Search-Result-Based Method for Unknown Term Translation in Cross-Language Information Retrieval

Jiun-Hung Lin Min-Shiang Shia Kao-Hung Lin
Shu-Jung Lin Scott Yu Wen-Hsiang Lu
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{jhlin, foreverdream, thexfile, shu-jung, scottyu }@csie.ncku.edu.tw
whlu@mail.ncku.edu.tw

Abstract

In this paper, we adopt two methods targeting NTCIR-5 Chinese-English CLIR task. First, to alleviate problems of unknown query terms, we combine dictionary-based and search-result-based methods to handle query translation for CLIR. Second, to reduce document retrieval time, we use a Chinese part-of-speech (*POS*) *tagger to extract only nouns, verbs, and foreign words as index terms. Additionally, we particularly focus on the evaluation of CLIR performance for TITLE queries because we are currently developing some techniques to handle translations of short unknown queries.*

Keywords: *cross-language information retrieval, search result, query translation, unknown term translation.*

1 Introduction

Query translation is one of the main challenges in Cross Language Information Retrieval (CLIR). Three major kinds of query translation methods have been proposed, including dictionary-, machine-translation-, and parallel-corpus-based methods. However, many proper names, such as person names, location names, and organization names, are still not correctly translated using the above methods. Some unknown proper names in NTCIR-5 Chinese queries are listed as follows: "柯恩"(Cohen), "祕魯"(Peru), and "時代華納" (Time Warner). Translating unknown terms is still a thorny challenge. We have proposed an effective search-result-based method to alleviate this problem [1]. Zhang and Vines [2] also proposed a method of mining web search results to deal with unknown term translation problem. Our strategy is to combine dictionary-based and search-result-based methods for query translation. If a query term can not be found in our dictionary, we use the search-result-based method to get translation for this unknown term. Currently, we use a representative Chinese POS tagger, provided by CKIP group of Academia Sinica [3], to handle Chinese word segmentation and extract only nouns, verbs, and foreign words as index terms for reducing document retrieval time.

2 Search-Result-Based Query Translation Method

To deal with translations of unknown query terms in Chinese-English CLIR tasks, we adopted a search-result-based method which has been shown effective to extract translations of unknown query term by exploring language-mixed search-result pages and utilizing the co-occurrence relation and context information. In this section, we will simply describe this method. For more details, please refer to our previous work [1].

(1) *Chi-square Test Method*: On the basis of co-occurrence relation, chi-square test (χ^2) is adopted to estimate similarity between the source term C and the target candidate E. The similarity measure is defined as

$$S_{\chi^2}(C,E)=\frac{N\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)}, \qquad (1)$$

where a, b, c and d are the numbers of pages retrieving from search engines by submitting Boolean queries: "C and E", "C and not E", "not C and E", and "not C and not E", respectively; N is the total number of pages, i.e., $N = a + b + c + d$.

(2) *Context-Vector Analysis Method***:** Due to the nature of Chinese-English mixed texts often appearing in Chinese pages, the source term C and the target candidate E may share common contextual terms in the Chinese search-result pages. The similarity between E and C will be computed based

on their context feature vectors in the vector-space model. The *tf-idf* weighting scheme is used and defined as

$$w_{t_i} = \frac{f(t_i, p)}{\max_j f(t_j, p)} \times \log(\frac{N}{n}), \qquad (2)$$

where $f(t_i, p)$ is the frequency of term t_i in search-result page p, N is the total number of Web pages, and n is the number of the pages containing t_i. Finally, we use the cosine measure to estimate the similarity as follows:

$$S_{CV}(C, \mathrm{E}) = \frac{\sum_{i=1}^{m} w_{e_i} \times w_{c_i}}{\sqrt{\sum_{i=1}^{m} (w_{e_i})^2 \times \sum_{i=1}^{m} (w_{c_i})^2}}. \qquad (3)$$

3 Monolingual Retrieval

At NTCIR-5 Chinese monolingual retrieval subtask, we submitted TITLE and DESCRIPTION runs. First, we introduce our indexing method in Chinese monolingual retrieval. We use CKIP tagger to deal with Chinese word segmentation and get POS tag information. To reduce processing time in document retrieval and analysis of keywords in queries, we only extract the POS tags of verbs, nouns, and foreign words as indexing terms and take words with high IDF as keywords. We heuristically select the first half terms with higher IDF to retrieve at most 10000 documents in the document corpus under consideration of retrieval time. We adopt a cosine measure with tf-idf weighting scheme to compute relevance between queries and documents. By separating one document into many passages, we believe that the similarity between queries and relevant documents will be more reliable if query terms appear at the same passage simultaneously. Inspired form Kwok [4], we separate a document into a few passages with constant length (550 bytes) and every two passages overlap 275 bytes. We use two different scoring strategies. One uses average score of all passages in a document as the final document score, and the other chooses the highest score of all passages in a document as the final document score.

