Query Expansion from Wikipedia and Topic Web Crawler on CLIR

Meng-Chun Lin, Ming-Xiang Li, Chih-Chuan Hsu 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

In this paper, we report various strategies for query expansion (QE) in the NTCIR-8 IR4QA subtask. We submit the results of twelve runs from the formal run, which include cross-language information retrieval from English to traditional Chinese, from English to simplified Chinese, and from English to Japanese in the official T-run, D-run and DN-run. Our approach uses Google translation and the Okapi BM25 pseudo relevance feedback as the basic retrieval system. We add more QE from Wikipedia and the result of QA analysis. In the additional runs, we use a topic web crawler to get more related web pages and to extract more keywords to act as candidates for QE.

Keywords

Wikipedia, query expansion, topic web crawler.

1. INTRODUCTION

In this paper, we discuss the use of our system in the NTCIR-8 IR4QA subtask, which is a cross-language information retrieval (CLIR) evaluation test bed. Our system is designed to assist question answering (QA) systems. Our approach uses Google translation and the Okapi BM25 pseudo relevance feedback as the basic retrieval system. Since the goal is to retrieve documents that might contain information that can answer the query topics, we used the results of the question analysis from QA participants as one of our QE information sources.

In our previous works, we used Wikipedia not only as a live dictionary to overcome the out-of-vocabulary (OOV) problem, but also as an information resource to find more query extensions via the anchor text of related pages. Su et al. [13] combined Wikipedia and online translation website services for use as a live dictionary to translate the query terms in the NTCIR multi -language information retrieval task. Lin et al. [4] extracted the anchor texts in relevant Wikipedia articles to act as the candidates of query extension to improve the recall of the pure Okapi BM25 pseudo relevance feedback algorithm [10,12]. Hsu et al. [5] combined Su's[13] and Lin's[4] methods for query term translation and query expansion.

Continuing with the previous works, we investigate more strategies for query extension. First, the candidate sets for query extension are further enlarged via various information sources. Second, the query terms are combined in various proportions to get better results. To investigate the system performance of different types of questions, we acquired the analysis results of question analysis from the NTCIR-8 official website. The results provided by the participation data WHUQA-EN-CT-T-01, WHUQA-EN-CS-01, and WHUQA-EN-JA-01 are used as our information source for QE. Another information source uses a topic web crawler to get more related articles from the web. These articles are treated as pseudo relevance feedback. Our system extracts more key terms to act as candidates for QE.

The following sections are organized as follows: Section 2 describes question processing, i.e. the extraction of question terms, segmentation, and indexing. Sections 3, 4, and 5 describe the question translation method, the query expansion method, and the system architecture of our system, respectively. Section 6 shows the experiment results. In the final section, we give the conclusions and discuss future work.

1.1 Related works

There are two major difficulties with query translation in CLIR -- word sense disambiguation (WSD) and Out Of vocabulary (OOV) terms. Ballesteros and Croft [2] proposed to eliminate translation disambiguation that finds the correct term translation and uses the co-occurrence statistics method[3]. Mirna [8] proposed the term-sense disambiguation technique. Mihalcea [7] used Wikipedia to solve the WSD. In addition, Ying, Phil, and Justin [16,17] collected co-occurrences from retrieved web text using the co-occurrence method for the OOV problem.

2. QUESTION AND DATA PROCESSING

2.1 Question preprocessing

Since the 100 topics provided by NTCIR are questions in English, in the preprocessing step, our system deletes stop words in the topics and tries to find the translation in Wikipedia (http://www.wikipedia.org/) and Google translation (http://translate.google.com/).

The translation results are treated as query terms. Wikipedia is a good source for finding translations of named entities, such as personal names, organizational names, place names, and terminologies in various professional areas. In March 2010, the amount of entries in the English, Chinese, and Japanese versions of Wikipedia was 3215333, 297207, and 662360, respectively.

Google translation, on the other hand, provides translation of common terms. Our system sends the English questions to the Google translation engine and filters out the stop words in the translation result. The rest of the words are treated as query terms. These two methods can complement each other in case a term might have different translations.

