Improving translation accuracy in web-based translation extraction

Chengye Lu Yue Xu 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

In this paper, we present some approaches to improve translation accuracy in web-based translation extraction. In previous work, the term extraction techniques that researchers used are proposed under large static corpus. We proposed some approaches that can improve the translation accuracy in web-based translation extraction which relies on small dynamic small corpus. We also analyzed the difference in using local text corpus and web corpus as disambiguation source.

Keywords: Cross-language Information retrieval, CLIR, query translation, translation disambiguation, OOV problem

1. INTRODUCTION

Dictionary based translation is one of the most common translation techniques that used for cross-language information retrieval (CLIR) because bilingual dictionaries are widely available and dictionary approaches are easy to implement. This approach shows high efficiency in word translation. However, as the coverage of a dictionary is always limited, there is always a challenge for cross-language information retrieval systems to discover the translation for the missing word. This problem is called Out of Vocabulary (OOV) problem.

The OOV problem usually happens to the translation of Multiword Lexical Units (MLU) such as proper names, phrases and new created words. Even in the best of dictionaries this is to be expected of course. As the length of input queries are usually less than 3 words, query expansion does not have enough information to help recover the missing words. Furthermore, it is precisely that sort of OOV term that is a key term 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 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. Obviously, a missing translation term in the query will affect the IR performance much. In NTCIR6, we are focusing on finding the translation of the query terms that do not in the dictionary.

In English-Chinese cross-language information retrieval, web-based translation extraction is a popular approach for OOV term translation. It is based on the observations that there are large numbers of 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 Chinese, the Chinese translation to this term and the original English term very often appear together in publications in an attempt to avoid misunderstanding. Mining this kind of web page can easily discover the translation of the new terms. Some earlier research already addressed the question of how those kinds of documents can be extracted by using web search engine such as Google and Yahoo. Popular search engines allow us to search English terms only for 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 is necessary to help find translations to the unknown query terms. Some studies [1, 9] have shown that such approaches are rather effective for proper name translation. Web-based translation extraction is commonly a three steps process.

1. Find documents: use web search engine to find the documents that contain both OOV term and target language. Collect the text in the result pages returned from the web search engine.

2. Extract terms: extract the words in the sentences where the OOV term appears. Record the words and their frequency appeared in the result summary.

3. Select translation: select the appropriate translation from the extracted terms.

Step 2 and step 3 are the core steps of Web-based translation extraction. The effectiveness of step 2 and step 3 will finally affect the final translation accuracy. By analyzing the previous approaches, we found that the main difficulty in step 2 is how to extract terms from sort corpora. The previous statistical based approaches are used in large static corpora. In Web-based translation extraction, the

corpora are search result summaries returned from search engine. The size of corpora is usually much smaller and is always dynamic. Therefore, the performance of term extraction is not always satisfactory. In the translation selection stage, the translation is usually selected by the highest rank. And the rank of a word is based on the calculation on word frequency and word length (Chen [1], Zhang [9]). Correct translation does not always have the highest frequency even thought it very often has a higher frequency. Therefore we argue that the correct translation is not necessarily the term with the highest rank.

2. System outline

In NTCIR6, our aim is to improve the translation accuracy of web-based translation extraction. We suggested several ideas using in step 2 and step 3 to improve the translation accuracy.

2.1 Term extraction in Chinese text

Extracting terms from Chinese text is more difficult than extracting terms from English text. 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 Web-based translation extraction, 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 [1, 3, 4, 5, 6, 9].

2.2 Frequency Change Measurement

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. According to our analyzing to small Chinese corpus, we found that while the frequencies of strings are low, Chinese characters still have a relatively high value. Therefore, we modify the local maxima based approaches into

$$R(s)=\frac{f(s)}{\sigma}=\frac{f(s)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (1)$$

Where, s is a Chinese sequence; *f(s)* is the frequency of s in the corpus. x_i is the frequency of the i_{th} Chinese character in the Chinese sequence and $\bar{x}$ is the average frequency of all the characters in the sequence

Let S be a Chinese sequence with n characters, S' is a substring of S with length n-1. If S is an MLU, we will have *f(S)* ≈*f(S')*. As S is an MLU, the longer is S, the smaller the average mean square error. We should have σ < σ '. 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')* σ > σ '.Therefore, *R(S)* <*R(S')*. In summary, if the Chinese string has higher R value, it is likely to be a Chinese MLU.

2.3 Term Extraction Strategy

Some Chinese terms consist of several small terms. For example, the Chinese term "天安門廣場"(Tiananmen Square) consists of two terms "天安門"(Tiananmen) and "廣場" (Square). The term extraction process should be able to extract not only the longest term but also the small terms. If we only keep the longest term, when we are looking for the translation of the sort term, we will not find it for the extracted term list. In this case, we suggest the following term extraction strategy:

Algorithm BUTE-M(*s*)

Input: $s=a_1a_2....a_n$ 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 $a_1a_2....a_m$, m≤n.

