Query Expansion via Link Analysis of Wikipedia for CLIR

Chih-Chuan Hsu, Yu-Te Li, You-Wei Chen 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 how we do the query expansion in the NTCIR-7 IR4QA subtask. We submit the results of nine runs of this subtask, which is cross-language information retrieval (CLIR) from English to traditional Chinese, simplified Chinese and Japanese in the official T-run, D-run and DN-run. In these runs, we use the Google online translation service to translate query terms and Wikipedia as an information resource for query expansion (QE), in addition to the OKAPI query expansion.
Keywords: *Wikipedia, query expansion, CLIR*

1. Introduction

In this paper, we show how Wikipedia can help with CLIR, such that a system can better satisfy users' needs for information and so that the users can get proper information in different languages. We submit the results of nine runs of the NTCIR-7 IR4QA subtask , which is cross-language information retrieval from English to traditional Chinese (EN-CT), simplified Chinese (EN-CS) and Japanese (EN-JA) in the official T-run, D-run and DN-run. Since we are not concerned with the question answering task, we treated the questions as IR queries and did not process the questions in advance.

There are two major difficulties with query translation in CLIR, word sense disambiguation (WSD) and Out Of vocabulary (OOV) terms. Without WSD, the query terms in the source language might be translated to the target language incorrectly. To solve the WSD problem, Ballesteros and Croft proposed the co-occurrence statistics method [2], Mirna proposed a term-sense disambiguation technique [5], Mihalcea proposed using Wikipedia [4]. In addition, Ying, Phil and Justin collected co-occurrences from the retrieved web text using the statistics method [12, 13] to translate the Chinese OOV terms.

In this paper, we focus on dealing with the OOV terms. In a previous research, Su et al. [9] adopted online translation website services as a fixed dictionary and Wikipedia as a live dictionary to translate query terms. Their method can translate OOV terms efficiently; we use this method in our system to translate query terms. However, there are terms for which translations cannot be found. In order to retrieve more relevant documents, we adopt the algorithm OKAPI BM25 [7, 8] to help our query expansion to raise recall. Furthermore, Lin et al. [3] purposed a method that combines OKAPI BM25 and Wikipedia anchor texts for query expansion. In this paper, we combine Su's and Lin's methods in our system.

The following sections are organized as follows: sections 2, 3, 4 and 5 describe the index methods, the translation methods, the query expansion methods, and the retrieval methods respectively. We show the experiment results in section 6 and give the conclusions and future work in section 7.

2. Index Method

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

Traditional Chinese Document Indexing

Our system adopts 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/)

Simplified Chinese Document Indexing

Our system adopts 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)

Japanese Document Indexing

For Japanese word segmentation, we use a free Japanese segmentation toolkit "JUMAN" (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html) development by Matsumoto et al. [6].

3. Query Translation

Dictionary-Based Translation

We use the dictionary-based translation method to translate the source language query into the target language query with a fixed bilingual dictionary. In this paper, the existing free online translation website services are regarded as the fixed dictionaries. Furthermore, we translated query terms with Google Translate (http://translate.google.com/). We translated the full texts, not segmented query terms. More details of the query translation in our system are given in section 5.

Wikipedia Translation

Wikipedia is a multilingual encyclopedia on the web and is composed and edited by volunteers all over the world. In addition, Wikipedia has varied and the latest information because it can be updated at any time. The number of English articles is more than 2.5 million, of Chinese articles, more than 0.2 million, and of Japanese articles, more than 0.5 million. In total, there are more than eleven million articles in 264 languages in Wikipedia. The numbers of articles is still increasing. (http://www.wikipedia.org)

Each entry in Wikipedia has links to entries in other languages if there are entries of the same topic in those languages [9]. The translation of an entry can be found just by following the link to the target language if the translation in the target language is available. Therefore, Wikipedia can be seen as a live dictionary having multiple languages. Additionally, Wikipedia entry titles are mostly proper nouns. Proper nouns help with IR than regular nouns since most query terms are proper nouns.

