Using Wikipedia to Translate OOV Terms on MLIR

Chen-Yu Su, Tien-Chien Lin and Shih-Hung Wu*
Department of Computer Science and Information Engineering
Chaoyang University of Technology
Taichung County 41349, TAIWAN (R.O.C)
*Contact author: shwu@cyut.edu.tw

Abstract

We deal with Chinese, Japanese and Korean multilingual information retrieval (MLIR) in NTCIR-6, and submit our results on the C-CJK-T and C-CJK-D subtask. In these runs, we adopt Dictionary-Based Approach to translate query terms. In addition to tradition dictionary, we incorporate the Wikipedia as a live dictionary.
Keywords: *MLIR, Wikipedia, OOV terms*

1. Introduction

MLIR is an important application in IR. Many information workers collect information from the global resources, which might be in different languages. MLIR system can help the users to query in their native language and retrieve information in various foreign languages.

Our approach is to translate the query terms from the source language into the target language. There are two major difficulties in this dictionary-based multilingual information retrieval research: word sense disambiguation (WSD) and the out-of-vocabulary (OOV) terms 0. In these cases, the query terms cannot be translated correctly into target language. Ballesteros and Croft proposed the co-occurrence statistics method [2], Mirna proposed a term-sense disambiguation technique [7], and Federico and Bertoldi proposed N-best translation method [4] to solve the WSD problem. In addition, Ying, Phil and Justin collected co-occurrence from the retrieved web text by statistics method [11][12] to translation the Chinese OOV terms.

We focus on dealing with OOV terms. In our approach, we adopt the online translation website services as a fixed dictionary and the Wikipedia as a live dictionary to translate query terms [10]. Most dictionaries do update periodically, but the updating frequency of Wikipedia is much faster. Since it is updated by volunteers all over the world everyday, and the amount of updates are quite steady. The dictionary translation and Wikipedia translation are merged as the query translation for document retrieval in our system. Based on the query translation method, in NTCIR-6, we submitted our results on C-CJK-T and C-CJK-D subtasks.

The following sections are organized as follows: Section 2, 3 and 4 describe the index methods, the translation methods, and the retrieval methods respectively. We show the experiment results in section 5 and give the conclusions and future work in section 6.

2. Index Method

Our index and retrieval system is built based on the Lemur (http://www.lemurproject.org/) IR toolkit. Since the official corpora are not segmented, a preprocessing of word segmentation is necessary for building the index.

2.1 Chinese Document indexing

Our system adopts a Chinese word segmentation toolkit developed by CKIP group (Chinese Knowledge and Information Processing) to segment Chinese corpus into indexing terms. The CKIP group is a research team formed by the Institute of Information Science and the Institute of Linguistics of Academia Sinica in 1986. The average accuracy of the toolkit is about 95%. (http://ckipsvr.iis.sinica.edu.tw/)

2.2 Japanese Document indexing

For Japanese word segmentation, our system just inserts white space between characters. Then build the index. This is a naïve approach for word segmentation due to the time constraint. However, to our knowledge, a free Japanese segmentation toolkit "JUMAN" (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html) is available to segment Japanese documents. In our ongoing work, this toolkit helps us to segment Japanese sentence into terms to improve the performances on C-J CLIR.

2.3 Korean Document indexing

Traditionally, authors for separating eojeols (which are a kind of compound word) insert spaces in Korean sentences. Since most documents in the NTCIR Korean corpus do follow the tradition, our system built the index of the Korean corpus without preprocessing.

3. Query Translation

3.1 Dictionary-Based translation

The dictionary-based translation method translates the source language query into target language query with a fixed bilingual dictionary 0. In this paper, the existing free online translation website services are regarded as the fixed dictionaries. More details of the query translation of our system are in section 4.

3.2 Wikipedia translation

Wikipedia is a multilingual, Web-based, free content encyclopedia project. Wikipedia are written by volunteers all over the world. Anyone can edit or create new articles. The number of English articles is more then 1.6 million. There are eleven languages that have more than 0.1 million articles. Total has more then six million articles in 250 languages. The numbers of articles still grow up. (http://www.wikipedia.org)

Each entry of Wikipedia has links to entries in other languages if there are entries describing the same topic in those languages. The translation of an entry can be found just follow the link to the target language if the translation in target language is available. Therefore, Wikipedia can be seen as a live dictionary with all kinds of languages. Additionally, the titles of Wikipedia entries are proper noun in the majority which helps more on IR than just word. Since most query terms are proper noun too.

