Wikipedia and Web document based Query Translation and Expansion for Cross-language IR

Ling-Xiang Tang[1], Andrew Trotman[2], Shlomo Geva[1], Yue Xu[1]
1Faculty of Science and Technology,
Queensland University of Technology,
Brisbane, Australia
{l4.tang,s.geva, yue.xu}@qut.edu.au
[2]Department of Computer Science,
University of Otago,
Dunedin, New Zealand
andrew@cs.otago.ac.nz

ABSTRACT

In this paper we describe our approaches to retrieving cross-lingual documents for question answering in the NTCIR ACLIA-IR4QA task. A few Chinese indexing techniques were used in our experiments. We mainly focused on using external recourses: web documents and Wikipedia for the key phrase identification, translation and query expansion. The evaluation shows encouraging results of our system.

Categories and Subject Descriptors

H.3.1 Content Analysis and Indexing – *Indexing methods, Linguistic processing*; H.3.3 Information Search and Retrieval – *Query formulation, Search process*.

General Terms

Algorithms, Design, Experimentation.

Keywords

NTCIR, Wikipedia, NGMI, Dual Indexing.

1. Introduction

In NTCIR-8, we participated in the Advanced Cross-lingual Information Access (ACLIA) IR4QA task, and submitted runs for the English to Chinese CLIR task (both English-to-Simplified Chinese and English-to-Traditional Chinese).

IR4QA is an embedded component of an overall cross-lingual question answering (CLQA) system [1, 2]. The answer for a question could be either the result of an End-to-End QA system, or a combination of different QA systems and IR systems. In both cases, the candidate answers for the question are extracted from documents retrieved in a CLIR stage. CLIR plays a very important role in an overall question answering system because the relevancy of retrieved passages or documents determines the accuracy of the answer.

A simple approach to achieving CLIR is to translate the query into the language of the documents and to use a mono-lingual IR system. However, for this it is essential to identify the key phrases in the question and to correctly translate them.

A good way to achieve high accuracy for query translation is to apply an English POS tagger[3] to the query, to extract the key phrases, and then to translate them. An alternative is to use a statistical machine translation toolkit[4] to obtain bi-lingual phrase pairs and to translate on a phrase by phrase basis. Nowadays, dictionary-based and statistical machine translation can achieve very high accuracy levels when translating general text. However, the complex phrases and possible ambiguity present in a question tax general purpose machine translation approaches. Out of vocabulary terms are particularly a problem.

Example 1: Question and three translations

Question	*What is the relationship between the movie "Riding Alone for Thousands of Miles" and ZHANG Yimou*
Google Translate [1] (traditional Chinese)	*什麼是電影的關係“單騎千里”和張藝謀*
Google Translate (simplified Chinese)	之间有什么电影“利民为千里单独的关系”和张艺谋
correct translation	张艺谋与电影“千里走单骑” 是什么关系

In example 1, neither of the Google translations for ZHANG Yimou's movie "Riding Alone for Thousands of Miles" (千里走单骑) is correct. For the traditional Chinese version of the movie name, the translation is partly correct, and it is reasonable to expect it will be an effective query. But for the simplified Chinese version, the translated movie title is nonsense and is full of noise words. In this example, "Riding Alone for Thousands of Miles" is an out- of-vocabulary phrase.

Rather than using the traditional machine translation methods described above, we adopted a far simpler strategy. We use a web search engine (Google) to locate a related Wikipedia entry for the question, and then use the title of Chinese page for translation. Query expansion is done using Chinese text (called *clue text*) encountered in the English results. A detailed description of the approach is presented in section 3.2.

2. Chinese Document Indexing

More than the case with western languages, the methods used in IR indexing of Chinese text vary between research groups.

[1] http://translate.google.com.

Unigrams, bigrams and words are all common tokens used when indexing Chinese text. Different word segmentation algorithms might be used as might different ranking functions. The performance of various IR systems combining different models can vary substantially [5, 6].

In our experiments for IR4QA, we used n-gram mutual information (NGMI) [7] to segment Chinese text. NGMI is an unsupervised n-gram word segmentation approach. It is derived from character-based mutual information, but can additionally recognize words longer than two characters.

