. The result returned by a web search engine is usually a long ordered list of document titles and summaries to help users locate information. Mining the result lists is a way to find translations to the unknown query terms. Some studies[3; 13] have shown that such approaches are rather effective for proper name translation. To distinguish such approaches from other web mining based translation approaches, we call those approaches "search engine based translation approaches". Search engine based approaches usually consist of three steps:

1. Document retrieval: use a web search engine to find the documents in target language that contain the OOV term in original language. For example, when finding the translation of an English term in Chinese, the English term will be put in the search engine and ask the search engine return Chinese result only. Collect the text (i.e. the summaries) in the result pages returned from the web search engine.
2. Term extraction: extract the meaningful terms in the summaries where the OOV term appears. Record the terms and their frequency in the summaries. As a term in one language could be translated to a phrase or even a sentence, the major difficulty in term extraction is how to extract correct MLUs from summaries.
3. Translation selection: select the appropriate translation from the extracted words. As the previous step may produce a long list of terms, translation selection has to find the correct translation from the terms.

Step 2 and step 3 are the core steps of search engine based translation. In this paper, we present our contribution to term extraction and translation selection. Specifically, we introduce a

statistics based approach to extraction of terms to improve the precision of the translations.

2. Previous Work

In this section, we briefly review some statistical based approaches on term extraction. Detailed analysis of these approaches will be given in the evaluation section.

Term extraction is mainly the task of finding MLUs in the corpus. The concept of MLU is important for applications that exploit language properties, such as Natural Language Processing (NLP), information retrieval and machine translation. An MLU is a group of words that always occur together to express a specific meaning. The minimal size of a MLU should be 2. For example, compound nouns like *Disney Land*, compound verbs like *take into account*, adverbial locutions like *as soon as possible*, idioms like *cutting edge*. In most cases, it is necessary to extract MLUs, rather than words, from a corpus because the meaning of an MLU is not always the compositional of each individual word in the MLU. For example, you cannot interprete the MLU 'cutting edge' by combining the meaning of 'cutting' and the meaning of 'edge'.

Finding MLUs from the summaries returned by a search engine is important in search engine based translation because a word in one language may be translated into a phrase or even a sentence. If only words are extracted from the summaries, the later process may not be able to find the correct translation because the translation might be a phrase rather than a word.

For Chinese text, a word consisting of several characters is not explicitly delimited since Chinese text contains sequences of Chinese characters without spaces between them. Chinese word segmentation is the process of marking word boundaries. The Chinese word segmentation is actually similar to the extraction of MLUs in English documents since the MLU extraction in English documents also needs to mark the lexicon boundaries between MLUs. Therefore, term extraction in Chinese documents can be considered as Chinese word segmentation. Many existing systems use lexical based or dictionary based segmenters to determine word boundaries in Chinese text. However, in the case of search engine based translation, as an OOV term is an unknown term to the system, these kinds of segmenters usually cannot correctly identify the OOV terms in the sentence. Incorrect segmentation may break a term into two or more words. Therefore, the translation of an OOV term cannot be found in a later process. Some researchers suggested approaches that are based on co-occurrence statistics model for Chinese word segmentation to avoid this problem [4; 5; 8; 9; 13].

One of the most popular statistics based extraction approaches is to use mutual information [6; 12]. Mutual information is defined as:

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{Nf(x,y)}{f(x)f(y)} \qquad (1)$$

The mutual information measurement quantifies the distance between the joint distribution of terms *X* and *Y* and the product of their marginal distributions [1]. When using mutual information in Chinese segmentation, *x, y* are two Chinese characters; *f(x), f(y), f(x,y)* are the frequencies that *x* appears, *y* appears, and *x* and *y* appear together, respectively; N is the size of the corpus. A string *XY* will be judged as a term if the *MI* value is greater than a predefined threshold.