4. Query Expansion

Query expansion is an important technology in IR systems since it can increase recall value. There are two major approaches: the thesaurus method and the Pseudo relevance feedback. The pseudo relevance feedback method extracts relevant terms from the result of the first retrieval and uses them as expanded queries to retrieve documents again. We combine these two methods in our experiments by treating Wikipedia as a kind of thesaurus.

OKAPI BM25

We adopt the OKAPI BM25 algorithm as the basic pseudo relevance feedback [7, 8]. 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

Wikipedia Query Expansion

In Wikipedia, every entry has links to related entries or other relevant web pages on other websites. The anchor texts of the hyperlinks must be related terms. Therefore, we treat these anchor texts as candidates for query expansion [3].

5. Retrieval System

Figure 1 shows a flowchart of our system. There are two parts in the translating query. In the first part, the query in the source language is translated into the target language using an online translation service. The system segments the query in the target language into terms. In the second part, the system segments the query in the source language into terms, and the query terms are translated into target language using Wikipedia. Finally, the translated query terms from the two parts are combined and the IR system retrieves documents in the target language based on these. In our EN-CT, EN-CS and EN-JA runs, the system follows the same flow.

On the one hand, to expand the query, our system segments these query terms and gets the relevant anchor texts from Wikipedia. On the other hand, our system uses the query of the target language to first retrieve the pseudo-relevant feedback (PRF) [11] to get the relevant documents of the top *R* and then uses the Okapi BM25 to rank these terms. Finally, we use the top *r* as the new QE terms.

In the official runs, the default OKAPI BM25 parameters are *k1*=1.2, *k3*=7, b=0.75, and the top 100 documents of the first search are treated as relevant documents. The feedback new terms number=50.

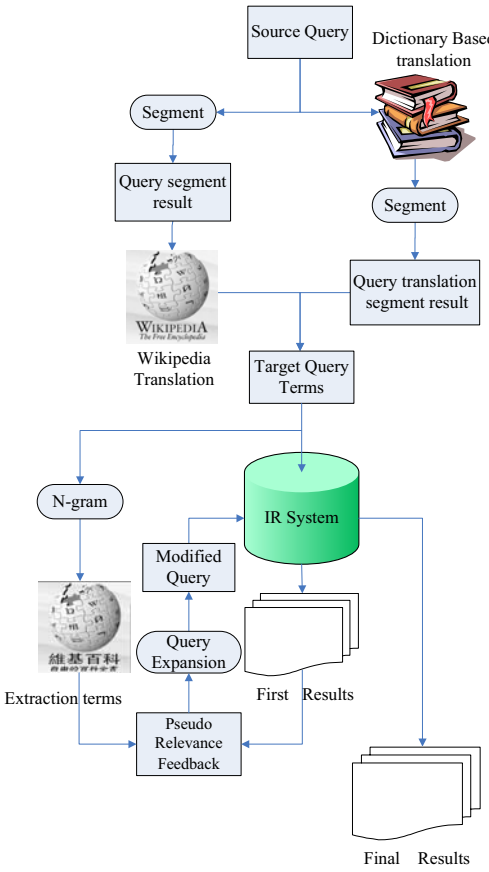

Figure 1. System flowchart

6. Experiment results

Official Run

In NTCIR-7 IR4QA task, we submitted nine runs: EN-CT-T-01, EN-CT-D-02, EN-CT-DN-03, EN-CS-T-01, EN-CS-D-02, EN-CS-DN-03, EN-JA-T-01, EN-JA-D-02, and EN-JA-DN-03. For each run, there are 100 queries.

Run Type:

T-run: only the QUESTION field is used.

D-run: only the NARRATIVE field is used.

DN-run: both the QUESTION and NARRATIVE fields are used.

The number of documents to be retrieved is listed in Table1.

Table 1. Data sets.

Language	File name	Number of the docs	Year
Chinese (Traditional)	CIRB020	249,203	1998-1999
	CIRB040	901,446	2000-2001
Chinese (Simplified)	Xinhua	535,610	1998-2001
	Lianhe Zaobao		
Japanese	Mainichi	419,759	1998-2001

The total number of queries in the official test set is 100 queries in each language. Due to the limitation of the official budget, in CS only 97 topics from 100 topics were picked for system analysis, in CT only 95 topics from 100 topics, and in JA only 98 topics from 100 topics. Official evaluation process reports use the mean AP, the mean Q-measure, and the mean nDCG to estimate the system's performance. More details of the task design and procedure are given in [10], since the candidates are extracted from different ways in our method and in Okapi. We mixed the candidates with 20% from Okapi and 80% from Wikipedia. The experimental results of our official runs are listed in Table 2.

