Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 61-90

Web-Based Query Translation for English-Chinese CLIR

Chengye Lu*, Yue Xu*, and Shlomo Geva*

Abstract

Dictionary-based translation is a traditional approach in use by cross-language information retrieval systems. However, significant performance degradation is often observed when queries contain words that do not appear in the dictionary. This is called the Out of Vocabulary (OOV) problem. In recent years, Web mining has been shown to be one of the effective approaches for solving this problem. However, the questions of how to extract Multiword Lexical Units (MLUs) from the Web content and how to select the correct translations from the extracted candidate MLUs are still two difficult problems in Web mining based automated translation approaches.

Most statistical approaches to MLU extraction rely on statistical information extracted from huge corpora. In the case of using Web mining techniques for automated translations, these approaches do not perform well because the size of the corpus is usually too small and statistical approaches that rely on a large sample can become unreliable. In this paper, we present a new Chinese term measurement and a new Chinese MLU extraction process that work well on small corpora. We also present our approach to the selection of MLUs in a more accurate manner. Our experiments show marked improvement in translation accuracy over other commonly used approaches.

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, users increasingly wish to explore documents that were written in either their native language

* Faculty of Information Technology, School of Software Engineering and Data Communications, Queensland University of Technology, Brisbane, QLD 4001, Australia
E-mail: {c.lu,yue.xu,s.geva}@qut.edu.au

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

or some language other than English. Cross-language information retrieval (CLIR) systems allow users to retrieve documents written in more than one language through queries written in a different language. This is a helpful end-user feature. Obviously, translation is needed in the CLIR process; either translating the query into the document language, or translating the documents into the query language. 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, due to the vocabulary limitation of dictionaries, very often the translations of some words in a query cannot be found in a dictionary. This problem is called the Out of Vocabulary (OOV) problem. Very often, the OOV terms are proper names or newly created words. Even using the best dictionary, the OOV problem is unavoidable. As input queries are usually short, query expansion does not provide enough information to help recover the missing words. Furthermore, in many cases, it is exactly the OOV terms that are the crucial words in the 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 published only a few years ago. If the word ***SARS*** is left out of the translated query, it is most likely that the user will be unable to find any relevant documents. Moreover, a phrase cannot always be translated by translating each individual word in the phrase. For example, an idiom is a phrase and should not be translated by combining translations of the individual words because the correct translation may be a specific word which is not the combination of individual word translations of the original phrase.

Another problem with the dictionary-based translation approach is the translation disambiguation problem. The problem is more serious for a language which does not have word boundaries, such as Chinese. Translation disambiguation refers to finding the most appropriate translation from several choices in the dictionary. For example, the English word STRING has over 20 different translations in Chinese, according to the Kingsoft online dictionary (www.kingsoft.com). One approach is to select the most likely translation [Eijk 1993] – usually the first one offered by a dictionary. However, even if the choices are ordered based on some criteria and the most likely *a-priori* translation is picked, in general, such an approach has a low probability of success. Another solution is to use all possible translations in the query with the OR operator. However, while this approach is likely to include the correct translation, it also introduces noise into the query. This can lead to the retrieval of many irrelevant documents which is, of course, undesirable. [Jang *et al.* 1999] and [Gao *et al.* 2001] report that this approach has precision that is 50% lower than the precision that is obtained by human translation.

In this paper, we present a Web-based approach to term extraction and translation selection. Specifically, we introduce a statistics-based approach to extracting terms and a translation disambiguation technique to improve the precision of the translations. The remainder of this paper is structured as follows: in Section 2, we present the existing approaches to query translation; in Section 3 we present our approach. Experimental evaluation and results discussion are presented in Section 4 and Section 5, respectively. Finally, we conclude the paper in Section 6.

2. PREVIOUS WORK

2.1 Translation

Dictionary-based query translation is one of the conventional approaches in CLIR. The appearance of OOV terms is one of the main difficulties arising with this approach. In very early years, OOV terms were not translated at all, leaving out the original terms in the translated query. This approach may significantly limit retrieval performance. In this section, several existing approaches to OOV translation are reviewed.

2.1.1 Transliteration

Proper names, such as personal names and place names, are a major source of OOV terms because many dictionaries do not include such terms. It is common for foreign names to be translated word-by-word based on phonetic pronunciations. In this manner, a name in one language will be pronounced similarly in another language – this is called transliteration. Such translation is usually done by a human when a new proper name is introduced from one language to another language.

Some researchers [Paola *et al*. 2003; Yan *et al*. 2003] have applied the rule of transliteration to automatically translate proper names. Basically, the transliteration will first convert words in one language into phonetic symbols, then convert the phonetic symbols into another language. Some researchers have found that transliteration is quite useful in proper name translation [Paola *et al*. 2003; Yan *et al*. 2003]. However, transliteration is useful in only a few language pairs. When dealing with language pairs for which there are many phonemes in one language that are not present in the other, such as Chinese and English, the problem is exacerbated. There are even more problems when translating English to Chinese. First, as there is no standard for name translation in Chinese, different communities may translate a name in different ways. For example, the word “Disney” is translated as “迪斯尼” in mainland China but is translated as“迪士尼” in Taiwan. Both are pronounced similarly in Chinese, but use different Chinese characters. Even a human interpreter would have difficulty in unambiguously choosing which character should be used. Second, at times, the Chinese

translation only uses some of the phonemes of the English names. For example, the translation of "American" is "美国" which only uses the second syllable of "American". Finally, the translation of a name is not limited to only using translation but also to transliteration. Sometimes, the translation of a proper name may even use a mixed form of translation and transliteration. For example, the translation of "New Zealand" in mainland China is "新西兰", where "新" is the translation of "New" and "西兰" is the transliteration of "Zealand".