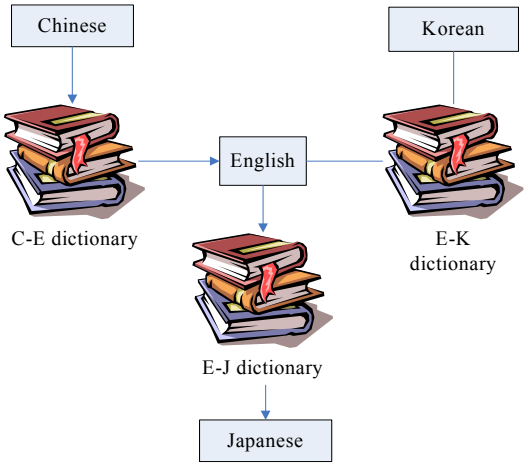

Fig 1. Using transitive translation method to translate Chinese into Japanese

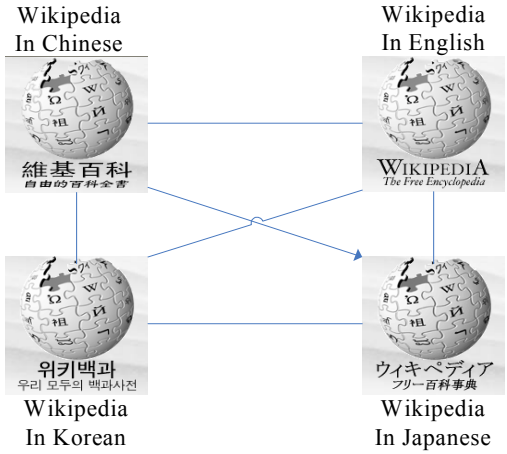

Fig 2. Using Wikipedia translation method to translate Chinese into Japanese

In MLIR, in order to handle all kinds of languages, Ballesteros introduced a transitive translation method [3] . The method based on the fact that there are already bilingual native-English and English-native dictionaries. A transitive translation can translate a query from source language A to target language B with the help of two bilingual dictionaries. For example, in Fig. 1, firstly, the query terms in Chinese are translated into English, and then the system translates the English query terms into target language Japanese. This method requires two translation steps; as a result, the decreasing of translation accuracy is unavoidable. Wikipedia translation method can direct translate query into target language without twice translation, as in Fig. 2, the query terms in Chinese are translated

into target language directly.

4. Retrieval System

The query flow of our system is shown in Fig 3. Firstly, the system segments the query in source language into terms. Secondly, the query terms are translated into target language using an online dictionary. Thirdly, the OOV terms are translated into target language using Wikipedia. At last, the IR system retrieves documents in target language based on the translated query terms.

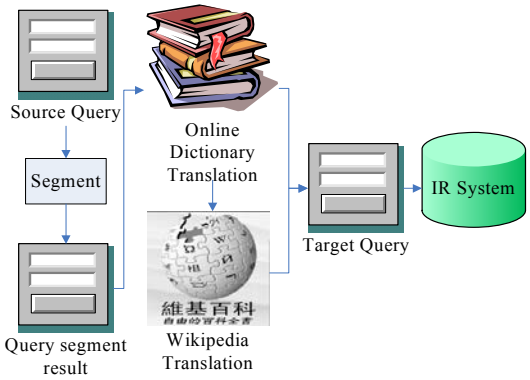

Fig 3. System flow chart

The ranking method in our system is the standard OKAPI BM25 algorithm [9] [8]. The OKAPI BM25 formulas are as follows. The similarity between a query Q and a document D_n can be computed by

$$Sim(Q,D_n)=\sum_{T\in Q} w^1 \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)}$$

Where

$$w^1=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

$$K=k_1((1-b)+b\frac{dl}{avdl})$$

N: Number of items (documents) in the collection
n: Collection frequency: number of items containing a specific term
R: Number of items known to be relevant to a specific topic
r: Number of these containing the term
tf: Frequency of occurrence of the term within a specific document
qtf: Frequency of occurrence of the term within a specific query
dl: Document length (arbitrary units)
$avdl$: Average document length
ki,b: Constants used in various BM functions

In our experiments, the default OKAPI BM25 parameters are: $k1$=1.2, $k3$=7, b=0.75, feedback new terms number=50. Retrieval process repeats three times for C-C, C-J and C-K. The system then merges the results of three Bi-lingua CLIR results into the final MLIR result. Fig. 4 shows the query translation process of our C-C, C-J, C-K runs. The details are in the following sub-sections.