Chien [6] suggests a variation of the mutual information measurement called significance estimation to extract Chinese keywords from corpora. The significance estimation of a Chinese string is defined as:

$$SE(c) = \frac{f(c)}{f(a) + f(b) - f(c)} \qquad (2)$$

Where *c* is a Chinese string with *n* characters; *a* and *b* are two longest composed substrings of *c* with length *n-1*; *f* is the function to calculate the frequency of a string. Two thresholds are predefined: *THF* and *THSE*. This approach identifies a Chinese string to be a MLU by the following steps. For the whole string c, if f(*c*)>*THF*, *c* is considered a Chinese term. For the two (*n*-1)-substrings *a* and *b* of *c*, if *SE(c)>=THSE*, both *a* and *b* are not a Chinese term. If *SE(c)<THSE*, and *f(a)>>f(b)* or *f(b)>>f(a)* , *a* or *b* is a Chinese term, respectively. Then for each *a* and *b*, the method is recursively applied to determine whether their substrings are terms.

However, all mutual information based approaches have the problem of tuning the thresholds for generic use. Silva and Lopes suggest an approach called Local Maxima to extract MLU from corpora to avoid using any predefined threshold [12]. The equation used in Local Maxima is known as SCP defined as:

$$SCP(s) = \frac{f(s)^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} f(w_1 \ldots w_i) f(w_{i+1} \ldots w_n)} \qquad (3)$$

S is an n-gram string, $w_{1,\ldots},w_i$ is the substring of *S*. A string is judged as an MLU if the SCP value is greater or equal than the SCP value of all the substrings of *S* and also greater or equal than the SCP value of its antecedent and successor. The antecedent of S is an (n-1)-gram substring of *S*. The successor of S is a string that S is its antecedent.

Although Local Maxima should be a language independent approach, Cheng et al.[5] found that it does not work well in Chinese word extraction. They introduced context dependency (CD) used together with the Local Maxima. The new approach is called SCPCD. The rank for a string is using the function:

$$SCPCD(s) = \frac{LC(s)RC(s)}{\frac{1}{n-1}\sum_{i=1}^{n-1} freq(w_1 \ldots w_i) freq(w_{i+1} \ldots w_n)} \qquad (4)$$

S is the input string, *w1..wi* is the substring of *S*, *LC()* and *RC()* are functions to calculate the number of unique left(right) adjacent characters of *S*. A string is judged as a Chinese term if the SCPCD value is greater or equal than the SCPCD value of all the substrings of *S*.

3. PROPOSED APPROACH

The term extraction approaches listed above have been used on large corpus. However, in our experiments, the performance of those approaches is not always satisfactory in search engine based

OOV term translation approaches. As described in introduction, web search engine result pages are used for search engine based OOV term translation. In most cases, only a few hundreds of top results from the result pages are used for translation extraction. As a result, the corpus size for search engine based approaches is quite small. In a small collection, the frequencies of strings very often are too low to be used in the approaches reviewed in Section 2. Moreover, the search engine results are usually part of a sentence, which makes the traditional Chinese word segmentation hard to be applied in this situation. That is why many researchers [4; 5; 7; 9; 13] try to apply statistical based approaches on search engine base translation for term extraction.

In this section, we describe a term extraction approach specifically designed for the search engine based translation extraction, which uses term frequency change as an indicator to determine term boundaries and also uses the similarity comparison between individual character frequencies instead of terms to reduce the impact of low term frequency in small collections. Together with the term extraction approach, we also describe a bottom-up term extraction approach that can help to increase the extraction quality.

3.1 Frequency Change Measurement

The approaches mentioned in Section 2 use a top-down approach that starts with examining the whole sentence and then examining substrings of the sentence to extract MLUs until the substring becomes empty. We propose to use a bottom-up approach that starts with examining the first character and then examines super strings. Our approach is based on the following observations for small document collections:

Observation 1: In a small collection of Chinese text such as a collection of search engine result pages, the frequencies of the characters in a MLU are similar. This is because in a small collection of text, there are a small number of MLUs, the characters appearing in one MLU may not appear in other MLUs. On the other hand, some times MLUs with similar meanings will share similar characters and those characters are unlikely to be used in other unrelated MLUs. For example, 戰機 (Fighter Aircraft) and 戰鬥機 have the same meaning in Chinese. They share similar Chinese characters. Therefore although the term's frequency is low, the individual characters of the term might still have relatively high and also similar frequencies. The high frequency can help in term extraction.