Since there are five more new question types (why, person, organization, location, date), in addition to the four types (definition, biography, relationship, event) in the last IR4QA, we believe that the results from the answer type analysis of CCLQA groups might help. Therefore, the query terms used in the question analysis are also used in our system.

2.2 Data processing by different Index method

The indexing tool of our system is the Lucene toolkit (http://lucene.apache.org/). Before it can be indexed by Lucene, corpora in different languages are segmented using different tools.

2.2.1 Traditional Chinese Document Indexing

Our system uses a traditional Chinese word segmentation toolkit developed by the CKIP group (Chinese Knowledge and Information Processing) to segment the traditional 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/)

2.2.2 Simplified Chinese Document Indexing

Our system uses a simplified Chinese word segmentation toolkit developed by ICTCLAS (Institute Computing Technology, Chinese Lexical Analysis System) to segment the simplified Chinese corpus into indexing terms. The average accuracy of the toolkit is about 98%. (http://ictclas.org/index.html)

2.2.3 Japanese Document Indexing

For Japanese word segmentation, our system uses a free Japanese segmentation toolkit JUMAN (a User-Extensible Morphological Analyzer for Japanese) development by Matsumoto et al. [9]. (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)

3. QUERY EXPANSION

Query expansion [4] usually is based on the thesaurus method and the Pseudo relevance feedback. QE can help to increase the recall of information retrieval. Okapi BM25 is the most widely-used pseudo relevance feedback algorithm, which uses the result of the first retrieval as a source to extract more query terms for the second retrieval. In this paper, we also try to use a topic crawler as another source of pseudo relevance feedback.

3.1 Okapi BM25

We use the OKAPI BM25 algorithm as the basic pseudo relevance feedback [9, 11]. The OKAPI BM25 formulas are as follows. The similarity between a query *Q* and a document D_n can be computed by using

$$Sim(Q, D_n) = \sum_{T \in Q} w^1 \frac{(k_1+1)tf(k_3+1)qtf}{(K+tf)(k_3+qtf)}$$ 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 occurrences of the term within a specific document

qtf: Frequency of occurrences of the term within a specific query

dl: Document length (arbitrary units)

avdl: Average document length

ki,b: Constants used in various BM functions

3.2 Topic web crawler

A topic web crawler is a Web spider program that can retrieve only the documents related to a give topic. This kind of crawler is called a focused crawler or thematic crawler. The key difference between a focused crawler and a general crawler lies in the ability of the focused crawler to find more related documents among all available links. In a previous research, G. Pant and P. Srinivasan [11] proposed a focused crawler based on a classifier. G. Almpanidis, C. Kotropoulos, I. Pitas [1] used the lantern semantic of a webpage text and link relation to design a focused crawler. Z. Chun, J. Ma, J. Lei [6] proposed a focused crawler for both English and Chinese, which used hierarchical taxonomy to describe the topic, and integrated the Shark-Search algorithm with four different relevance prediction strategies to find related documents.

In our system, we incorporate a topic web crawler to get more related documents from the web and use them as another source of query expansion.

4. ARCHITECTURE OF OUR INFORMATION RETRIEVAL

Fig.1 shows the flow of our retrieval system. In the first stage, our CLIR system translates the queries into target language via Wikipedia and Google translate. The translated terms are used as the query terms in the first retrieval.

In the second stage, our system uses the query expansion strategy to improve our search results. In addition to the OKAPI BM25 [15], there are three more sources of query extension in our experiment. The first source is the released answer type analysis from a CCLQA group. There are many NEs that are useful, such as a query of an event with a specific time or date. The time expression, such as ” 2005年7月7日” , or ” 2003年4月24日” , can be used as part of a query in addition to the events, i.e. “倫敦地鐵爆炸事件” , and “和平醫院因SARS封院” . The second source is the anchor texts in related Wikipedia pages.

In additional runs, a third source is used. This third source is the keywords extracted from the search results of a topic web crawler.