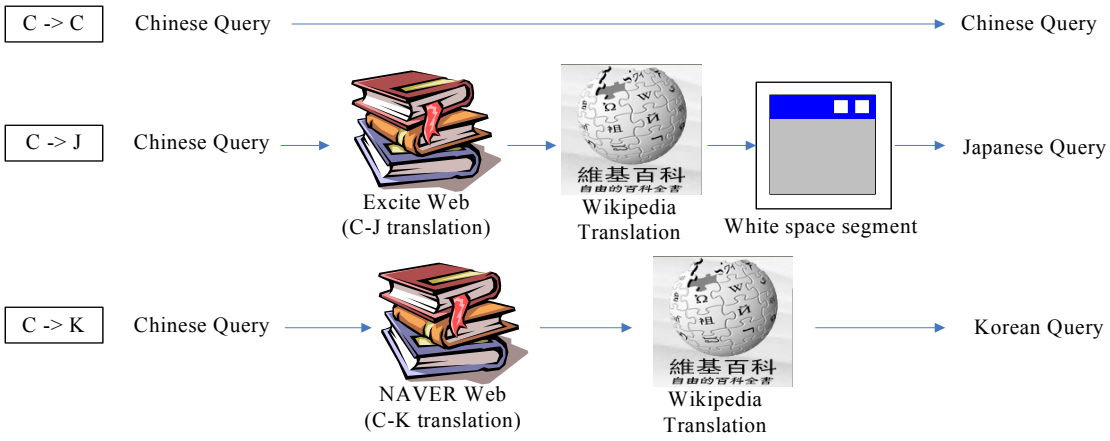

Fig 4. Query translation, the Chinese query was segmented already.

4.1 C-C Retrieval

For C-C run, Chinese query must be segmented into terms first. Because the Chinese query here is a whole sentence without space. The system then gets the retrieval results: the ranking and scoring of related Chinese documents.

4.2 C-J Retrieval

As C-C runs, the Chinese query must be segmented into several terms. For C-J translation, our system chooses the Excite (http://www.excite.co.jp/dictionary/chinese_japanese/) online translation website as the fixed Chinese-Japanese dictionary to translate the Chinese query terms into Japanese terms, and Wikipedia is adopted to translate the OOV terms. Finally, insert white space into each Japanese term. This segment process is the same as building Japanese index. These Japanese terms are target queries. The system then gets the retrieval results: the ranking and scoring of related Japanese documents.

4.3 C-K Retrieval

As C-C runs, the Chinese query must be segmented into terms. For C-K translation, our system chooses the NAVER (http://cndic.naver.com/) online translation website as the fixed Chinese-Korean dictionary to translate the Chinese terms into Korean terms, and again Wikipedia is adopted to translate the OOV terms. These Korean terms are target queries. The system then gets the retrieval results: the ranking and scoring of related Korean documents.

4.4 Merge retrieval results

Via the step in subsection 4.1, 4.2 and 4.3, the system gets three language's document rankings and scorings in Chinese, Japanese and Korean and then merges the three document rankings and sorts the document rankings according to the value of the scoring computed by OKAPI BM25. See Fig 5 for example, where Ch-001 is the ID of a document and 25.0 is the similarity to the query according to the OKAPI BM25 algorithm.

For C-CJK run, our system picked up the top 1000 document rankings for each query. And submit the top 1000 ranking results. The experiment results are shown in the following section.

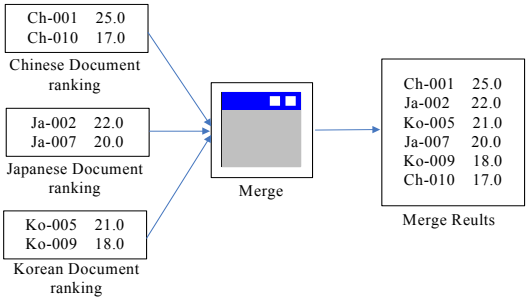

Fig 5. Ranking merge example

5. Experiment results

In NTCIR-6 MLIR task, we submitted two runs: C-CJK-T-01 and C-CJK-D-02. For each run, there are 140 queries.

1.C-CJK-T-01 run

This run considered the sentence in <TITLE> tag as query.

2.C-CJK-D-02 run

This run considered the sentence in <DESC> tag as query.

The number of documents to retrieve is listed in the following table.