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

[3] Let $t_1= a_ba_2....a_e$, $t_2= a_ba_2....a_{(e+1)}$.

If $R(t_1) >> R(t_2)$,

then M:=M∪ $\{t_1\}$

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.

[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 ω 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.

2.4 Translation selection

Translation selection is relatively simple by comparison 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. 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 another paper [6], the traditional translation selection approaches select the translation on the basis of word frequency and word length (Chen [1], Zhang [9]). 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 [6] 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 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 terms, we can apply the translation disambiguation technique to select the most appropriate term as the translation of the English terms. 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

$$C(x_1x_2x_3...x_n)=\log_2\frac{f(x_1x_2x_3...x_n)+1}{(f(x_1)+1)(f(x_2)+1)...(f(x_n)+1)} \quad (2)$$

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 appears in the equation when a word's frequency is 0.

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

3. EVALUATION

3.1 Term Extraction

We have conducted two sets of experiments to evaluate the term extraction and translation selection approaches. The web search engine we used in the experiments is Google. The result pages returned from Google are stored for later processing.

The first set of experiments is designed to evaluate the effectiveness of term extraction approaches. 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 OOV term is translated via the following steps:

1. From the result page downloaded from Google, use the 3 different term extraction approaches to produce 3 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. 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 2.4.

Finally we have 5 sets of OOV translations and then we compare the translation accuracy.

The term extraction approaches we used are abbreviated as:

- SE for the approach introduced by Chien[2] which represents Mutual Information (MI) based approach.
- SCP for the Local Maxima introduced by Silva and Lopes[7].
- SQUT for our suggested approach.

3.2 Translation Selection

The second set of experiments is designed to evaluate the effectiveness of translation disambiguation approaches. We use NTCIR5 CLIR task topics. The following runs were performed in our English-Chinese CLIR experiments:

- 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 human. This run provides the baseline result for all other runs.
- Local: use the approach introduced in 2.4 with local corpus for translation disambiguation.
- Web: use the approach introduced in 2.4 with web corpus for translation disambiguation.

We compare the retrieval performance of different runs.

The documents were indexed using a character-based inverted file index. The retrieval model that is used in the system is Boolean model with *tf-idf* weighting schema. We do not employ the relevance feedback in the retrieval system. And all the retrieval results are initial search results without query expansion.

3.3 Results and discussion

3.4 Experiment 1

Table 1 Translation Accuracy of OOV Terms 1

	Correct	Accuracy (%)
SE	41	59.4
SCP	53	76.8
SQUT	59	85.5

For the 69 OOV terms, by using the 5 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 provided similar performance. The approaches based on mutual information provided lowest performance.

Most of the translations can be extracted by the SQUT algorithm. As our approach monitors the change in R (the result of equation 1) 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.

3.5 Experiment 2

Table 2 Retrieval performance

	MAP	Percentage of Mono
Mono	0. 3526	100%
Local	0.2576	73.1%
Web	0.2488	70.6%

As we can see from table 2, using local corpus for translation disambiguation is better then using web corpus. Because there is no standard for name translation in Chinese, it is quite common for a person's name to be translated into different form with similar pronunciation (akin to phonetic form). Different people may choose different translation due to their custom. 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 in Hong Kong and is not suitable for a collection from Taiwan. By using local corpus to verify the translation, our translation approach can minimize the problem of inappropriate translation.

4. CONCLUSION AND FUTURE WORK

In this paper, we have described some approaches to improve the translation accuracy in web-based translation extraction. We also introduce 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. Our experiments show that this approach is effective in web mining for translation extraction of unknown query terms. High translation accuracy of OOV term may help a CLIR process to increase the retrieval precision. In the translation selection step, using local corpus for translation disambiguation is better then using web corpus.

While the experimental result is promising, our retrieval performances are low. It has been reported by NTCIR5 relevance feedback can greatly improve the retrieval performance. Our future work will be focus on improving the CLIR retrieval performance.

5. References

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

[2] Chien, L.-F. (1997). 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 Philadelphia, Pennsylvania, United States ACM Press.

[3] Cheng, P.-J., J.-W. Teng, et al. (2004). Cross-language information retrieval: Translating unknown queries with web corpora for cross-language information retrieval. Proceedings of the 27th annual international conference on Research and development in information retrieval.

[4] Gao, J., J.-Y. Nie, et al. (2001). Improving query translation for cross-language information retrieval using statistical models. SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, ACM Press.

[5] Jang, M.-G., Myaeng, S.H., and Park, S.Y. (1999). 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.

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

[7] Silva, JF and Lopes, JGP (1999), A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword units, In Proceedings of the 6th meeting on the mathematics of language.

[8] Wikipedia. (2006). "Mutual information." from http://en.wikipedia.org/wiki/Mutual_information.

[9] Zhang, Y., and Vines, P (2004). 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.