Query Translation in Chinese-English Cross-Language Information Retrieval

Zhang Yibo, Sun Le, Du Lin, Sun Yufang
Chinese Information Processing Center,
Institute of Software, Chinese Academy of Sciences,
P.O.Box 8718, Beijing, 100080, P.R. China
e-mail: zyb, lesun, ldu, yfsun@sonata.iscas.ac.cn

Abstract

This paper proposed a new query translation method based on the mutual information matrices of terms in the Chinese and English corpora. Instead of looking up a bilingual phrase dictionary, the compositional phrase (the translation of phrase can be derived from the translation of its components) in the query can be indirectly translated via a general-purpose Chinese-English dictionary look-up procedure. A novel selection method for translations of query terms is also presented in detail. Our query translation method ultimately constructs an English query in which each query term has a weight. The evaluation results show that the retrieval performance achieved by our query translation method is about 73% of monolingual information retrieval and is about 28% higher than that of simple word-by-word translation way.

Introduction

With the rapid growth of electronic documents and the great development of network in China, there are more and more people touching the Internet, on which, however, English is the most popular language being used. It is difficult for most people in China to use English fluently, so they would like to use Chinese to express their queries to retrieval the relevant English documents. This situation motivates research in Cross Language Information Retrieval (CLIR).

There are two approaches to CLIR, one is query translation; the other is translating original language documents to destination language equivalents. Obviously, the latter is a very expensive task since there are so many documents in a collection and there is not yet a reliable machine translation system that can be used to process automatically. Most researchers are inclined to choose the query translation approach [Oard. (1996)]. Methods for query translation have focused on three areas: the employment of machine translation techniques, dictionary based translation [Hull & Grefenstette (1996); Ballesteros & Croft (1996)], parallel or comparable corpora for generating a translation model [Davis & Dunning (1995); Sheridan & Ballerini (1996); Nie, Jian-Yun et al.(1999)]. Machine translation (MT) method has many obstacles to prevent its employment into CLIR such as deep syntactic and semantic analysis, user queries consisting of only one or two words, and an arduous task to build a MT system. Dictionary based query translation is the most popular method because of its easiness to perform. The main reasons leading to the great drops in CLIR effectiveness by this method are ambiguities caused by more than one translation of a query term and failures to translate phrases during query translation. Previous studies [Hull & Grefenstette (1996); Ballesteros & Croft (1996)] have shown that automatic word-by-word (WBW) query translation via machine readable dictionary (MRD) results in a 40-60% loss in effectiveness below that of monolingual retrieval. With regard to the use of parallel corpora translation method, the critiques one often raises concern the availability of reliable parallel text corpora. An alternative way is that making use of the comparable corpora because they are easier to be obtained and there are more and more bilingual even multilingual documents on the Internet. From analyzing a document collection, an associated word list can be yielded and it is often used to expansion the query in monolingual information retrieval [Qiu & Frei(1993); Jing & Croft(1994)].

This research was supported by the National Science Fund of China for Distinguished Young Scholars under contact 69983009.

In this paper, a new query translation is presented by combination dictionary based method with the comparable corpora analyzing. Ambiguity problem and phrase information lost are attacked in dictionary based Chinese-English Cross-Language information Retrieval (CECLIR). The remainder of this paper is organized as follows: section 1 gives a method to calculate the mutual information matrices of Chinese-English comparable corpora. Section 2 develops a scheme to select the translations of the Chinese query terms and introduces how the compositional phrase can be kept in our method. Section 3 presents a set of preliminary experiment on comparable corpora to evaluate our query translation method and gives some explanations.

1 Mutual information matrices calculation

We hypothesize that the words in a sentence after being removed the stop words be associated with each other and work together to express a query requirement. The association relationship between two words can be indicated by their mutual information, which can be further used to discover phrases [Church & Hanks (1990)]. If two words are independent with each other, their mutual information would be close to zero. On the other hand, if they are strongly related, the mutual information would be much greater than zero and they would be much like to be a phrase; if they occur complementarily, the mutual information would be negative. In conclusion, the bigger the mutual information of word pair, the more probable the word phrase would be a phrase. According to [Fano (1961)], we can define the mutual information $MI(t_1,t_2)$ of term t_1 and t_2 as formula (1).

$$MI(t_1,t_2)=\log_2\frac{P(t_1,t_2)}{P(t_1)P(t_2)} \quad (1)$$

Where

$P(t_1,t_2)$ is the co-occurrence probability of t_1 and t_2 in a Chinese sentence. The reason we select a Chinese sentence to be a window other than a fixed length window is that a full Chinese sentence can keep more linguistic information and consequently, it is more reasonable that we can regard t_1 and t_2 to be a phrase when they co-occur in a sentence. $P(t_1)$ and $P(t_2)$ are the occurrence probabilities of term t_1 and t_2 in a sentence. These probabilities can be calculated by the occurrence of term t_1 and t_2 in the collection as equation (2), (3) and (4).