Data sets:

Genre	Language	File name	Number of the documents	Year
News articles	Chinese (traditional)	CIRB040	901,446	2000-2001
	Korean	Hankookilbo	85,250	
		Chosunilbo	135,124	
	Japanese	Mainichi	199,681	
		Yomiuri	658,719	

Although, the total number of queries in the

official test set is 140. Due to the limitation of the official's budget, only 50 topics from 140 topics are picked for system's analysis. Official evaluation process reports MAP and R-Precision to estimate the system's performance. More details of the task design or procedure are in [6]. The experiment results of our official runs are shown in Table 1 and Table 2.

Table 1. The MAP of official runs

	C-CJK-T-01		C-CJK-D-02	
Recall	Rigid	Relax	Rigid	Relax
0.00	0.4884	0.6986	0.4651	0.6463
0.10	0.2278	0.3315	0.1887	0.2767
0.20	0.1384	0.2130	0.1101	0.1533
0.30	0.0888	0.1330	0.0716	0.0826
0.40	0.0489	0.0531	0.0359	0.0470
0.50	0.0256	0.0268	0.0235	0.0185
0.60	0.0129	0.0139	0.0089	0.0047
0.70	0.0051	0.0061	0.0027	0.0043
0.80	0.0016	0.0000	0.0021	0.0036
0.90	0.0000	0.0000	0.0000	0.0000
1.00	0.0000	0.0000	0.0000	0.0000
MAP	**0.0704**	**0.0992**	**0.0584**	**0.0802**

Table 2. The R-Precision of official runs

	R-Precision	
	Rigid	Relax
C-CJK-T-01	0.1528	0.1811
C-CJK-D-02	0.1357	0.1579

The rigid relevance assessment MAP values for title and description runs were 0.0704 and 0.0584. The relax relevance assessment MAP values for title and description runs were 0.0992 and 0.0802. Moreover, the rigid relevance assessment R-Precision values for title and description runs were 0.1811 and 0.1579. The relax relevance assessment R-Precision values for title and description runs were 0.1528 and 0.1357.

In addition to the official evaluation, we analyze our system further. The MAP of C-C, C-J, C-K is shown in Table 3. We can observe that the performances are quite different. We believe that were proportional to the level of the understanding of word segmentation in each language. Currently, our system segments Chinese well and segments Japanese, Korean poorly.

Table 3. The MAP of each run

	MAP	
	Rigid	Relax
C-C-T	0.2183	0.3194
C-C-D	0.1893	0.2784
C-J-T	0.0842	0.1140
C-J-D	0.0259	0.0409
C-K-T	0.0440	0.0660
C-K-D	0.0374	0.0579

Figure 6 shows the average precision of each topic in C-CJK-T-01 and C-CJK-D-02 runs. There are several queries that get bad retrieval results; we speculate the reasons are:

1. **The word segmentation result of query sentence is not correct.** Like the query 020, the noun "Y2K" in Chinese cannot be segmented correctly. As a result, the term cannot be translated correctly, either.

2. **The query terms are all common words.** Like query 019 "International incidents at Sea", the terms "International", " incidents" and "Sea" are all common words. Therefore, the top 1000 retrieval results did not include the real related document. We should deal with the query terms as compound words.

3. **The simple word segmentation strategy of Japanese is not working.** From table 3, we find our system's MAP of C-J run was much lower than the NTCIR-6 average MAP. We find that our system retrieved too many documents in Japanese. That was caused by our poor segmentation strategy of Japanese. Because we just inserted white space after each Japanese character without doing the right word segmentation, the keywords in Japanese become the common words. The retrieval results cannot focus on the keywords and the precision decrease.