2.1.2 Parallel Text Mining

Parallel text is a text in one language together with its translation in another language. The typical way to use parallel texts is to generate translation equivalence automatically, without using a dictionary. It has been used in several studies [Eijk 1993; Kupiec 1993; Smadja *et al*. 1996; Nie *et al*. 1999] on multilingual related tasks such as machine translation or CLIR.

The idea of parallel text mining is straightforward. Since parallel texts are texts in two languages, it should be possible to identify corresponding sentences in two languages. When the corresponding sentences have been correctly identified, it is possible to learn the correspondence translation of each term in the sentence using statistical information since the term's translation will always appear in the corresponding sentences. Therefore, an OOV term can be translated by mining parallel corpora. Many researchers have also reported that parallel text mining based translation can significantly improve the CLIR performance [Eijk 1993; Kupiec 1993; Smadja *et al*. 1996; Nie *et al*. 1999].

In the very early stages, parallel text based translation approaches were word-by-word based and only domain specific noun terms were translated. In general, these approaches [Eijk 1993; Kupiec 1993] first align the sentences in each corpus, then noun phrases are identified by a part-of-speech tagger. Finally, noun terms are mapped using simple frequency calculations. In such translation models, phrases, especially verb phrases, are very hard to translate. As phrases in one language may have different word order in another language, phrases cannot be translated on a word-by-word basis. This problem in parallel text based translation is called the collocation problem.

Some later approaches [Smadja *et al.* 1996; Nie *et al*. 1999] started to use more complex strategies such as statistical association measurement or probabilistic translation models to solve the collocation problem. Smadja *et al.* [Smadja *et al*. 1996] proposed an approach that can translate word pairs and phrases. In particular, they used a statistical association measure of the Dice coefficient to deal with the problem of collocation translation. Nie *et al*. [Nie *et al*. 1999] proposed an approach based on a probabilistic model that demonstrates another approach to solving the collocation problem. Using parallel texts, their translation model can return $p(t|S)$, which is the probability of having the term t of the target language in the translation of the source sentence S. As the probability model does not consider the order and

the position of words, collocation is no longer a problem.

Some of the advantages of the parallel text based approaches include the very high accuracy of translation without using bilingual dictionaries and the extraction of multiple transitions with equivalent meaning that can be used for query expansion. However, the sources of parallel corpora tend to be limited to some particular domain and language pairs. Currently, large-scale parallel corpora are available only in the form of government proceedings, *e.g.* Canadian parliamentary proceedings in English and French, or Hong Kong government proceedings in Chinese and English. Obviously, such corpora are not suitable for translating newly created terms or domain-specific terms that are outside the domains of the corpora. As a result, current studies of parallel text based translation are focusing on constructing large-scale parallel corpora in various domains from the Web.

2.1.3 Web Mining

Web mining for automated translation is based on the observation that there are a large number of Web pages on the Internet that contain parallel text in several languages. Investigation has found that when a new English term, such as a new technical term or a proper name, is introduced into Chinese, the Chinese translation to this term and the original English term very often appear together in literature publications in an attempt to avoid misunderstanding. Some earlier studies have already addressed the problem of extracting useful information from the Internet using Web search engines such as Google and Yahoo. These search engines search for English terms on pages in a certain language, *e.g.*, Chinese or Japanese. The results of Web search engines are normally a long, ordered list of document titles and summaries to help users locate information. Mining the result lists can help find translations to the unknown query terms. Some studies [Cheng *et al*. 2004; Zhang *et al*. 2004] have shown that such approaches are rather effective for proper name translation.

Generally, Web-based translation extraction approaches consist of three steps:

1. Web document retrieval: use a Web search engine to find the documents in the target language that contain the OOV term in the original language and 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 and 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 the identification of correct MLUs in the summaries (refer to Section 2.2 for the definition of MLUs).
3. Translation selection: select the appropriate translation from the extracted words. As the previous steps may produce a long list of terms, translation selection has to find the correct translation from the extracted terms.

The existing term extraction techniques in the second step fall into two main categories: approaches that are based on lexical analysis or dictionary-based word segmentation, and approaches that are based on co-occurrence statistics. When translating Chinese text into English, Chinese terms should be correctly detected first. As there are no word boundaries in Chinese text, the mining system has to perform segmentation of the Chinese sentences to find the candidate words. The quality of the segmentation greatly influences the quality of the term extraction because incorrect segmentation of the Chinese text may break the correct translation of an English term into two or more words so that the correct word is lost. The translation selection in the third step also suffers from the problem that selection of the most frequent word or the longest word, which is the more popular techniques, does not always produce a correct translation. The term extraction and translation selection problems will be further addressed in subsequent sections.

2.2 Term Extraction

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 convey a specific meaning. For example, compound nouns like *Disneyland*, phrasal verbs like *take into account*, adverbial locutions like *as soon as possible*, and idioms like *cutting edge* are MLUs. In most cases, it is necessary to extract MLUs rather than individual words from a corpus because the meaning of an MLU is not always the combination of individual words in the MLU. The meaning of the MLU 'cutting edge' is not the combination of the meaning of individual words, 'cutting' and 'edge'.

Finding MLUs from the summaries returned by a search engine is important in Web mining based automated translation. If only words are extracted from the summaries, the following process may not be able to find the correct translation since 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 as the MLU extraction in English documents also needs to mark the lexical boundaries between MLUs. Therefore, term extraction in Chinese documents can be considered as Chinese word segmentation. Many existing systems use lexicon-based or dictionary-based segmentation techniques to determine word boundaries in Chinese text. However, in the case of Web mining for automated translation, as an OOV term is an unknown term to the system, the dictionary-based segmenters usually cannot correctly identify the OOV terms in the sentence. Therefore, the translation of an OOV term cannot be found in

a later process. Some researchers have suggested approaches that are based on co-occurrence statistics model for Chinese word segmentation to avoid this problem [Chen *et al*. 2000; Maeda *et al*. 2000; Gao *et al*. 2001; Pirkola *et al*. 2001].