Observation 2: When a correct Chinese term is extended with an additional character, the frequency of the new term very often drops significantly.

According to Observation 1, the frequencies of a term and each character in the term should be similar. We propose to use the root mean square error (RMS error) given in Equation (5) to measure the similarity between the character frequencies.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

For a given Chinese character sequence, x_i is the frequency of each character in the sequence, $\bar{x}$ is the average frequency of all the characters in the sequence. Although the frequency of a string is low in small corpora, the frequencies of Chinese characters still have relatively high values. According to Observation 1, if a given sequence is an MLU, the characters in the sequence should have a similar frequency, in other words, s should be small. If the frequencies of all the characters in a Chinese sequence are equal, then $s = 0$. Because s represents the average frequency error of individual characters in the sequence, according to observation 1, in an MLU, the longer substring of that MLU will have smaller average frequency error.

According to our observation 1, an MLU can be identified by Equation 5. However, as Equation 5 only measures the frequency similarity between individual characters, any character combinations may be identified as MLU if their frequencies are similar; even when they are not occurring together. To avoid this problem, we introduce sequence frequency *f(S)* into the formula. Therefore, if the characters are not occurring together, they won't be considered as a sequence and therefore *f(S)* = 0. Thus any character combinations can be identified if they appear together as a sequence in the corpra.

Finally, we combine the sequence frequency and the RMSE measurement together. We designed the following equation to measure the possibility of *S* being a term:

$$R(S) = \frac{f(S)}{s+1} = \frac{f(S)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} + 1} \tag{6}$$

Where, *S* is a Chinese sequence; *f(S)* is the frequency of *s* in the corpus. We use $s+1$ as the denominator instead of using s to avoid 0 denominators.

Let *S* be a Chinese sequence with *n* characters; S= $a_1a_2\ldots a_n$. And *S'* is a substring of *S* with length *n*-1; S' = $a_1a_2\ldots a_{n-1}$.

According to observation 1, we should have:

- If *S* is an MLU, we will have $f(S) \approx f(S')$.
- If S is an MLU, the longer is S, the smaller the average mean square error is.

Therefore, in the case *S'* is a substring of *S* with length *n*-1, we would have$\sigma<\sigma'$. As a result we will have *R(S)>R(S').* In another case where *S'* is a substring of *S* and S' is an MLU while *S* is not. In other words, *S* has an additional character to an MLU. In this case, we will have *f(S) <f(S')* and the frequency of the additional character makes the RMSE value larger, so$\sigma>\sigma'$.Therefore, *R(S) <R(S').*

In summary, for a string S and its substring S', the one with higher R value would most likely be an MLU. Table 1 gives the R value of each possible term in a Chinese sentence chosen from a small collection of summaries returned from a search engine: "隱形戰機/是/一種/靈活度/極差/的/戰機" ("/" indicates the lexicon boundary given by a human).

Table 1 Chinese strings and *R(S)*

String *S*	*R(S)*
隱形	26.00
隱形戰	0.94

戰機	2.89
戰機是	0.08
一種	0.44
一種靈	0.21
靈活	2.00
靈活度	2.00
靈活度極	1.07
極差	0.8
極差的	0.07
戰機	2.89

This example clearly shows that if a Chinese MLU has an additional character, its R value will be significantly smaller than the R value of the MLU. For example, R(一種)=0.44>R(一種/靈)=0.21, R(靈活)=R(靈活度)=2.00>R(靈活度極)=1.07. It is reasonable that if we segment the Chinese sentence at the position that the string's R value drops greatly. For the example sentence, it would be segmented as: "隱形/戰機/是/一種/靈活度/極差/的/戰機" by the proposed method. The only difference between the human segmented sentence and the automatic segmented sentence is that "隱形戰機" (Stealth Fighter) is segmented into two words "隱形" (Stealth) and "戰機" (Fighter) by the proposed method. However, this is still an acceptable segmentation because those two words are meaningful.