4 Chinese-English CLIR

To translate queries in our Chinese-English CLIR system, we combine the dictionary-based and search-result-based methods. To handle translation for Chinese transliterated foreign name, especially Japanese and English names, we expand CEDICT [5] with 26k entries by combining a list of Japanese-English names [6] and a list of Chinese-English name pairs which were extracted from web search results by a semi-surprised method [7]. Currently, our dictionary contains about 260k entries. We simply describe the process of query translation. If any Chinese query term can be found in our dictionary, we just look at its corresponding translation in our expanded Chinese-English dictionary. If query terms can not be found in our dictionary, we use the search-result-based method to translate it (see Section 2).

5 Experimental Results

5.1 Chinese Monolingual Retrieval

The evaluation data provided from the NTCIR-5 Chinese monolingual retrieval subtask contains 901,446 Chinese news articles and 50 topic descriptions. We submitted two "TITLE" runs with the different ranking strategies, named T-Avg and T-Max, respectively. The T-Avg run uses only the title of the topics as queries and uses the average score of all passages in a document to represent the document's ranking score. The T-Max run uses only the title of the topics as queries and selects the maximum score of all passages in a document to represent the document's ranking score. Some results are shown in Table 1. We find that the strategy of selecting the maximum score seems to achieve better performance on the Title runs.

Table 1: Results of evaluation using two different ranking strategies in TITLE runs at Chinese monolingual retrieval task

Run	Relax			Rigid		
	MAP	P@10	R.Pre	MAP	P@10	R.pre
T-Avg	.2650	.4586	.2890	.2141	.3949	.2400
T-Max	.3019	.4908	.3174	.2421	.4251	.2469

Our evaluation results of Chinese monolingual retrieval task are not satisfactory. We think our heuristic strategy selecting the first half keywords with higher IDF and retrieving at most 10000 documents may miss a few relevant documents. We make a simple analysis by computing the average coverage rate (retrieved relevant documents with at most 10000 documents divided by relevant documents) for 50 queries. Table 2 shows the current average coverage rate.

In addition, we only extract the POS tags of verb, noun, and foreign words as indexing terms. It may loss some keywords with other POS tags (like adjective). Moreover, we do not employ characters as indexing terms. These two reasons may lead to worse performance in Chinese monolingual task. We will conduct more experiments to investigate the effects using these heuristic strategies in the future.

Table 2: Coverage analysis of retrieved relevant documents using our heuristic strategy in Chinese monolingual retrieval

Number of Retrieved Relevant Documents for 50 queries	3052
Number of Retrieved Relevant Documents with our Strategy of Selecting at most 10000 documents for 50 queries	2535
Average Coverage Rate	0.8306

5.2 Chinese-English CLIR

In the evaluation of the Chinese-English CLIR task, we focus on determining the effectiveness of translating unknown short queries (Title queries) by using search-result-based methods. The total number of TITLE queries is 49, and the total number of English documents is 259,050.

Table 3: Evaluation result in TITLE runs of the Chinese-English CLIR task

Run	Relax			Rigid		
	MAP	P@10	R.Pre	MAP	P@10	R.pre
Title	.1279	.2321	.1651	.1084	.1994	.1314

Table 3 shows the evaluation result of our TITLE runs at Chinese-English CLIR task. Although our performance is not good, our search-result-based method is still effective in dealing with translation of some unknown terms. For example, for TITLE query #02 "祕魯總統, 藤森, 醜聞, 賄選" (President of Peru, Alberto Fujimori, scandal, bribe), its results of Chinese word segmentation are "祕魯, 總統, 藤森, 醜 聞 , 賄 選 ". Our search-result-based method correctly translates the unknown query term "祕魯" into "Peru", but get wrong translation "bbc" of the unknown query term "賄選". Besides them, "藤森" is translated into "Fujinomori, Fujimori" correctly by our expanded dictionary containing a list of transliterated Japanese-English names. We make some detailed performance analysis for this query in Table 4. We observe the effects on three kinds of translation resources or methods, "C" means to use only common bilingual dictionary CEDICT dictionary, "JE" means Japanese-English name dictionary, and "S" means search-result-based method. We can find that the CLIR performance is significantly improved by bilingual name dictionaries and the search-result-based method. Although search-result-based method also get incorrect translations, but their negative effects in performance seem relatively trivial.