2.2.1 Mutual Information and its Variations

One of the most popular statistics-based extraction approaches is to use mutual information [Chien 1997; Silva *et al*. 1999]. Mutual information is defined as:

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

The mutual information measurement quantifies the distance between the joint distribution of terms *X* and *Y* and the product of their marginal distributions. 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 [Chien 1997] 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:

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

where *c* is a Chinese string with *n* characters; *a* and *b* are the 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 as an 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.

2.2.2 Local Maxima Based Approaches

All mutual information based approaches require tuning the thresholds for generic use. Silva and Lopes suggest an approach called Local Maxima to extract MLU from corpora without using any predefined threshold [Silva *et al*. 1999]. The equation used in Local Maxima is known as SCP and is defined as follows:

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

where S is an n-gram string and $w_1,\ldots,w_i$ is the substring of S. A string is judged as an MLU if the SCP value is greater than or equal to the SCP value of all the substrings of S and also greater than or equal to 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 where S is its antecedent.

Although Local Maxima should be a language-independent approach, Jenq-Haur Wang *et al*. [Cheng *et al*. 2004] 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 uses the function:

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

where S is the input string, $w_1..w_i$ is the substring of S, and $LC()$ and $RC()$ are functions to calculate the number of unique left or right adjacent characters of S. A string is judged as a Chinese term if the SCPCD value is greater than or equal to the SCPCD value of all the substrings of S.

In summary, statistics-based approaches are widely used in Chinese term extraction for translation. The main reason is that Chinese terms are required to be extracted from search engine result pages. However, search engine results are usually partial sentences, which makes the traditional Chinese word segmentation hard to apply in this situation.

Current statistics-based approaches still have weaknesses. Web pages returned from a search engine are used for search engine based OOV term translation. In most cases, only a few hundred of the top result snippets on the result pages are used for translation extraction. Consequently, 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 3.2, we will describe our approach to addressing this difficulty in detail.

3. Web-Based Query Translation

Our approach is based on earlier work by Chen [Chen *et al*. 2000] and Zhang [Zhang *et al*. 2004]. Both approaches submit English queries (usually an English term) to a Web search engine and the top returned results (*i.e*., summaries in Chinese) are segmented into a word list. Each of the words in the list is then assigned a rank calculated from term frequency. The word with the highest rank in the word list is selected as the translation of the English term. However, observations have shown that the correct translation does not always have the highest frequency, even though it very often has a high frequency. The most appropriate translation is not necessarily the term with the highest rank.

As in previous work, we adopted the same idea of finding the OOV term's translation

through a Web search engine. However, our approach differs in term ranking and selection strategy. The aim of our approach is to find the most appropriate translation from the word list regardless of term frequency, which is the basic measurement used in previous work. Our approach combines translation disambiguation technology and Web-based translation extraction technology. Web-based translation extraction usually returns a list of words in a target language. As those words are all extracted from the result snippets returned by the Web search engine, it is reasonable to assume that these words are relevant to the English terms that were submitted to the Web search engine. If we assume all these words are potential translations of the English terms, we can apply the translation disambiguation technique to select the most appropriate word as the translation of the English terms.

Our translation extraction approach contains three major modules: collecting Web document summaries, word extraction, and translation selection. For easier understanding, we will use an example of finding the translation to the term "Stealth Fighter" to demonstrate our approach.

3.1 Collecting Web Document Summaries

First, we collect the top 100 document summaries returned from Google that contain both English and Chinese words. The English queries entered into Google will be enclosed in double quotation marks to ensure Google only returns results with the exact phrase. Sample document summaries are shown in Figure 1.


PChome Online 網路家庭-下載 - [Translate this page]
其中包括了隱形戰機(Stealth fighter)、Su-27、F-16、Sr-71、Glider、X-29等。安裝好後，到控制台中的顯示器內容設定螢幕保護裝置為"3D-Terrain Flight"就可以了。然後，在其設定值中設定所要出現的"戰鬥機種"(總共有六種機型) 及螢幕的視野 ...
toget.pchome.com.tw/intro/desktop_ssaver/desktop_ssaver_military/5317.html - 16k -
Cached - Similar pages

SOC GAMING > [分享]Allegiance - [Translate this page]
Stealth Fighter(隱形戰機): 隱形戰機是一種靈活度極差的戰機,因此短距離的纏鬥決不是Stealth Fighter的強項,Stealth Fighter的威力在於其裝載的Hunter Misslie,一種長距離且高殺傷力的飛彈, Stealth Fighter ...
www.socgame.com.tw/bbs/lofiversion/index.php/t56968.html - 15k - Supplemental Result -
Cached - Similar pages

蜻蜓的俱樂部- Yahoo!奇摩部落格 - [Translate this page]
1991年一月十七日凌晨，F-117A隱形戰機（stealth fighter），由沙烏地阿拉伯基地 ... 《詳全文》. 回應 (0) 引用 (0). 米格-29支點系列戰機. 分類：特種部隊. 2006/11/29 15:18. MIG-29的進氣道具保護裝置，注意滑行時機翼下方進氣道的保 ... 《詳全文》 ...
tw.myblog.yahoo.com/jw!LXskEjiWHwXt3jX2TYlqG8BQ2w--/archive?l=f&id=32&page=3 - 32k -
Supplemental Result - Cached - Similar pages

Figure 1. Three sample document summaries for "Stealth Fighter"

Figure 1 shows that *Stealth Fighter* and its translation in Chinese 隱形戰機 always appear together. The Chinese translation of *Stealth Fighter* appears either before or after the English words. In the example summaries shown in Figure 1, the translation and the English term "Stealth Fighter" are highlighted in red.