Table 2. The performances of official runs

Run	MAP	M-Q	M-nDCG
EN-CT-T-01	0.2590	0.2747	0.4752
EN-CT-D-02	0.2458	0.2620	0.4612
EN-CT-DN-03	0.2516	0.2648	0.4638
EN-CS-T-01	0.3781	0.3936	0.6115
EN-CS-D-02	0.3726	0.3880	0.6057
EN-CS-DN-03	0.4238	0.4386	0.6578
EN-JA-T-01	0.2543	0.2528	0.4252
EN-JA-D-02	0.2294	0.2300	0.4124
EN-JA-DN-03	0.2568	0.2545	0.4366

Additional Runs

In addition to the results of the official runs, we would like to know if our method is helpful in single language information retrieval. We also want to know the effect of different numbers of QE terms and different proportions of QE terms from Okapi and Wikipedia. We conducted more experiments as follows:

1. The first experiment tests our method on single language runs: CT-CT, CS-CS, and JA-JA. The performance will be compared with CLIR.

2. The second compares the results from runs of 10, 20, 30, 40, and 50 terms on query expansion.

3. Finally, we compare the different proportions in QE terms from Okapi and Wikipedia. Proportions are ranged from 0% to 100%.

In Table 3, 4, and 5, the representative results of experiments in CT, CS, and JA are given, respectively. The translating queries, especially in EN-CS performs better than the others. The value of MAP in EN-CS and CS-CS is very close and is shown in Table 4. Although the performance of EN-JA is not bad, our MAP of the EN-JA run was much lower than the NTCIR-7 average.

With regards to the ratio in Wikipedia and Okapi, QE terms from Okapi give better results in most cases. However, QE terms from Wikipedia give better result in DN-runs in each language. Furthermore, the MAP of QE terms from Okapi only and the MAP of QE terms from Wikipedia only are quite close. This is an interesting result, because the QE terms are not very close. As we show in Table 6, the intersection rate of QE terms from two different methods are quite low. The extreme case is in the CT-CT-DN run. Whereas the MAP of OKAPI QE is 0.3920 and the MAP of Wikipedia is 0.3955, the intersection of the expanded query terms is only 10%. That is, our system gets the same performance with almost different query terms.

We also discover that in our experiments, the MAP values of many DN-runs are better than those of T-runs. The phenomenon is most prominent in CS-runs and is shown in Table 4.

Table 3. The performances of CT-runs; QE term=20; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
EN-CT-T	0.2681	0.2770	0.2786	0.2776	0.2765	0.2806	0.2814	**0.2840**	0.2797	0.2739	0.2680
EN-CT-D	0.2635	0.2705	0.2746	0.2723	0.2695	0.2729	0.2734	**0.2765**	0.2661	0.2588	0.2561
EN-CT-DN	0.2484	0.2570	0.2616	0.2628	0.2619	0.2627	**0.2639**	0.2627	0.2563	0.2483	0.2416
CT-CT-T	0.4211	**0.4318**	0.4264	0.4185	0.4122	0.4123	0.4119	0.4160	0.4131	0.4061	0.4008
CT-CT-D	0.3922	**0.4018**	0.3980	0.3897	0.3824	0.3774	0.3779	0.3795	0.3766	0.3713	0.3650
CT-CT-DN	0.3920	0.4037	0.4043	0.4017	0.4027	0.4026	0.4059	**0.4088**	0.4061	0.4017	0.3955