3.2 A Bottom-up Term Extraction Strategy

As mentioned in Section 3.1, the top-down strategy is firstly to check whether the whole sentence is an MLU, then reduce the sentence size by 1 and recursively check sub sequences. It is reported that over 90% of meaningful Chinese terms consist of less than 4 characters [9], and on average, the number of characters in a sentence is much larger than 4. Obviously, a whole sentence is unlikely to be an MLU. Therefore, checking the whole sentence for an MLU is unnecessary. In this section, we describe a bottom-up strategy that extracts terms starting from the first character in the sentence. The basic idea is to determine the boundary of a term in a sentence by examining the frequency change (i.e., the change of the *R* value defined in Equation (6)) when the size of the term is increasing. If the *R* value of a term with size *n*+1 drops compared with its largest sub term with size *n*, the sub term with size *n* is extracted as an MLU. For example, in Table 1, there is a big drop between the *R* value of the third term "靈活度" (2.00) and its super term "靈活度極" (1.07). Therefore, "靈活度" is considered as an MLU.

The following algorithm describes the bottom-up term extraction strategy:

Algorithm BUTE(*s*)

Input: $s=a_1a_2\ldots a_n$ is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

1. Check each character in *s*, if it is a stop character such as 是 (is, are), 的(of) ，了…, remove it from *s*. After removing all stop characters, s becomes $a_1a_2\ldots a_m$, $m\le n$.
2. Let b=2, e=2, and M= f
3. Let $t_1= a_ba_2\ldots a_e$, $t_2= a_ba_2\ldots a_{(e+1)}$.
   If $R(t_1) > R(t_2)$, then $M=M\cup(t_1)$, b=e+1.
4. e=e+1, if e+1>m, return M, otherwise go to step 3.

The algorithm makes the sub sequence uncheckable once it is identified as an MLU (i.e., b=e+1 in step 3 ensures that the next valid checkable sequence doesn't contain t_1 which was just extracted as an MLU). However, when using the bottom-up strategy described above, some longer term might be missed since the longer term contains several shorter terms. As showed in our example, "隱形戰機" (Stealth Fighter) consists of two terms "隱形" and "戰機". When using bottom-up strategy, "隱形戰機" would not be extracted because the composite term has been segmented into two terms. To avoid this problem, we set up a fixed number w which equals specifies the maximum number of characters to be examined before reducing the size of the checkable sequence. The modified algorithm is given below:

Algorithm BUTE-M(*s*)

Input: $s=a_1a_2\ldots a_n$ is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

1. Check each character in *s*, if it is a stop character such as 是，了，的…, remove it from *s*. After removing all stop characters, s becomes $a_1a_2\ldots a_m$, $m\le n$.
2. Let b=2, e=2, First-term = true, and M= f
3. Let $t_1= a_ba_2\ldots a_e$, $t_2= a_ba_2\ldots a_{(e+1)}$.
   If $R(t_1) > R(t_2)$,
   then $M:=M\cup\{t_1\}$
   If First-term = true
   then first-position:= e and First-term:= false
   If $e\text{-}b+1 \ge w$
   then e:=first-position, b:=e+1, First-term:=true.
4. e=e+1, if e+1>m, return M, otherwise go to step 3

In algorithm **BUTE-M**, the variable first-position gives the ending position of the first identified MLU. Only when w characters have been examined, the first identified MLU will be removed from the next valid checkable sequence, otherwise the current sequence is still being checked for a possible MLU even it contains an extracted MLU. Therefore, not only the term "隱形" and "戰機" will be extracted but also the longer term "隱形戰機" (Stealth Fighter) will be extracted.

3.3 Translation selection

Translation selection is relatively simple compared with term extraction. The translation of a word in a source language is typically determined according to the ranking of the extracted terms. Each of the terms is assigned a rank, usually calculated

based on term frequency and term length[13]. The term with the highest rank in the extracted term list is selected as the translation of the English term.

As we have described in other papers [9; 10], the traditional translation selection approaches select the translation on the basis of word frequency and word length [4; 13]. We have suggested an approach to finding the most appropriate translation from the extracted word list regardless of term frequency. In our scheme even a low frequency word will have a chance to be selected. Our experiments in that paper show that in some cases, the most appropriate translation is the low frequency word. In this paper, we only give a brief description of our translation selection technique. The reader is referred to [9] for a more complete discussion.