Although the query submitted to Google is asking for Chinese documents, Google may still return some documents purely in English. Therefore, we need to filter out the documents that are written in English only. The documents that contain both the English terms and Chinese characters are kept. Also, all the html tags are removed, and only the plain text is kept.

Second, from the document summaries returned by the search engine, we collect the sentences in the target language; for example, we can collect three Chinese sentences from the three sample document summaries in Figure 1. Each sentence must contain the English term and the Chinese characters before and after the term. From the summaries given in Figure 1, we get the following Chinese sentences shown in Figure 2.

Stealth Fighter(隱形戰機): 隱形戰機是一種靈活度極差的戰機,
因此短距離的纏鬥決不是 Stealth Fighter 的強項,
Stealth Fighter 的威力在於其裝載的 Hunter Misslie
一種長距離且高殺傷力的飛彈,

其中包括了隱形戰機(Stealth fighter)、Su-27、F-16、Sr-71、Glider、X-29 等。
安裝好後，到控制台中的顯示器內容設定螢幕保護裝置為"3D-Terrain Flight"就可以了。

1991 年一月十七日凌晨，F-117A 隱形戰機（stealth fighter），由沙烏地阿拉伯基地

Figure 2. Sample output of Chinese string collection

3.2 Word/Phrase Extraction

In order to calculate the statistical information of the Chinese terms, the Chinese sentences have to be correctly segmented. The term extraction approaches reviewed in Section 2.2 have been widely used on large corpora. However, in our experiments, the performance of those approaches is not always satisfactory for search engine based OOV term translation approaches.

In this section, we describe a term extraction approach specifically designed for 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 algorithm that can help to increase the extraction quality.

3.2.1 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 using a bottom-up approach that starts with examining the first character and then examining 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 Web pages returned from a search engine, the frequencies of the characters in an MLU are similar. This is due to the nature of the sample: in a small collection of text, there are a small number of MLUs and the characters appearing in one MLU may not appear in other MLUs. We also found that some different MLUs with similar meanings very often 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 frequency is low in a small collection, the individual characters of the term might still be relatively high and also have 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 extended 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 (RMSE) given in Equation (5) to measure the similarity between the character frequencies.

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

For a given Chinese character sequence with n characters, x_i is the frequency of each character in the sequence and $\overline{x}$ 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, σ should be small.

If the frequencies of all the characters in a Chinese sequence are equal, then σ = 0. Since σ represents the average frequency deviation from the mean 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 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 MLUs 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. With this addition, if the characters are not occurring together, they will not be considered as a sequence, causing f(S) = 0. Thus, any character combination can be identified if it appears as a sequence in the corpus.

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

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

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

Let S be a Chinese sequence with n characters; S= a1a2….an. And S' is a substring of S with length n-1; S' = a1a2….an-1. According to Observation 1, if S is an MLU, we should have f(S) ≈ f(S'), and the longer S is, the smaller σ should be. Therefore, in the case where 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'). Consider another case where S' is a substring of S and S' is an MLU while S is not. In other words, S adds an additional character to an MLU. In this case, we will have f(S) <f(S') and the frequency of the additional character makes σ larger, so $\sigma > \sigma'$ and 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, 一種, 靈活, and 靈活度 are valid MLUs, but 一種靈 and 靈活度極 are not.

In Table 1, we have:

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,

which shows the R value drop from 一種 to 一種靈, and from 靈活 and 靈活度 to 靈活度極.

This example indicates that it is reasonable to segment the Chinese sentence at the positions where 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 words in Chinese and have the same meaning as the combination of the two words.

3.2.2 A Bottom-Up Term Extraction Strategy

As mentioned in Section 3.1, the top-down strategy is to first 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 [Wu 2004], 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=a1a2….an is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

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 a1a2….am, m≤n.

Let b=1, e=1, and M=φ

Let t1= aba2….ae, t2= aba2….a(e+1).

If R(t1) >R(t2), then M=M∪(t1), b=e+1.

e=e+1, if e+1>m, return M, otherwise go to step 3.

Once a sequence is identified as an MLU, the algorithm BUTE will not check its subsequences for other possible MLUs (*i.e.*, b=e+1 in step 3 makes it so the next valid checkable sequence doesn't contain t1, which was just extracted as an MLU). However, when using the bottom-up strategy, some longer terms might be missed when the longer term contains several shorter terms. As shown 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 ω which 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=a1a2….an is a Chinese sentence with n Chinese characters

Output: M, a set of MLUs

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 a1a2….am, m≤n.

Let b=1, e=1, First-term = true, and M= φ

Let t1= aba2….ae, t2= aba2….a(e+1).

If R(t1) >R(t2),

then M:=M ∪ {t1)

If First-term = true

then first-position:= e and First-term:= false

If e-b+1 ≥ ω

then e:=first-position, b:=e+1, First-term:=true.

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 ω characters have been examined will the first identified MLU be removed from the next valid checkable sequence, otherwise the current sequence is still being checked for a possible MLU even if it contains an extracted MLU. Therefore, not only will the terms "隱形" and "戰機" be extracted, but also the longer term "隱形戰機" (Stealth Fighter) will be extracted.

3.3 Translation Selection

At this point, we have a list of translation candidates for the query term. The final step is to find the correct translation from the candidate list.

As we have described in another paper [Lu *et al*. 2007], the traditional translation selection approaches select the translation based on word frequency and word length [Chen *et al*. 2003; Zhang *et al*. 2004]. We have proposed an approach to determining the most appropriate translation from the extracted word list using the documents in the collection dataset regardless of term frequency. Using this approach, even a low-frequency word might be selected. Our experiments in that paper show that in some cases, the most appropriate translation can be a word with low frequency.