Table 4: Performance comparison for different translation resources or methods

Query #02 (祕魯總統，藤森，醜聞，賄選)			
Translation Resources/ Methods	Relax		
	MAP	P@10	R.Pre
C	.0001	.0000	.0000
C+JE	.4286	.5608	.4315
C+JE+S	.6002	.8000	.6225

Table 5 also shows correct translations of other unknown terms in TITLE queries extracted by our search-result-based method. The number in the parentheses means the rank of translation candidate returned by our search-result-based method. For example, MLB (1) means that MLB is the top-1 translation candidate.

Table 5: Correct translations of unknown query terms in TITLE queries

Query Number	Chinese Unknown Query Terms	Extracted English Translations of (Rank)
02	祕魯	Peru (1)
04	柯恩	Cohen (2)
11	大聯盟	MLB (1)
19	協和號	Concorde (3)
40	哈利波特	Harry Potter (1)

In the following, we do some error analyses. For TITLE query #19 "超音速飛機，協和號，墜機" (supersonic airliner, Concord, airplane crash), its results of Chinese word segmentation are "超音速，飛機，協和號，墜機". The unknown terms "協和號" and "墜機" gets wrong translations "jets" and "china books" using our search-result-based method. Thus, the MAP value of this query is 0.0001. Currently, we choose only the top-1 candidate as translation under the consideration of reliability. In the future, we are considering choosing top-k translation candidates to obtain better performance. For example, the top-3 translation candidates of the unknown query term "協和號" are "Jets", "Volvo", and "Concorde", and "Concorde" is the correct translation.

The same as Chinese monolingual retrieval task, we also use heuristic strategy to select the first half translated English terms with higher IDF to retrieve at most 10000 documents. This heuristic strategy may

also decrease the retrieval accuracy at Chinese-English CLIR task. Moreover, this strategy may have more risks because translated English terms may be incorrect, and using the first half terms with higher IDF to retrieve at most 10000 documents may cause more errors. We made an experimental analysis for this strategy. We gather the retrieved relevant documents with at most 10000 documents retrieved by our heuristic strategy. Then, we compute the average coverage rate (retrieved relevant documents with at most 10000 documents divided by relevant documents) for 49 queries. We list these statistics in Table 6. We will also make more experimental analysis for investigating the effectiveness using these heuristic strategies in the future.

Table 6: Coverage analysis of retrieved relevant documents using our heuristic strategy in Chinese-English CLIR

Number of Retrieved Relevant Documents for 50 queries	4064
Number of Retrieved Relevant Documents with our Strategy of Selecting at most 10000 documents for 50 queries	2458
Average Coverage Rate	0.6048

6 Conclusion and Future Work

In this paper, we use a search-result-based method to help translate unknown query terms. And we use the POS tag information to reduce document retrieval time. Translating unknown query terms is still a thorny problem to CLIR. In the future, we will try to combine Web-based transliteration models to improve our search-result-based term translation method [7].

References

[1] P. J. Cheng, J. W. Teng, R. C. Chen, J. H. Wang, W. H. Lu, L. F. Chien. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In *Proceedings of ACM SIGIR*, 2004.

[2] Y. Zhang and P. Vines. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proceedings of ACM SIGIR,* 2004.

[3] CKIP website: http://ckipsvr.iis.sinica.edu.tw/demo.htm

[4] K. L. Kwok. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In *Proceedings of the second NTCIR Workshop*, 2002.

[5] CEDICT website: http://www.mandarintools.com/cedict.html

[6] ENAMDICT website: http://www.csse.monash.edu.au/~jwb/enamdict_doc.html

[7] M. S. Shia, J. H. Lin, S. Yu, W. H. Lu. Improving Translation of Unknown Proper Names Using a Hybrid Translation Extraction Method. In *Proceedings of ROCLING,* 2005.