Table 4. The performances of CS-runs; QE term=50; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
EN-CS-T	0.4032	**0.4039**	0.3964	0.3912	0.3900	0.3866	0.3840	0.3789	0.3781	0.3684	0.3470
EN-CS-D	**0.4017**	0.4009	0.3944	0.3889	0.3869	0.3861	0.3838	0.3750	0.3726	0.3640	0.3413
EN-CS-DN	0.4047	0.4077	0.4094	0.4126	0.4176	0.4189	0.4208	0.4226	**0.4238**	0.4161	0.4037
CS-CS-T	**0.4231**	0.4139	0.4035	0.3978	0.3984	0.3998	0.3981	0.3997	0.4005	0.3956	0.3891
CS-CS-D	**0.4206**	0.4136	0.4049	0.3998	0.4015	0.4027	0.4042	0.4022	0.4021	0.3954	0.3851
CS-CS-DN	0.4250	0.4247	0.4232	0.4249	0.4305	0.4344	0.4407	0.4429	0.4435	**0.4443**	0.4409

Table 5. The performances of JA-runs; QE term=40; the different proportion in QE term from Okapi and Wikipedia.

	Okapi QE : Wikipedia QE										
Run	**100:0**	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**	**40:60**	**30:70**	**20:80**	**10:90**	**0:100**
EN-JA-T	**0.2744**	0.2687	0.2606	0.2586	0.2617	0.2625	0.2611	0.2590	0.2557	0.2451	0.2310
EN-JA-D	**0.2391**	0.2330	0.2290	0.2264	0.2297	0.2310	0.2302	0.2297	0.2305	0.2251	0.2115
EN-JA-DN	0.2571	0.2550	0.2577	0.2571	0.2580	0.2566	**0.2580**	0.2571	0.2558	0.2487	0.2328
JA-JA-T	**0.2702**	0.2545	0.2485	0.2408	0.2423	0.2410	0.2407	0.2418	0.2397	0.2407	0.2220
JA-JA-D	**0.1855**	0.1812	0.1721	0.1703	0.1720	0.1685	0.1684	0.1679	0.1675	0.1613	0.1503
JA-JA-DN	**0.1740**	0.1696	0.1683	0.1679	0.1692	0.1691	0.1685	0.1681	0.1669	0.1629	0.1490

Table 6. In CT-runs, the intersection rate QE terms by using full okapi QE and full Wikipedia QE; QE term=20; We calculated Intersection Rate.

$$\text{Intersection Rate} = \frac{1}{T}\sum_{i \in T} \frac{\#\,\text{WikiQE}_i \cap \text{OkpaiQE}_i}{\#\,\text{WikiQE}_i \cup \text{OkapiQE}_i} \times 100\%$$

Where T is number of topics, WikiQE is Wikipedia QE terms, OkapiQE is QE terms and # is number of terms.

Run	Intersection Rate (%)
EN-CT-T	31.88
EN-CT-D	31.67
EN-CT-DN	10.90
CT-CT-T	30.73
CT-CT-D	29.77
CT-CT-DN	10.38

7. Conclusion

In this paper, we make use of Wikipedia, a good information resource, to improve our precision in CLIR. We combine Su's [9] and Lin's [3] methods for query term translation and query expansion. When Wikipedia, with multilingual characteristics, is used, a system can translate English OOV terms to traditional Chinese, simplified Chinese, and Japanese. We find that this is especially useful in EN-CS runs. Furthermore, we use the Wikipedia as the thesaurus of QE by treating the anchor texts as candidates for QE. We discovered that it could be helpful in several cases in Wikipedia, especially in CT-runs and DN-runs in each language.

Future works

In our experiment, the MAP values from using Wikipedia QE are unstable. This might be due to the segmentation method of the anchor texts in CT, CS, and JA. Our system did not segment them well and led to the low TF-IDF of some relevant terms. We will try better segmentation methods on the anchor texts and observe if this will improve the performance.

We only use the "LongQE" method [3] in our system. We will use the "ShortQE" in the future and compare the performances of these two methods in EN-CT, EN-CS, EN-JA, CT-CT, CS-CS, and JA-JA .

Acknowledgement

This research was partly supported by the National Science Council under NSC 96-2221-E-324-046.

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

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

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

[6] 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)

[7] 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.

[8] 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.

[9] 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.

[10] Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Donghong Ji, Kuang-Hua Chen, Eric Nyberg, , "Overview of the NTCIR-7 ACLIA IR4QA Subtask", In Proceedings of the 7th NTCIR Workshop, 2008.

[11] 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.

[12] 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.

[13] 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.