First, we retrieve the documents that contain each candidate translation from the collection. Then, we calculate the frequency of each candidate translation in the collection.

For instance, suppose we have an English query with three terms A,B,C and A1,A2…, B1,B2…, and C1,C2… are the candidate translations for A, B, and C, respectively, and the frequency of A1, A2, …, B1, B2, …, C1,C2… in the collection is f(A1), f(A2),… f(B1), f(B2)…, and so on. Second, we retrieve the documents that contain all the possible combinations of the candidate translations and calculate the frequencies. For example, the frequency of combination A1B1C1 is f(A1B1C1), A1B2C1 is f(A1B2C1), and A1B2C3 is f(A1B2C3)… and so on. Finally, we calculate the co-occurrence of all the possible combinations using the following equation:

$$C(x_1x_2x_3...x_n) = \log_2 \frac{N^{n-1} f(x_1x_2x_3...x_n)}{f(x_1) f(x_2) f(x_3)... f(x_n)}, \quad (7)$$

where x_i is a candidate translation for the ith query term, $f(x_i)$ is the frequency of word x_i appearing in the corpus, $x_1x_2…x_n$ is a combination of the candidate translation, $f(x_1x_2x_3...x_n)$ is the frequency that $x_1x_2…x_n$ appears in the corpus. N is the size of the corpus.

For the example query with three terms A, B, C, the co-occurrence of three candidate translation $A_1B_1C_1$ is calculated by:

$$C(A1B1C1) = \log_2 \frac{N^2 f(A1B1C1)}{f(A1) f(B1) f(C1)}. \quad (8)$$

The translation combination with the highest total correlation value C is selected as the correct translation for that query.

4. Experiments

We have conducted experiments to evaluate our proposed query translation approach. The Web search engine used in the experiments was Google.

4.1 Test Set

Queries, document collection, and relevance judgments provided by NTCIR (http://research.nii.ac.jp/ntcir/) are used in the experiments. The NTCIR6 Chinese test document collection was used as our test collection. The articles in the collection are news articles published on United Daily News (udn), United Express (ude), MingHseng News (mhn), and Economic Daily News (edn) in 2000-2001, for a total of 901,446 articles.

Queries used in the experiments are from NTCIR5 and NTCIR6 CLIR tasks. Altogether, there are 100 queries created by researchers from Taiwan, Japan, and Korea. NTCIR provided both English queries and corresponding Chinese queries. The Chinese queries are translated by human translators and are, thus, correct translations of the corresponding English queries.

In our experiments, English queries are extracted from English description fields by human experts. The corresponding Chinese translations are transcribed from the Chinese title fields by humans.

Yahoo's online English-Chinese dictionary (http://tw.dictionary.yahoo.com/) is used in the experiments. We first translate the English queries using the Yahoo's online English-Chinese dictionary. The terms that could not be translated by the online dictionary were used as the input queries to evaluate the performance of our proposed Web-based query translation approach. There are 108 OOV terms that cannot be translated by the online dictionary and, therefore, are used in the experiments.

4.2 Retrieval System

The documents were indexed using a character-based inverted file index. In the inverted file, the indexer records each Chinese character, its position in the document, and the document ID. Chinese phrase is determined by each Chinese character position and document ID. Only when character positions are consecutive and have the same document ID will the character sequence be considered as a phrase in the document. English words and numbers in the document are also recorded in the inverted file.

The retrieval model that is used in the system is an extended Boolean model with tf-idf weighting schema which is used in GPX by [Geva 2006]. Document rank for a query Q is calculated by the equation below:

$$D_{rank} = n^5 \sum tf_i * idf_i$$

Here, n is the number of the unique query terms in the document. tfi is the frequency of the ith term in the document and $idfi$ is the inverse document frequency of the ith term in the collection. This equation can ensure two things: first, the more unique query terms that match in a document, the higher rank the document has. For example, the document that contains five unique query terms will always have higher rank than the document that contains four query terms, regardless of the query terms frequency in the document; second, when documents contain the same number of unique terms, the score of a document will be determined by the sum of query terms' *tf-idf*, as traditional information retrieval does.

We do not employ relevance feedback in the retrieval system. Also, all the retrieval results are initial search results without query expansion.

4.3 Experiment Design

We designed two sets of experiments to evaluate our approach. The first set of experiments was designed to evaluate the effectiveness of term extraction for OOV translation, and the second set of experiments was designed to evaluate the effectiveness of translation selection

for OOV translation.

4.3.1 Experiment Set 1

In this experiment, we compared the performance of our proposed translation extraction approach (denoted as SQUT) with the approaches reviewed in Section 2.2, including the Mutual Information method (denoted as MI), the approach introduced by [Chien 1997] (denoted as SE), the Local Maxima method introduced by [Silva *et al*. 1999] (denoted as SCP), and the approach introduced by [Cheng *et al*. 2004] (denoted as SCPCD).

The OOV term is translated via the following steps:

Send the OOV term as a query to Google; from the result pages returned from Google, use the five different term extraction approaches to produce five Chinese term lists.

If a Chinese word can be translated to an English word using a dictionary, the English word must not be an OOV word. This means, the Chinese word must not be a translation of the queried English OOV word. Therefore, for each term list obtained in Step 1, remove the terms which can be translated to English by Yahoo's online dictionary. After this step, only OOV terms remain.

Select the top 20 terms in the new term list as translation candidates. Select the final translation from the candidate list using our translation selection approach described in 3.3.

Finally, we have five sets of OOV translations produced by the five approaches, respectively. A sample of the translation is given in Appendix 1.

Translation accuracy will be determined by human experts. Chinese queries will be used as reference only. As we were using the same corpus and the same translation selection approach, the difference in translation accuracy is the result of using different term extraction approaches. Thus, we can claim that the approach with the higher translation accuracy has higher extraction accuracy.