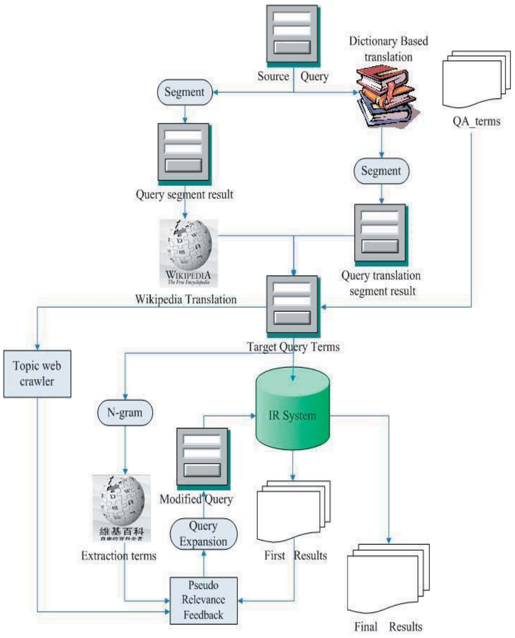

Fig.1 Architecture of retrieval system

5. EXPERIMENT RESULTS

5.1 Official Runs

Table 1 shows the different settings of our system in official runs. Table 2 lists the news corpus size of the three target languages.

Table1. Settings of official runs

Run	Setting
CYUT-EN-CT-T-01 CYUT-EN-CS-T-01 CYUT-EN-JA-T-01	Use only QUESTION field in Topic files as query terms
CYUTEN-CT-T-02 CYUT-EN-CS-T-02 CYUT-EN-JA-T-02	Adding more terms from answer type analysis of CCLQA to the first setting
CYUT-EN-CT-D-03 CYUT-EN-CS-D-03 CYUT-EN-JA-D-03	Use the NARRATIVE field in Topic file as the query terms
CYUT-EN-CT-DN-04 CYUT-EN-CS-DN-04 CYUT-EN-JA-DN-04	Combine the terms in QUESTION field and NARRATIVE field as the query terms

Table2. Data sets

Language	Data Name	Number of documents	Year
Chinese (Traditional)	UDN	1,663,517	2002-2005
Chinese (Simplified)	Xinhua	308,845	2002-2005
Japanese	Mainichi	377,941	2002-2005

Table3 shows the result of the twelve runs in three performance metrics: mean AP, mean Q-measure, and mean nDCG [14]. In the official runs, the default OKAPI BM25 parameters were: k1=1.2, k3=7, b=0.75, and the top 100 documents of the first search were treated as relevant documents. The new feedback term number was 50.

Table3. Performances of Official Runs (CS/JA results BEFORE bug fix.**)**

Run	MAP	M-Q	M-nDCG
CYUT-EN-CT-T-01	0.1733	0.1923	0.3672
CYUT-EN-CT-T-02	0.1941	0.2137	0.3963
CYUT-EN-CT-D-03	0.1362	0.1509	0.321
CYUT-EN-CT-DN-04	0.1486	0.1677	0.3516
CYUT-EN-CS-T-01	0.1955	0.2225	0.4152
CYUT-EN-CS-T-02	0.1996	0.2263	0.429
CYUT-EN-CS-D-03	0.1445	0.1674	0.3622
CYUT-EN-CS-DN-04	0.1562	0.1817	0.3933
CYUT-EN-JA-T-01	0.1708	0.1776	0.3613
CYUT-EN-JA-T-02	0.1719	0.1788	0.3638
CYUT-EN-JA-D-03	0.1023	0.1027	0.2565
CYUT-EN-JA-DN-04	0.0999	0.0985	0.2449

5.2 Additional Runs

We designed two experiments as additional runs. The evaluation toolkit was designed according to the NTCIR MAP evaluation tool. In the first experiment, we compared the proportion of the expanded query terms from two sources: Okapi BM25 and Wikipedia. We tried a total of 20 or 50 terms in different proportions between 0 to 100%.

In the second experiment, we compared the proportion of the expanded query terms from two sources: Okapi BM25 and a Topic web crawler. We tried a total of 30 or 50 terms in different proportions between 0 to 100%.

5.2.1 Experiment 1

We tested this setting on all EN-JA, EN-CT, EN-CS runs and reported the MAP.