The idea of our approach is to use the translation disambiguation technology to select the translation from the extracted term list. As the extracted terms are from the result set returned by the web search engine, it is reasonable to assume that those terms are relevant to the English query term that was submitted to the web search engine. If we assume all those terms are translations of the English term, we can apply the translation disambiguation technique to select the most appropriate term as the translation of the English term. We also introduced a filtering technique in our approach to minimize the length of the extracted term list.

In our approach, the correct translation will be selected using a simple translation disambiguation technique that is based on co-occurrence statistic. We use the total correlation which is one of several generalizations of the mutual information to calculate the relationship between the query words.

Our modified total correlation equation is defined as

$$C(x_1x_2x_3..x_n)=\log_2\frac{f(x_1x_2x_3..x_n)+1}{(f(x_1)+1)(f(x_2)+1)...(f(x_n)+1)} \quad (7)$$

Here, x_i are query words, $f(x_i)$ is the frequency that the query word x_i appears in the corpus, $f(x_1x_2x_3...x_n)$ is the frequency that all query words appears in the corpus. For each word frequency, we add 1 because we want to avoid 0 appearing in the equation when a word's frequency is 0.

The frequency information required by equation 7 can be easily collected from local corpus.

4. EVALUATION

We have conducted experiments to evaluate our proposed query translation approach. Using the algorithm described in Section 3.2, we firstly extract Chinese terms from the summaries returned by a search engine with an English term as the query. These Chinese terms are considered the candidate translations of the query English term. We then use the method described in Section 3.3 to select the most appropriate translation. The web search engine that we used in the experiments is Google, and the top 300 summaries returned were used for later processing. The English queries entered into Google will be enclosed by double quotation to ensure Google only returns result with exact phrases. Also specified result pages written in Chinese.

4.1 Test set

140 English queries from the NTCIR6 CLIR task were used. Query terms were first translated using Yahoo's online dictionary. (http://tw.dictionary.yahoo.com/). The remaining OOV terms which could not be translated were used to evaluate the performance of our web based query translation approach described in Section 3.2 and 3.3 by comparing with other approaches. There are 69 OOV terms altogether. The correct translation of each OOV term is given by NTCIR.

4.2 System setup

In our experiments, we evaluated the effectiveness of term extraction approaches in OOV translation. All the approaches described in section 2.1 were used in the experiment. The abbreviations are: MI for Mutual information, SE for the approach introduced by Chien, SCP for the Local Maxima introduced by Silva and Lopes, and SCPCD for the approach introduced by Jenq-Haur Wang et al.. Our extraction approach with BUTE-M extraction strategy is abbreviated as SQUT

The OOV term is translated via the following steps:

1. From the result pages downloaded from Google, use the 5 different term extraction approaches to produce 5 Chinese term lists.
2. For each term list, remove a term if it can be translated to English by Yahoo's online dictionary. This leaves only OOV terms.
3. From each remaining term list which contains only OOV terms, select the top 20 terms as translation candidates. Select the final translation from the candidate list using our translation selection approach described in 3.3.

Finally we have 5 sets of OOV translations, as shown in the Appendix.

4.3 Results and discussion

For the 69 OOV terms, by using the 5 different term extraction approaches, we obtained the translation results shown in Table 2. Details of the translation are showed in appendix.

As we were using the same corpus and the same translation selection approach, the difference in translation accuracy is the result of different term extraction approaches. Thus we can claim that the approach with the higher translation accuracy has higher extraction accuracy.

As we can see from table 2 below, SQUT has the highest translation accuracy. SCP and SCPCD provided similar performance. The approaches based on mutual information provided lowest performance.

Table 2. OOV translation accuracy

	Correct	Accuracy (%)
MI	30	43.5
SE	41	59.4
SCP	53	76.8
SCPCD	52	75.4
SQUT	59	85.5

4.3.1 Mutual information based approaches

In the experiment, MI based approach cannot determine the Chinese term boundaries well. The term lists produced by MI based approaches contain a large number of partial Chinese terms. It is quite often that partial Chinese terms were chosen as the translation of OOV terms. Some partial Chinese terms selected by our system are listed in table 3

Table 3 Some Extracted terms by MI

OOV Terms	Extracted terms	Correct terms
Embryonic Stem Cell	胚胎幹細	胚胎幹細胞
consumption tax	費稅	消費稅
Promoting Academic Excellence	卓越發	卓越發展計畫

Mutual information based term extraction approaches, such as MI and SE, are affected by many factors. These approaches rely on the predefined thresholds to determine the lexicon boundaries. Those thresholds can only be adjusted experimentally. Therefore, they can be optimized in static corpus. However, in OOV term translation, the corpus is dynamic. The result pages returned from search engine will be different for different term query entered. It is impossible to optimized thresholds for generic use. As a result, the output quality is not guaranteed.