4.3.2 Experiment Set 2

This experiment is to retrieve Chinese documents for a given English query. The following experiments were conducted:

1. Mono: in this run, we use the original Chinese queries form NTCIR5. Only the title field is used and the Chinese terms are segmented by a human. This run provides the baseline result for comparison with all other runs.
2. IgnoreOOV: in this run, the English queries are translated using the online Yahoo

English-Chinese dictionary with the disambiguation technology proposed in 3.3. If a translation is not found in the dictionary, the query will keep the original English word.

3. SimpleSelect: similar to IgnoreOOV, English queries are translated using the online Yahoo English-Chinese dictionary with disambiguation technology. If a term cannot be translated by the dictionary, it will be translated by the proposed Web mining based approach. However, in the translation selection step, the longest and the highest frequency string were selected as its translation. This run simulates the previous Web translation selection approaches.
4. TQUT: like SimpleSelect, except that the translation for the "missing word" is selected with the disambiguation technology that is discussed in 3.3. Actually, TQUT uses the same translation technology as SQUT which we used in Experiment Set 1. We named it TQUT here simply to distinguish the concept that TQUT is an information retrieval task while SQUT is a translation task.

Although NTCIR gives 190 queries, only 100 of them have relevance judgments. Therefore, we are only able to evaluate the retrieval performance using those 100 queries in Experiment Set 2.

5. Results and Discussion

5.1 Experiment Set 1

For the 108 OOV terms, using the five different term extraction approaches, we obtained the translation results shown in Table 2. 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	48	44.4
SE	58	53.7
SCP	73	67.6
SCPCD	74	68.5
SQUT	84	77.8

5.1.1 Mutual Information Based Approaches

In the experiment, the MI based approaches were unable to determine the Chinese term boundaries well. The term lists produced by the MI based approaches contain a huge number of partial Chinese terms. It is quite often the case 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

The performance of the Mutual Information based term extraction approaches, such as MI and SE, is affected by many factors. These approaches rely on predefined thresholds to determine the lexicon boundaries. Those thresholds can only be adjusted experimentally. Therefore, they can be optimized in fixed corpora. However, in OOV term translation, the corpus is dynamic. It is almost impossible to optimize thresholds for general 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 is based on Chinese characters but not words as in English. On average, a high school graduate in the U.S. has a vocabulary of 27,600 words [Salovesh 1996]. Unless stemming or lemmatizing is used, the number of English word variations in a corpus is much greater. In contrast, the cardinality of the commonly used Chinese character set is under 3000. Due to the small set of Chinese characters, Chinese characters have much higher frequencies than English words. This means that one Chinese character could be used in many MLUs while an English word will have a much lower probability of being used in Multiple MLUs. As a result, an English MLU will have much higher MI value than a Chinese MLU. The subtle difference in MI values between MLUs and non-MLUs in Chinese makes the thresholds hard to tune for general use.

Some filtering techniques are used in SE to minimize the affect of thresholds. In our experiment, there is a 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.

5.1.2 Local Maxima Based Approaches

Without using thresholds, Local Maxima based approaches have much better flexibility than the MI based approaches in various corpora, achieving higher translation accuracy in our experiment. Comparing the two, 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 difficulty with them. As over 90% of the Chinese terms are short terms, this is a problem for SCP in Chinese term extraction. Conversely, SCPCD has difficulty in extracting long terms. Overall, the two Local Maxima based approaches have similar performance. However, since most of the translation of OOV terms are long terms in our experiment, 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".

5.1.3 SQUT Approach

Most of the translations can be extracted by the SQUT algorithm. As our approach monitors the change in R value to determine MLUs rather than 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 calculation makes our approach suitable in small corpora. Therefore, we have much higher translation accuracy than the MI-based approaches and also about 10% improvement over the 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 a very high-frequency character in the summaries. It is 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.

5.2 Experiment Set 2

Table 4 below gives the results from the four runs defined in Section 4.3.2.

Table 4. NTCIR 5 retrieval performance

	Average precision	Percentage of MonoRun
Mono	0. 3713	-
IgnoreOOV	0.1312	35.3%
SimpleSelect	0.2482	66.8%
TQUT	0.2978	79.3%

5.2.1 IgnoreOOV

The performance of the IgnoreOOV is 0.1312 which is only 35.3% of the monolingual retrieval performance. This result shows the extent to which an OOV term can affect a query. By looking at the translated queries, we found that 62 queries out of 100 have OOV terms. By removing all 62 queries, the Mono's average precision becomes 0.3026 and the IgnoreOOV's average precision becomes 0.2581 which is about 85.3% of the Mono's precision. This is a reasonable result and indicates that our disambiguation technique works well to find the correct translations. The reason that we cannot get 100% precision is mainly due to the limited coverage of the dictionary introducing inappropriate translations. By "inappropriate translation", we mean that the translation is a valid translation in some other context but not in the current query context. In query 24: for "space station, Mir", 儲存信息暫存器 (Memory Information Register) is the only translation returned from the dictionary. However, it should be translated to 和平號太空站 here. In this case, when a dictionary only returns one translation, it is difficult to tell if it is suitable in the context. As the dictionary only gives one translation, we have no opportunity to correct this translation error using a disambiguation technique. Some translations from the dictionary are inappropriate in some given contexts because the translations are different in different regions. For example, the query "mad cow disease" is translated to 瘋牛病 in the dictionary which is used in mainland China and Hong Kong. However, in the NTCIR collection which is obtained from Taiwan, "mad cow disease" is translated to 狂牛症 or to 狂牛病. We also find the same problem in query 24 "syndrome". Its translation is 症候群 in Taiwan. The translations given in the dictionary, though, are 併發症狀 and 綜合症狀, which are used in Hong Kong and mainland China. With these inappropriate translations, the retrieval precision for these queries is very low, thus it is impossible to achieve 100% of Mono performance.