$$P(t_1)=\frac{n_{t_1}}{N} \quad (2)$$

$$P(t_2)=\frac{n_{t_2}}{N} \quad (3)$$

$$P(t_1,t_2)=\frac{n_{t_1,t_2}}{N} \quad (4)$$

Where

n_{t_1}, n_{t_2} is the individual term frequency of term t_1 and t_2 respectively if either of them occur in a sentence of the collection. n_{t_1,t_2} is the co-occurrence frequency of term t_1 and t_2 if they are all in a sentence of the collection. N is the number of sentences of the collection. Replacing (1) with equation (2), (3) and (4), the mutual information of term t_1 and t_2 can be expressed by following formula.

$$MI(t_1,t_2)=\log_2\frac{n_{t_1,t_2}N}{n_{t_1}n_{t_2}} \quad (5)$$

Table 2 and table 3 show the occurrence frequency values and mutual information values calculated by formula (5) for three Chinese compositional phrases and their corresponding English phrases respectively found in our comparable corpora.

$t_1 \mid t_2$	n_{t_1}	n_{t_2}	n_{t_1,t_2}	MI
文件|系统	106	84	45	9.28
用户|管理	45	97	21	9.21
图形|界面	73	22	19	10.51

Table 2: Mutual information of three Chinese phrases ($N=123{,}000$)

$t_1 \mid t_2$	n_{t_1}	n_{t_2}	n_{t_1,t_2}	MI
File | system	158	126	52	8.91
User | management	59	112	18	8.97
Graphic | interface	92	41	34	10.70

Table 3: Mutual information of three English phrases ($N=184{,}000$)

Analyzing the Chinese-English comparable corpora in this way, we can get two mutual information value matrices to indicate which two terms (as to the Chinese collection, they are almost Chinese words after segmentation) would be most possible to be a phrase. A word list associated to each Chinese query term can be obtained by looking up the mutual information value matrix of the Chinese corpus with a cutoff of $MI = 1.50$. As discussed above, the bigger the mutual information value between two terms, the more possible the two words would be a phrase. We can infer that the associated word list of the query term contains the terms that are the most possible components of a compositional phrase. In other words, the phrase information can be kept by this way. The Chinese query is translated into English via looking up the English senses of Chinese query term and words in its associated word list in a Chinese-English dictionary. The procedures how to select appropriate tranlations and to construct the English query are discussed in section 2.

2 Translations selection and phrase keeping

It is a naive method to translate a Chinese query only by looking up each Chinese term to get its English senses in a Chinese-English dictionary. This method, however, results in too many ambiguities during the query translation and offers no path to select appropriate ones among the translations. In addition, phrases in the query can not be translated effectively. Previous study has showed that failure to translate phrases greatly reduces the performance by up to 25% over automatic word-by-word (WBW) query translation [Ballesteros & Croft (1996)].

In our method, those English translations most likely co-occur with each other can be obtained via looking up the mutual information value matrix of the English corpus with a cutoff $MI = 1.50$. In this way, the English senses of terms in the associated word list can provide a good context for the translation of the Chinese query term and give a significant clue for its translations selection. In addition, the information of two terms (either Chinese or English) to be a phrase can also be stored in the associated word list. In the following, we firstly describe our method to select translations in detail, and then we give an example to demonstrate how to keep the phrase information in our method.

Supposing the Chinese query is expressed by $(e_1, e_2, \cdots, e_r)$. $e_1, e_2, ..., e_r$ are the segmented Chinese words of the query after removing the stop words. The translations of e_m $(m = 1, ..., r)$ by looking up the Chinese-English bilingual dictionary can be ordered in descending by following formula.

$$w(f_m^l) = \log_{10}(\alpha \cdot i\_MI(f_m^l) + \beta \cdot o\_MI(f_m^l)) \quad (6)$$

$$i\_MI(f_m^l) = \frac{\sum_{k=1}^{|E_m|} \sum_{j=1}^{|F_{mk}|} \left(MI(e_m, e_{mk}) \cdot MI(f_m^l, f_{mk}^j) \right)}{\sum_{k=1}^{|E_m|} |F_{mk}|} \quad (7)$$

$$o\_MI(f_m^l) = \frac{\sum_{i=1, i \neq m}^{r} \sum_{k=1}^{|F_i|} MI(f_m^l, f_i^k)}{\sum_{i=1, i \neq m}^{r} |F_i|} \quad (8)$$