In addition, mutual information based approaches seem unsuitable in Chinese term extraction. As there are no word boundaries between Chinese words, the calculation of MI values in Chinese are based on Chinese characters but not words as it does in English. The average high school graduate in the U.S. has a vocabulary of 27,600 words [11], while the cardinality of the commonly used Chinese character set is under 3000 [2]. Since Chinese characters have much higher frequencies than English words, one Chinese character will be used in many MLUs while an English word will have less chance to be used in Multiple MLUs. As a result, an English MLU will have much higher MI value than a Chinese MLU. The subtle differences in MI values between Chinese MLUs and non-MLUs make the thresholds hard to tune for generic use.

Some filtering techniques are used in SE to minimize the affect of thresholds. In our experiment, there is 17.2% improvement in translation accuracy. Obviously the improvement comes from the higher quality of extracted terms. However, the limitation of thresholds is not avoidable.

4.3.2 Local Maxima based approaches

Without using thresholds, local maxima based approaches have much better flexibility than MI based approaches, achieving higher translation accuracy in our experiment. In comparison, the SCP approach tries to extract longer MLUs while the SCPCD approach tries to extract shorter ones. The translation of "Autumn Struggle", "Wang Dan", "Masako" and "Renault" are all 2-character Chinese terms. SCPCD can extract the translation with no problem while SCP always has trouble with them. As over 90% of the Chinese terms are short terms, this is a problem for SCP in Chinese term extraction. In the mean time, SCPCD has trouble in extracting long terms. Overall, the two local maxima based approaches have similar performance. However, since in our experiment, most of the translations of OOV terms are long terms, SCP's performance is a little better than that of SCPCD.

Local maxima based approaches use string frequencies in the calculation of $\frac{1}{n-1}\sum_{i=1}^{n-1} f(w_1...w_i) f(w_{i+1}...w_n)$. In a small corpus, the frequency of a string becomes very low which makes the calculation of string frequencies less meaningful. Local Maxima based approaches are not effective in a small corpus. In comparison, our approach calculates the difference between character frequencies. In a small corpus, characters still have a relatively high value. As a result, our approach performs better than Local Maxima based approaches in small corpora. For example, local maxima based approaches were unable to extract the translation of "Nissan Motor Company" because the corpus is too small - Google only returns 73 results for the query "Nissan Motor Company".

4.3.3 SQUT Approach

Most of the translations can be extracted by the SQUT algorithm. As our approach monitors the change in R to determine a string to be an MLU instead of using the absolute value of R, it does not have the difficulty of using predefined thresholds. In addition, the use of single character frequencies in RMSE calculations makes our approach usable in small corpora. Therefore, we have much higher translation accuracy than MI based approaches and also about 10% improvement over Local Maxima based approaches.

However, the SQUT algorithm has difficulty in extracting the translation of "Wang Dan". In analyzing the result summaries, we found that the Chinese character "王"("Wang") is not only a very frequent character in the summaries but also used in other terms such as "霸王"(the Conqueror), "帝王"(regal); "國王"(king); "女王" (queen) and "王朝" (dynasty). Those terms also appear frequently in the result summaries. In our approach, where we are using the count of individual characters, the very high frequency of "王" breaks observation 2. Thus the translation of "Wang Dan" cannot be extracted. However, in most cases, our observations are true in small corpora as demonstrated by the high translation accuracy of our approach in query expansion from Chinese/English web search summaries.