5.2.2 SimpleSelect

The performance of SimpleSelect, which achieved 0.2482 in precision, was much better than IgnoreOOV and it is 66.8% of the Mono performance. This result shows quite clearly that some of the OOV terms in English are found and translated to Chinese correctly.

Table 5. Retrieval performance on queries that contains OOV terms only

	Average precision	Percentage of Mono Run
Mono	0.4134	-
SimpleSelect	0.2149	52.0%
TQUT	0.2946	71.3%

The results of the 62 queries that have OOV terms are given in Table 5. From Table 5, we can see that the precision of Mono is 0.4134 and the precision of SimpleSelect is 0.2149 which is 52.0% of the Mono's precision. This indicates that just choosing the longest and highest frequency terms as the translation of OOV terms results in performance that is actually lower than looking them up the dictionary. The performance is quite close to the performance of looking up terms in a dictionary without translation disambiguation technology reported by other researchers. However, some of our results show that this approach is quite useful in looking up proper names. As there is no standard for name translation in Chinese, it is quite common that a person's name might be translated into different forms with similar pronunciation (akin to phonetic form). Different people may choose different translations due to their custom. As our test collection contains articles from four different news agents, if we only choose one of the translations, we may not retrieve all the relevant documents.

For example, in query 12, the precision of SimpleSelect is 0.3528 and the precision of Mono is 0.0508 which means SimpleSelect's performance is vastly superior to Mono. This is a notable performance boost. The English OOV term in query 12 is Jennifer Capriati (the name of a tennis player). The human translation is 卡普莉雅蒂. The translations from our approach are 卡普裏亞蒂, 卡普莉雅蒂, 卡普裏雅蒂 and 雅蒂. They are all correct translations. It is clear that we miss many relevant documents when we only use the translation 卡普莉雅蒂. When we take a deep look into the collection, actually three out of four news agents have sports news. Those three news agents use three different translations for Jennifer Capriati. These translations are 卡普莉雅蒂 in the mhn, 凱普莉雅蒂 in the ude and 卡普莉亞蒂 in the udn. Obviously, our translated query takes advantage of adding 雅蒂. Since we use a character-based index for our collection, the documents containing 雅蒂 will include the documents that contain both 卡普莉雅蒂 and 凱普莉雅蒂. Therefore, although we cannot find the correct translation 凱普莉雅蒂, we can still retrieve the documents that contain 凱普莉雅蒂 by using 雅蒂.

Although using part of the translation might improve the retrieval performance, it also

introduces noise information and the noise information may make it harder for the search engine to find the relevant documents. For example, America Online is translated as 美國線上 in Taiwan but 美國在線 in mainland China. If we only choose 美國 (American) as the translation, we lose the information of the term. If it is the only term in the query, obviously, we are not going to retrieve any relevant documents.

5.2.3 TQUT

Table 6. OOV translation accuracy NTCIR5&6

	Correct	No. of OOV	Accuracy (%)
TQUT	50	71	70
SimpleSelect	43	71	60

Table 6 shows that, using translation disambiguated technology in Web Translation Extraction, we can get more accurate translation than in previous approaches. We have 65% accuracy of the translation while the simulation of previous approach only achieves 51%. The IR performance of disambiguated queries achieved 79.3% of the Mono which is 0.2978. If we only look at the results of 62 queries that contain OOV terms, the precision is 0.2846 which is 71.3% of the Mono's precision. This result is much higher than the result in SimpleSelect, which is only 52% of Mono. There are 71 OOV terms over 100 queries. 50 of the OOV terms' translations can be found using our proposed approach. And 43 of the translations are equivalent to the human translation. It is about 70% in precision.

There are many reasons for not being able to get 100% precision. The first reason is the different translation customs that we described earlier. Since we cannot control from where the Web search engine gets the documents and to whom the Web search engine returns documents, we cannot guarantee the translation will be suitable for the collection. For example, we may be able to find the translation for an OOV term from the Internet, but this translation may be used only in Hong Kong and is not suitable for a collection from Taiwan. The translation of the term "Kursk" is a good example. Our Web translation extraction method only returns one translation 庫爾斯克 as the translation of "Kursk". This result shows that most of the documents over the Internet use 庫爾斯克 as the translation of "Kursk". However, the NTCIR5 collection uses 科斯克 as its translation. This kind of inappropriate translation is very hard to avoid even by human interpreters. Another good example is the translation of "National Council of Timorese Resistance". We believe 帝汶抵抗全國委員會 (from our Web translation extraction system) and 東帝汶人抗爭國家委員會 (from NTCIR human translation) are both correct. The difference of the two translations comes from the different customs of translation. However, when using the two translations as two queries, our IR system cannot retrieve any documents. This means that the documents in the NTCIR5

collection use a different translation for "National Council of Timorese Resistance". Actually the translation in the NTCIR5 collection is: 東帝汶全國反抗會議.

Another reason that we cannot get 100% precision is that our Web translation extraction system does not consider the query context. As we described before, we only put the OOV terms into a Web search engine. This may lead to a situation where we get a translation suitable for other context. For instance, in query 36, we are looking for some articles about the use of a robot for remote operation in a medical context. "Remote operation" is an OOV term in this query. Our Web translation extraction method returns the term 遠程操作服務 as its translation. Disregarding the query context, this is a correct translation. However, this translation is only correct when it is used in computer science. If we do not consider the query context, 27 of the translations are correct with about 87% precision. This result is close to the disambiguated queries of dictionary translations which is 85%.

6. Conclusion and Future Work

In this paper, first, we reviewed some existing popular OOV translation approaches. Then, we described an approach to tackling the OOV problem in English-Chinese information retrieval. As the first step of this approach, we proposed a bottom-up term extraction approach suitable for small corpora for generating candidate translations for query OOV terms. This 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 to a 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 based on the absolute value of R. Our experiments show that this approach is effective for translation extraction of unknown query terms.