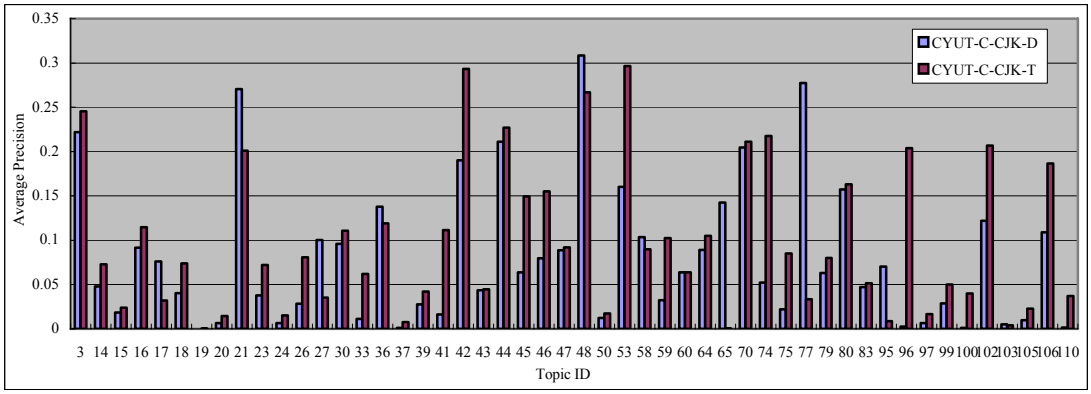

Fig 6. The average precision of each query in C-CJK-T-01 and C-CJK-D-02 runs

6. Conclusion

In MLIR, the translation of query terms requires dictionary in various languages. In this paper, we regard Wikipedia as an additional live-dictionary to translate the OOV terms since the articles in Wikipedia contain various languages and volunteers worldwide update the content daily. No free live-dictionary contains so many languages before. We find that the Wikipedia translation is a good resource that can improve the performance of a MLIR system. With the help of Wikipedia, some OOV terms can be translated into target language and the precision of MLIR increases in our experiments.

Future works

Our MAP of C-J run was much lower than the NTCIR-6 average. We speculate that was cause by the lack of right segmentation the Japanese documents. We just insert white space after each Japanese character during building Japanese index. We will segment the Japanese documents more accurately and re-building the Japanese index.

Another technique to improve the C-J run accuracy is to combine both the original query terms and the translated query terms to do the query [5]. Since Japanese documents also contain many Chinese characters. In C-J run, after translate source query into target query, we could add the originally Chinese query terms into target query. This method increases the C-J retrieval accuracy.

The document sizes of different languages are different; therefore, merging the similarity rankings for documents in different language might not be appropriate. We consider setting different document weight according to the document size before merging the CJK results to balance the effect of different corpus sizes.

Acknowledgement

This research was partly supported by the National Science Council under GRANT NSC 95-2520-S-324-001.

Reference

[1] L. Ballesteros, and W.B. Croft, “Dictionary-based Methods for Cross-Lingual Information Retrieval”, Proc. of International Conference on Database and Expert System Applications, 1996, pp 791-801.

[2] L. Ballesteros, and W.B. Croft, “Resolving Ambiguity for Cross-Lingual Information Retrieval”, Research and Development in Information Retrieval, 1998, pp 64-71.

[3] L. Ballesteros, Cross language retrieval via transitive translation. In W. B. Croft (Ed.), Advances in information retrieval: Recent research from the CIIR, Dordrecht: Kluwer Academic Publishers, 2000, pp. 203–234.

[4] M. Federico, and N. Bertoldi, “Statistical Cross-Language Information Retrieval Using N-Best Query Translations”, Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieva, Tampere, Finland, 2002, pp 167-174.

[5] F. C. Gey, “Chinese and Korean Topic Search of Japanese News Collections”, Proc. of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, April 2003 – June 2004.

[6] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen, and Sung Hyon Myaeng, “Overview of CLIR Task at the Sixth NTCIR Workshop”, In Proceedings of the Sixth NTCIR Workshop, 2007.

[7] A. Mirana, Using statistical term similarity for sense disambiguation in cross-language information retrieval. Information Retrieval 2, 1, 2000, pp.67–68.

[8] Tetsuji Nakagawa, and Mihoko Kitamura, “NTCIR-4 CLIR Experiments at Okapi”, Proc. of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, April 2003 – June 2004.

[9] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford, “Okapi at TREC-3”. In Proceedings of the Third Text Retrieval Conference (TREC-3), NIST, 1995.

[10] Chen-Yu Su, and Shih-Hung Wu, “Using Wikipedia to translate OOV term on CLIR- using Chinese-Japanese CLIR for example (in Chinese) ”, In Proceedings of the 11th Conference on Artificial Intelligence and Applications, Taiwan Kaohsiung, December 15-16 2006, pp 395-401.

[11] Ying Zhang, Phil Vines, and Justin Zobel, “Chinese OOV Translation and Post-translation Query Expansion in Chinese-English Cross-lingual Information Retrieval”, ACM Transaction on Asian Language Information Processing, Vol. 4, No. 2, June 2005, pp 55-77.

[12] Ying Zhang, and Phil Vines, “Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval”, Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom, July 25 - 29, 2004, pp 162-169.