5. Conclusion and Future Work

In this paper, we proposed a bottom-up term extraction approach to be used in small corpora. The method introduces a new measurement of a Chinese string based on frequency and RMSE, together with a Chinese MLU extraction process based on the change of the new string measurement that does not rely on any predefined thresholds. The method considers a Chinese string as a term based on the change of R's value when the size of the string increases rather than the absolute value of R. Our experiments show that this approach is effective in web mining for translation extraction of unknown query terms.

Although the proposed approach shows impressive accuracy for OOV term translation, there are still some works to be conducted in the future. Our experiments were conducted using a small scale test set which only has 69 OOV terms from NTCIR6 CLIR task queries. It might be necessary to test our approach under larger scale test set such as a test set that has over 1000 OOV terms. Although there are 140, only 50 of them have document relevance

results given by NTCIR. There are only 11 OOV terms in those 50 queries. As a result, the number of OOV terms is not enough to distinguish the different approaches. While many researchers [4; 5; 7; 9; 13] had showed that better query translation quality should improve the CLIR performance, it might be necessary to test actual benefits of high translation accuracy in CLIR in the future when we have appropriate testing data.

6. REFERENCES

[1] Mutual information, Wikipedia.

[2] The List of common use Chinese Characters, Ministry of Education of the People's Republic of China.

[3] A. Chen, H. Jiang, and F. Gey, Combining multiple sources for short query translation in Chinese-English cross-language information retrieval, Proceedings of the fifth international workshop on on Information retrieval with Asian languages, ACM Press, Hong Kong, China, 2000.

[4] A. Chen, and F. Gey, Experiments on Cross-language and Patent retrieval at NTCIR3 Worksho, Proceedings of the 3rd NTCIR Workshop, Japan, 2003.

[5] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien, Translating unknown queries with web corpora for cross-language information retrieval, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, Sheffield, United Kingdom, 2004.

[6] L.-F. Chien, PAT-tree-based keyword extraction for Chinese information retrieval, Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval ACM Press, Philadelphia, Pennsylvania, United States 1997.

[7] J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang, Improving query translation for cross-language information retrieval using statistical models, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New Orleans, Louisiana, United States, 2001.

[8] M.-G. Jang, S.H. Myaeng, and S.Y. Park, Using mutual information to resolve query translation ambiguities and query term weighting, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, College Park, Maryland, 1999.

[9] C. Lu, Y. Xu, and S. Geva, Translation disambiguation in web-based translation extraction for English-Chinese CLIR, Proceeding of The 22nd Annual ACM Symposium on Applied Computing, 2007.

[10] C. Lu, Y. Xu, and S. Geva, Improving translation accuracy in web-based translation extraction, Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies:Information Retrieval, Question Answering and Cross-Lingual Information Access, Japan, 2007.

[11] M. Salovesh, How many words in an "average" person's vocabulary?, http://unauthorised.org/anthropology/anthro-l/august-1996/0436.html, 1996.

[12] J.F.d. Silva, and G.P. Lopes, A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units, International Conference on Mathematics of Language, 1999.

[13] Y. Zhang, and P. Vines, Using the web for automated translation extraction in cross-language information retrieval, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, Sheffield, United Kingdom, 2004.

7. Appendix Sample of translated terms

OOV term	SQUT	SCP	SCPCD	SE	MI
Autumn Struggle:	秋鬥大遊	從秋鬥	秋鬥	秋鬥	秋鬥
Jonnie Walker:	約翰走路	約翰走路	黑次元	高雄演唱	高雄演唱
Charity Golf Tournament:	慈善高爾夫球賽	慈善高爾夫球賽		慈善高	慈善高
Embryonic Stem Cell:	胚胎幹細胞	胚胎幹細胞	胚胎幹細胞		
Florence Griffith Joyner:	花蝴蝶	葛瑞菲絲	葛瑞菲絲	花蝴蝶	花蝴蝶
FloJo:	佛羅倫薩格里菲斯	花蝴蝶	花蝴蝶	花蝴蝶	花蝴蝶
Michael Jordan:	麥可喬丹	麥可喬丹	喬丹	喬丹	喬丹
Hu Jin tao:	胡錦濤	胡錦濤	胡錦濤	胡錦濤	胡錦濤
Wang Dan:		天安門	王丹	王丹	王丹
Tiananmen	天安門廣場	天安門	天安門	天安門	天安門
Akira Kurosawa:	黑澤明	黑澤明	黑澤明	黑澤明	黑澤明
Keizo Obuchi:	小淵惠三	小淵惠三	小淵惠三	小淵惠三	小淵惠三
Environmental Hormone:	環境荷爾蒙	環境荷爾蒙	環境荷爾蒙	環境荷爾蒙	
Acquired Immune Deficiency Syndrome:	後天免疫缺乏症候群	愛滋病	愛滋病	愛滋病	愛滋
Social Problem:	社會問題	社會問題	社會問題		