We also proposed a simple translation selection approach to improve translation accuracy. Our experimental results show that OOV terms can significantly affect the performance of CLIR systems. Using the translation extraction method proposed in this paper, the overall performance can be boosted by almost 174% relative to the case of not processing OOV terms. With our proposed translation selection approach, the accuracy of OOV term translation can be improved by up to 85%. The overall performance shows about 200% improvement relative to the case of not processing OOV terms. Also, it is about 120% relative to our implementation of previous approaches.

Although our proposed approach shows impressive accuracy for OOV term translation, there is still some work to be done in the future. First, our experiments were conducted using a relatively small scale test set from NTCIR5 and NTCIR6 along with CLIR task queries which only have 108 OOV terms. It is necessary to test our approach to a larger-scale test set such as a test set that has over 1000 OOV terms. Second, inappropriate translation is still a problem in

query translation. The main reasons include the limited size of the dictionary, different customs of translation, and ignoring query context. Some work should be done to minimize these problems. Our experiments provide hints for some possible approaches. If we have a large amount of resources, we may be able to find all the possible translations. For translation selection, if some of the translations hit a similar number of documents, we may keep all of them as correct translations. It may be useful to include more results from the Google search for instance or combining different translation result together. We will validate these ideas in the future.

REFERENCES

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

Chen, A. and F. Gey, "Experiments on Cross-language and Patent retrieval at NTCIR3 Workshop," in *Proceedings of the 3rd NTCIR Workshop*, Japan, 2003.

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

Cheng, P.-J., 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," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, ACM Press, 2004.

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

Eijk, P. v. d., "Automating the acquisition of bilingual terminology," in Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, Utrecht, The Netherlands, 1993.

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

Geva, S.,"Gardens Point XML IR at INEX 2006," Comparative Evaluation of XML information Retrieval Systems 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, Dagstuhl Castle, Germany, Springer, 2006.

Jang, M.-G., 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*,

College Park, Maryland, Association for Computational Linguistics, 1999.

Kupiec, J. M., "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," in *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, 1993.

Lu, C., Y. Xu, and S. Geva, "Translation disambiguation in web-based translation extraction for English-Chinese CLIR," in *Proceeding of The 22nd Annual ACM Symposium on Applied Computing*, 2007, pp. 819-823.

Maeda, A., F. Sadat, M. Yoshikawa, and S. Uemura, "Query term disambiguation for Web cross-language information retrieval using a search engine," in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, Hong Kong, China, ACM Press, 2000, pp. 25-32.

Nie, J.-Y., M. Simard, P. Isabelle, and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, ACM Press, 1999, pp. 74-81.

Paola, V. and K. Sanjeev, "Transliteration of proper names in cross-lingual information retrieval," *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15*, Association for Computational Linguistics, 2003.

Pirkola, A., T. Hedlund, H. Keskustalo, and K. Järvelin,"Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings," *Information Retrieval*, 4(3-4), 2001, pp. 209 - 230.

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

Silva, J. F. d., G. Dias, S. Guilloré, and J.G. Pereira, "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units," *Progress in Artificial Intelligence: 9th Portuguese Conference on Artificial Intelligence*, 1999.

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

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: a statistical approach," *Computational Linguistics,* 22(1), 1996, pp. 1-38.

Wu, G., "Research and Application on Statistical Language Model," *Computer science and technology*, Beijing, Tsinghua University, China, 2004.

Yan, Q., G. Gregory, and D.A. Evans., "Automatic transliteration for Japanese-to-English text retrieval," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, ACM Press, 2003.

Zhang, Y. 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*, Sheffield, United Kingdom, ACM Press, 2004.

Appendix 1

Sample translations of OOV terms from NTCIR

OOV term	SQUT	SCP	SCPCD	SE	MI
Chiutou:					
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
Torrijos Carter Treaty:					
Viagra:					
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	隱形戰鬥機	隱形戰	隱形戰	隱形戰
Stealth Fighter:	隱形戰機	隱形戰機	形戰鬥機	形戰鬥機	形戰鬥機
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
ST1:					
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
European monetary union:	歐洲貨幣聯盟	歐洲貨幣聯盟	歐洲貨幣	歐洲貨幣	歐洲貨幣
capital tie up:					
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

Pol Pot:	波布	紅高棉	紅高棉	紅高棉	紅高棉
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
Clinton:	克林頓	克林頓	克林頓		
New Year Holiday:	新年假期	新年假期	新年假期		
Drunken Driving:	醉後駕車	醉後駕車	醉後駕車	醉後駕車	後駕車
Science Camp:	科學營	科學營	科學營	科學營	
Nelson Mandela:	曼德拉	曼德拉	曼德拉	曼德拉	曼德拉
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		
Kim Jong Il:	金正日	金正日	金正日	金正日	金正日
Time Warner	時代華納	時代華納	時代華納	時代華納	時代華納
American Online	美國線上	美國線上	美國線上	美國線上	美國線上
Alberto Fujimori	藤森	藤森	藤森	藤森	藤森
Taliban	塔利班	塔利班	塔利班	塔利班	塔利班
Tiger Woods	老虎伍茲	老虎伍茲	老虎伍茲	老虎伍茲	伍茲
Harry Potter	哈利波特	哈利波特	哈利波特	哈利波特	哈利波特
Greenspan	葛林斯班	葛林斯班	葛林斯班	葛林斯	
monetary policy	貨幣政策	貨幣政策	貨幣政策	貨幣政策	
abnormal weather	天氣異常	天氣異常	天氣異常	天氣異常	天氣
National Council of Timorese Resistance	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會	帝汶抵抗全國委員會