In Tables 4, 5, and 6, the representative results of the experiments in EN-JA, EN-CT, and EN-CS are given, respectively. The result shows that more query terms from OKAPI BM25 and less query terms from Wikipedia will get a better MAP. The best proportion is about 80:20.

5.2.2 Experiment 2

We built a topic web crawler in Chinese only, because of the limitation of the language resource. Therefore, we tested this setting only on EN-CT and EN-CS runs and reported the MAP.

The original query terms were used to get the seed URLs by sending them to the Google search engine. The topic web crawler then followed the seed URLs to get more related documents. These documents were used as another kind of pseudo relevance feedback. Our system extracted the keyword in the titles and anchor texts from these documents as query expansion candidates.

In Tables 7 and 8, the representative results of experiments in EN-CT and EN-CS are given, respectively. The results show that a topic web crawler can improve the search MAP. The proportion of query terms from OKAPI BM25 or a Topic web crawler is not clear. The best proportion can be 70:30 or 40:60.

6. CONCLUSIONS

This paper reports the results of combining query terms from different sources on query expansion in CLIR. We tested this idea on EN-JA, EN-CT, and EN-CS pairs. The method in official runs combines the translation results from Wikipedia and Google translation. We conducted several additional runs to show that the combined QE is better than QE from a single source. In additional runs, we added a topic web crawler for further query expansion in EN-CT and EN-CS. The titles and anchor texts in related pages were treated as another source of QE. The experiment results show that this further expansion improved performance.

6.1 Future work

The question types of the IR4QA task increased from 4 in NTCIR-7 to 9 in NTCIR-8. This change makes the task more difficult. In the future, the IR system must use more information on the question types, such as building classifiers to relate documents to particular question types.

Table4. The performances of JA-runs; QE term=20; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE(QE term=50)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-JA-T-01	0.1628	**0.1636**	0.161	0.1603	0.1594	0.1561	0.154	0.1515	0.1428	0.1321	0.1034
CYUT-EN-JA-T-02	0.1617	**0.1625**	0.1601	0.1594	0.1583	0.155	0.1528	0.1503	0.1414	0.131	0.1024
CYUT-EN-JA-D	0.0881	0.0928	**0.0929**	0.0917	0.0907	0.0893	0.0877	0.0849	0.0822	0.079	0.058
CYUT-EN-JA-DN	0.0857	0.0904	0.0895	0.0904	**0.0905**	0.0875	0.0851	0.0822	0.0813	0.077	0.0569

Table5. The performances of CT-runs; QE term=20; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE(QE term=20)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CT-T-01	0.1738	0.1738	0.1746	0.1762	**0.1782**	0.1768	0.1752	0.1704	0.1667	0.1648	0.153
CYUT-EN-CT-T-02	0.1938	0.1935	0.1943	**0.1948**	0.1971	0.1959	0.1938	0.1911	0.1877	0.1842	0.1697
CYUT-EN-CT-D	0.1382	0.1406	**0.141**	0.1379	0.1395	0.1396	0.1381	0.1352	0.1313	0.123	0.1137
CYUT-EN-CT-DN	0.1559	0.1567	**0.1571**	0.1565	0.1567	0.1555	0.153	0.152	0.149	0.1427	0.1343

Table6. The performances of CS-runs; QE term=50; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE(QE term=50)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CS-T-01	0.2006	0.1984	0.1999	**0.2014**	0.2003	0.1965	0.1948	0.1926	0.186	0.1865	0.1707
CYUT-EN-CS-T-02	0.202	0.2014	**0.2031**	0.2028	0.2005	0.2001	0.196	0.1941	0.1894	0.1943	0.1806
CYUT-EN-CS-D	**0.1601**	0.1575	0.1566	0.156	0.1538	0.1472	0.1434	0.1421	0.1386	0.1291	0.1136
CYUT-EN-CS-DN	**0.1696**	0.1668	0.1673	0.1655	0.165	0.1572	0.1565	0.1563	0.1546	0.1489	0.1311

Table7. The performances of CS-runs; QE term=20; the different proportion in QE term from Okapi and Topic web crawler.