Kia Motors:	起亞汽車	起亞汽車	起亞汽車	起亞	起亞
Self Defense Force:	自衛隊	自衛隊	自衛隊	自衛隊	自衛隊
Animal Cloning Technique:	動物克隆技術	動物克隆技術			
Political Crisis:	政治危機	政治危機	政治危機		
Public Officer:	公職人員	公職人員	公職人員	公職人員	
Research Trend:	研究趨勢	研究趨勢	研究趨勢	研究趨勢	
Foreign Worker:	外籍勞工	外籍勞工	外籍勞工	外籍勞工	
World Cup:	世界盃	世界盃	世界盃	世界盃	世界盃
Apple Computer:	蘋果公司	蘋果電腦	蘋果電腦	蘋果電腦	蘋果電腦
Weapon of Mass Destruction:	大規模毀滅性武器	大規模毀滅性武器	性武器		
Energy Consumption:	能源消費	能源消費	能源消費		
International Space Station:	國際太空站	國際太空站	國際太空站		
President Habibie:	哈比比總統	哈比比總統	哈比比總統	哈比比	
Underground Nuclear Test:	地下核試驗	地下核試驗	地下核試		
F117:	戰鬥機	隱形戰機	隱形戰	隱形戰	隱形戰
Stealth Fighter:	隱形戰機	隱形戰機	形戰鬥機	隱形戰	隱形戰
Masako:	雅子	太子妃	雅子	雅子	雅子
Copyright Protection:	版權保護	版權保護	版權保護	版權保護	版權保護
Daepodong:	大浦洞	大浦洞	大浦洞	大浦洞	大浦洞
Contactless SMART Card:	智慧卡	非接觸式智慧卡	非接觸式智慧卡	非接觸式	非接觸式
Han Dynasty:	漢朝	大漢風	漢朝	漢朝	漢朝
Promoting Academic Excellence:	學術追求卓越發展計畫	卓越計畫	卓越發展計畫	卓越發展計畫	卓越發
China Airlines:	中華航空	中華航空	中華航空	中華航空	長榮
El Nino	聖嬰	聖嬰現象	聖嬰現象	聖嬰	聖嬰
Mount Ali:	阿里山	阿里山	阿里山	阿里山	阿里山
Kazuhiro Sasaki:	佐佐木主浩	佐佐木主浩	佐佐木	佐佐木	佐佐木
Seattle Mariners:	西雅圖水手	西雅圖水手	西雅圖水手		
Takeshi Kitano:	北野武	北野武	北野武	北野武	北野武
Nissan Motor Company:	日產汽車公司	汽車公司	汽車公司	處經濟	處經濟
Renault:	雷諾	休旅車	雷諾	雷諾	雷諾
war crime:	戰爭罪	戰爭罪	戰爭罪	戰爭罪	
Kim Dae Jung:	金大中	金大中	金大中	金大中	金大中
Medecins Sans Frontieres:	無國界醫生	無國界醫生	無國界醫生		
volcanic eruptions:	火山爆發				
Clinton:	克林頓	克林頓	克林頓		
Science Camp:	科學營	科學營	科學營	科學營	
Kim Il Sung:	金日成	金日成	金日成	金日成	金日成
anticancer drug:	抗癌藥物				
consumption tax:	消費稅	消費稅	消費稅	消費稅	費稅
Uruguay Round:	烏拉圭回合	烏拉圭回合	烏拉圭回合		
Economic Collaboration:	經濟整合	經濟整合	經濟整合	經濟整合	經濟整合
Kim Jong Il:	金正日	金正日	金正日	金正日	金正日