Where

f_m^l is one sense of the English translation set F_m of the word e_m $(l = 1, ..., |F_m|)$. E_m is the association word set of e_m. The size of E_m is $|E_m|$ and its element is $e_{mk} (k = 1, ..., |E_m|)$. F_{mk} is the English translation set of e_{mk}, its element is f_{mk}^j. α is the coefficient to emphasize the inner mutual information between the English sense f_m^l of the single Chinese query term e_m and the English sense f_{mk}^j of the e_m's association word e_{mk}. The first part of the formula (6) $i\_MI(f_m^l)$ reflects the probability of English translation f_m^l and f_{mk}^j to be a phrase. β is the coefficient to emphasize the outside mutual information between f_m^l and the English sense f_i^k of the other Chinese terms included in the query. The second part of the formula (6) $o\_MI(f_m^l)$ reflects the relevant value between the English sense f_m^l of e_m and the whole query concept.

Our method of translations selection can be described as follows: if the weight of any translation of the Chinese query term is greater than 1.00, the sense is selected to construct the English query. If there is no weight of any translation of the Chinese query term greater than 1.00, the sense with biggest one is selected to construct the English query. In this way, we can make an English query by the following Boolean expression.

$$Query = \bigwedge_{m=1}^{r}\left(\bigvee_{l=1}^{|F_m|}\left(g_m^l, w(g_m^l)\right)\right) \qquad (9)$$

Where g_m^l is set element after the English translation sense set F_m which is detruncated by our translation selection method.

In order to demonstrate the procedure of our method, we give an example and explain how the English translations are selected and how the phrase information is kept. Given a simple Chinese query "用户，管理，命令 (user, management, command)" after segmentation and removing stop words, the associated word list of term "用户 (user)" is "管理 (management)，信息 (information)，手册 (manual)" and the associated word list of term "管理(management)" is "用户(user)，硬盘(hard disk)，文件(file)". We process the associated word "管理(management)" of the query term "用户(user)" in a special way by adding an appropriate value to their mutual information value to let theirs be the biggest in the associated word list, because the associated word "管理(management)" also occurs in the original query. Similar way is done with the associated word "用户(user)" of the query term "管理(management)". In this way, , the compositional phrase "用户管理 (user management)" can be kept in both associated word list of term "用户(user)" and term "管理(management)".

When term "用户" is translated into English by looking up the general-purpose Chinese-English bilingual dictionary, we get its English sense set "user, consumer" ordered by the formula (6). When term "管理" is translated into English, we get its English sense set "management, administration, supervision, run" ordered by the formula (6). We can find the first positions of the English translation set of the query term "用户" and term "管理" are "user" and "management" respectively. From the point of view of translation, the phrase "user management" can be regarded as the English phrase translation of "用户管理". According to our translation selection and formula (9), we can construct the English Boolean query as follows, in which each query term has a weight.

Query = (user, 1.86) and ((management, 1.83) or (administration, 1.63)) and (command, 1.92).

3 Evaluation and discussion

To evaluate our query translation method, we did a set of experiment to compare it to the word-by-word (WBW) translation method and manual translation method. In the word-by-word translation method, the Chinese queries are automatically segmented and the Chinese terms included in them are translated into English only by looking up the general-purpose Chinese-English bilingual dictionary. In the manual translation method, the Chinese queries are translated into English by a Ph.D. student. The segmentation we used is based on a small general-purpose Chinese-English bilingual dictionary that only contains 46,570 pairs in which each Chinese word has several English translations. The forward and backward maximum matching algorithm is used to segment the texts and find the combinatorial ambiguities. Of all the combinatorial ambiguities, 91.2% are removed with the word uni-gram prior probabilities. A stop word list of 1210 elements is set up, which contains frequently used functional words as well as symbols [Du & Sun (2000)]. Our Chinese query translation process contains following steps:

(1) Segment the Chinese query according to the method introduced above.
(2) Get the associated word list of each Chinese term included in the query from the Chinese mutual information matrix.
(3) Look up the English sense set of each Chinese term and its associated word in the general-purpose Chinese-English bilingual dictionary.
(4) Select the English translation sense by the method introduced in section 2 (in formula (6) the coefficents α and β are selected by 1.0 and 0.5 respectively in our experiment) and construct the English query on the basis of the formula (9).

The document collection used in our experiments consists of several Chinese and corresponding English computer manuals, which include Linux-HOWTO, PostgreSQL handbook, Mysql handbook, Linux kernel* and Linux Gazette 17 volumes (from July, 1998 to Dec., 1999)**. In order get a large number document Chinese and English collections, we decomposed these manuals and let every document no more than 15 sentences. As a

* http://www.linuxforum.net/books/index.html

**http://www.linuxgazette.com.cn

result, Chinese-English bilingual comparable corpora are obtained in which contain about 8,200 Chinese documents and 12,500 English documents. We design 13 Chinese queries, the average length is about 7 single Chinese character (about three Chinese words). All work in this study was performed on the Search2000 information retrieval system [Du & Zhang (2000)], which can process both Chinese and English Boolean queries.