	Okapi QE : Topic web crawler QE(QE term=20)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CS-T-01	0.2006	0.205	0.2077	**0.2071**	0.2041	0.1945	0.1949	0.1929	0.1865	0.1846	0.1729
CYUT-EN-CS-T-02	0.202	0.2073	0.208	**0.2084**	**0.2084**	0.1998	0.2001	0.1965	0.1932	0.1937	0.1767
CYUT-EN-CS-D	0.1601	0.1638	0.1641	**0.1652**	0.1612	0.1556	0.156	0.1537	0.1496	0.1447	0.1343
CYUT-EN-CS-DN	0.1696	**0.1707**	0.1688	0.1704	0.1681	0.1606	0.1613	0.1623	0.1609	0.159	0.1472

Table8. The performances of CT-runs; QE term=30; the different proportion in QE term from Okapi and Topic web crawler.

	Okapi QE : Topic web crawler QE(QE term=30)										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
CYUT-EN-CT-T-01	0.1735	0.1769	0.1791	0.1801	0.1816	0.1824	**0.1839**	0.1808	0.1793	0.181	0.1682
CYUT-EN-CT-T-02	0.1946	0.1995	0.1999	0.2021	0.2024	0.2044	**0.206**	0.2003	0.1972	0.1974	0.1798
CYUT-EN-CT-D	0.1375	0.1388	0.1431	0.141	0.1457	**0.1461**	0.1449	0.1462	0.1413	0.1409	0.1275
CYUT-EN-CT-DN	0.1566	0.1589	0.1588	0.1614	0.1654	0.1669	**0.1676**	0.1667	0.1651	0.1651	0.1508

ACKNOWLEDGEMENT

This research was partly supported by the National Science Council under Grant No. NSC98-2221-E-324 -025.

REFERENCE

[1] G. Almpanidis, C. Kotropoulos, I. Pitas, “Combining text and link analysis for focused crawling—An application for vertical search engines”, *Information Systems*, Volume 32, Issue 6, September 2007, Pages 886-908.

[2] 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.

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

[4] Tien-Chien Lin, Shih-Hung Wu, “Query Expansion via Wikipedia Link”, *International Conference on Information Technology and Industrial Application*, 2008.

[5] Chih-Chuan Hsu, Yu-Te Li, You-Wei Chen, Shih-Hung Wu, “Query Expansion via Link Analysis of Wikipedia for CLIR”, *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19, 2008,

pp.125-131.

[6] Zhumin Chen, Jun Ma, Jingsheng Lei, Bo Yuan, Li Lian, Ling Song, “A cross-language focused crawling algorithm based on multiple relevance prediction strategies ”, *Computers & Mathematics with Applications,* Volume 57, Issue 6, March 2009, Pages 1057-1072.

[7] Rada Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation", *Proceedings of NAACL HLT*, 2007, pp. 196–203.

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

[9] Yuji Matsumoto, Sadao Kurohashi, Yutaka Nyoki, Hitoshi Shinho, and Makoto Nagao. "User's Guide for the Juman System, a User-Extensible Morphological Analyzer for Japanese. Version 0.5", Kyoto University. (in Japanese).

[10] 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.

[11] G. Pant, P. Srinivasan, “Learning to Crawl: Comparing Classification Schemes” , *ACM Transactions on Information Systems(TOIS)*, Vol. 23, No. 4, October 2005, Pages 430–462.

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

[13] Chen-Yu Su, Tien-Chien Lin, Shih-Hung Wu, “Using Wikipedia to Translate OOV Terms on MLIR”, *Proceedings of NTCIR-6 Workshop Meeting*, May 15-18, 2007, pp. 109-115.

[14] Tetsuya Sakai, Hideki Shima, Noriko Kando, Ruihua Song, Chuan-Jie Lin, Teruko Mitamura*,* Miho Sugimoto, “Overview of NTCIR-8 ACLIA IR4QA ”, *In Proceedings of the 8th NTCIR Workshop*, 2010.

[15] Fan, Weiguo, Luo, Ming, Wang, Li, Xi, Wensi and Fox, Edward A. , “Tuning Before Feedback: Combining Ranking Discovery and Blind Feedback for Robust Retrieval”, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 138-145.

[16] 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.

[17] 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.