In order to find the most suitable segmentation strategy for our CLIR system, unigram indexing and dual indexing (unigrams and NGMI segmentation) were used in different experimental runs.

3. QUERY PROCESSING

3.1 Question Analysis

Example 2: English-to-Chinese NTCIR8- ACLIA topics

topic id	ACLIA2-CS-0002
question(en)	What is the relationship between the movie "Riding Alone for Thousands of Miles" and ZHANG Yimou?
question(zh)	《千里走单骑》和张艺谋是什么关系？
narrative(en)	The user wants to know the relationship between the movie "Riding Alone for Thousands of Miles" and ZHANG Yimou.
narrative(zh)	使用者想知道《千里走单骑》和张艺谋之间有何关系。
topic id	ACLIA2-CT-0024
question(en)	Give short descriptions to the isolation of the Heping Branch of Taipei city hospital caused by SARS on 24th of April 2003.
question(zh)	[請簡述 2003 年 4 月 24 日，和平醫院因 SARS 封院的始末。
narrative(en)	The analyst would like to know the truth that Taipei City Hospital - Heping Branch had to be isolated from others because of group SARS cases.
narrative(zh)	使用者想知道 2003 年 4 月 24 日，和平醫院爆發集體感染 SARS 病例而封院的經過

Example 2 shows two queries from the English-to-Chinese NTCIR8-ALCIA topic set. They are complex questions and it is important to be able to precisely translate them.

Commonly, topics are classified into different domains and receive special processing accordingly. The quality of the query can be improved by providing additional terms for query expansion or removing noise words. For example, recognising a biography question can help in query expansion by appending the query with the term “born”; or removing the noise phrase “是什么关系 (what is the relation between)” for relationship questions. In NTCIR 8 ACLIA there are a few pre-defined question types: DEFINITION, BIOGRAPHY, RELATIONSHIP, EVENT, WHY, PERSON, ORGANIZATION, LOCATION and DATE.

Rather than using complex natural language processing, we employ a naive question processing method that simply breaks the sentence down into term chunks by removing stop words. The stop word list contains 313 words extracted from Medline[8]. The benefit of this method is that it is question independent. Table 1 presents two examples of the term-chunks after stop words have been removed.

Table 1. Example term chunks after stop word removal

Topic ID	Terms Chunk
ACLIA2-CT-0001	so-called steroid
ACLIA2-CT-0024	give short descriptions, isolation, Heping Branch, Taipei city hospital caused, SARS, 24 th April, 2003

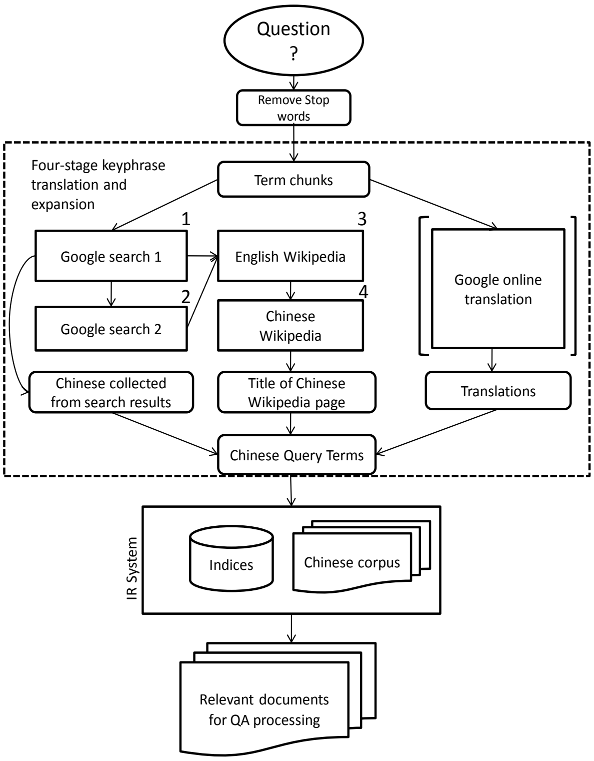
Question
?
Remove Stop words
Four-stage keyphrase translation and expansion
Term chunks
1
3
Google search 1
English Wikipedia
Google online translation
2
4
Google search 2
Chinese Wikipedia
Chinese collected from search results
Title of Chinese Wikipedia page
Translations
Chinese Query Terms
IR System
Indices
Chinese corpus
Relevant documents for QA processing

Figure 1. The design of CLIR system.