Table 4 shows the precision and recall table for the three methods. The first column in table 4 contains precision values averaged 13 queries and interpolated to eleven recall points from 0.0 to 1.0 in steps of 0.1. The third column contains precision values achieved by our translation method (QT).

Recall	Precision (WBW)	Precision (Manual)	Precision (QT)
at 0.00	0.5831	0.8975	0.6642
at 0.10	0.5132	0.7884	0.5825
at 0.20	0.4036	0.6573	0.5174
at 0.30	0.3771	0.6206	0.4728
at 0.40	0.3128	0.5840	0.4163
at 0.50	0.2816	0.5118	0.3838
at 0.60	0.2143	0.4876	0.3104
at 0.70	0.1641	0.3833	0.2645
at 0.80	0.1110	0.2114	0.1702
at 0.90	0.0741	0.1667	0.1020
at 1.00	0.0212	0.0428	0.0342
Avg.	0.2778	0.4865	0.3562

Table 4: The results of the three methods

The results in table 4 suggest that in this case, the WBW query translation leads to a great drop in effectiveness of 42.90% below that for monolingual retrieval (manual translation method). The result of our query translation method greatly improves effectiveness by 28.22% over the WBW method, and its effectiveness is about 73.21% of that for monolingual retrieval. Although phrase translation is not executed directly in our method, the phrase information is kept effectively in the associated word list. Therefore, the phrase can be well translated. The associated word list also provides a good context for translation of the Chinese query terms (corresponding to the first part of formula (6) $i\_MI(f_m^l)$) and a good English translation is given a relatively high weight. The results in table 4 show that our query translation method can construct a good English query and indeed improve the effectiveness.

Conclusion

Automatic word-by-word query translation is an attractive method because it is easy to perform, resources are readily available, and performance is similar to that of other CLIR methods. However, there are a lot of ambiguities in translation of the query terms and failures to translate phrases correctly, which are mainly responsible for the large drops in effectiveness below monolingual retrieval performance. Aiming to tackle with these problems, we develop a new scheme for how to select translations in this paper. In addition, rather than using a bilingual phrase dictionary, we also put forward a new method to translate phrases indirectly by using the mutual information between two words in a full sentence and keep the phrase information in the associated word list effectively. As a result of our query translation method, an English query is constructed in which each query term has a weight.

In this study, our method leads to improve the effectiveness by 28.22% over the word by word query translation method, but is still about 27% below the monolingual retrieval performance. If query expansion is employed in our method, we expect the performance should be further improved. A shortcoming of our method is that the cost of calculation of the mutual information matrices is very large. We are currently exploring an algorithm to generate the matrices more efficiently and the selection of coefficients in formula (6) also needs further research.

Acknowledgements

The authors wish to express their appreciation to those interpreters of computer manuals. Without their selfless contribution, our experiment would be impossible. Thanks to the anonymous reviewers for their helpful comments.

References

Ballesteros, L. and Croft, W. B.(1996). Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*,pp.791-801.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics,* 16(1), pp. 22-29.

Davis, M. and Dunning, T. (1995). Query translation using evolutionary programming for multi-lingual information retrieval. In *Proceedings of the 4th Annual Conference on Evolutionary Programming,* pp. 175-185.

Du, Lin and Sun, Yufang. (2000). A new indexing method based on word proximity for Chinese text retrieval. *Journal of Computer Science and Technology,*15(3),pp.280-286.

Du, Lin; Zhang, Yibo and Sun, Yufang. (2000). The Design and Implementation of WEB-Based Chinese Text Retrieval System Search2000, (in Chinese). In *Proceedings of 2000 International Conference on Multilingual Information Processing,*pp.44-50.

Fano,. R. (1961). *Transmission of Information: A statistical theory of Communications.* MIT Press, Cambridge, MA.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual informaiton retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval,*pp.49-57.

Jing, Yufeng and Croft, W. Bruce. (1994). An association thesaurus for information retrieval. In *Proceedings of RIAO 94,*pp.146-160.

Nie, Jian-Yun; Brisebois M. and Ren, Xiaobo. (1996). On Chinese text retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,*pp.225-233.

Oard, D. W. (1996). A survey of multilingual text retrieval. *Technical Report UMIACS-TR-96-19,*http://www.ee.umd.edu/medlab/filter/papers/sigir96.ps.

Qiu, Yonggang, and Frei , H. P. (1993). Concept based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,*pp.160-169.

Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval,*pp.58-65.