3.2 CLIR System Architecture

We adopted a simple all-in-one strategy for query key term recognition, translation, query expansion, and to address the out- of-vocabulary problem. It is depicted in Figure 1.

A Google search for the best English Wikipedia pages is done using the English question. The Chinese equivalent page is found by following the language link. Finally the title of the Chinese Wikipedia page is used as the translation of the query. If this fails we use Google Translate. This approach relies on following observations:

- The Chinese Wikipedia has over 100,000 entries describing various events, people, organizations, locations, and facts. Most importantly, there are links between English articles and their corresponding Chinese counterparts.
- When people post information on the Internet, they often provide a translation (where necessary) in the same web documents. These pages contain bi-lingual phrase pairs.

- A web search engines such as Google identify Wikipedia entries and bi-lingual web documents that are closely related to a query.

In summary, we use three different sources for query translation in our strategy. They are listed in Table 2.

Table 2: Translation sources

Name	Method
Google	bi-lingual web documents
Wikipedia	multilingual encyclopedia
Google Translate	machine translation [9]

3.3 The Query Translation Algorithm

The Chinese translation of the English question is generated from the following steps:

1. Remove stop words from the English question to generate term chunks.

2. Search Google using the term chunks with options set to return only the Chinese and English page. If there is a Wikipedia page at the top 10 then go to step 4, else go to step 3. In both cases Chinese text present in the English results list is collected – this clue text is used for query expansion.

3. Calculate the frequency of words appearing in the title of the results from step 2. Choose the top ranking ones with a frequency above a threshold. The threshold is set to 3 to begin with then linearly decreased by one until at least one term is found. Search Google using these words, but restrict the search to the English Wikipedia. The highest ranking result containing any term from the question is considered the correct result.

4. Retrieve the English Wikipedia page.

5. Retrieve the corresponding Chinese Wikipedia page by following the languages link from the English page. If there is no Chinese link found, then jump to step 7.

6. Extract the title of the Chinese Wikipedia page, and use it as the query.

7. Process the clue text if there is any for query expansion. Either the whole or some frequent words of the clue text could be used as expansion terms. Append these additional terms to the Chinese query. The frequent words are chosen from the NGMI segmented terms in the clue text. The frequency of those words has to be above a threshold. The threshold is set to 5 to begin with, then linearly decreased by one until at least one term is found.

8. If no Chinese query terms have been found then use the translation from Google Translate on the term chunks.. Otherwise, the results of this machine translation are optionally appended to the query.

9. Search the corpus using either: single character segmentation; or single character segmentation plus NGMI segmented terms.

The resulting query can be thought-of as:

Chinese query

=

The title of the Chinese Wikipedia document found by searching Google for the English Wikipedia document and following the languages link

+

Chinese clue text collected from Google search (all or high frequency words)

+

Result from Google Translate (optional)

4. Weighting Model

A slightly modified BM25 ranking function was used for document ordering.

When calculating the inverse document frequency, we use:

$$\mathrm{IDF}(q_i) = \log\frac{N}{n} \qquad (1)$$

Where N is the number of documents in the corpus, and n is the document frequency of query term q_i. The retrieval status value of a document d with respect to query $q(q_1, \dots, q_m)$ is calculated as:

$$\mathrm{rsv}(q,d) = \sum_{i=0}^{m} \frac{\mathrm{tf}(q_i,d) * (k_1+1)}{\mathrm{tf}(q_i,d) + k_1 * \left(1 - b + b * \frac{\mathrm{len}(d)}{\mathrm{avgdl}}\right)} * \mathrm{IDF}(q_i) \qquad (2)$$

Where $\mathrm{tf}(q_i, d)$ is the term frequency of term q_i in document d; $\mathrm{len}(d)$ is the length of document d in words and avgdl is the mean document length. When indexing using the dual indexing strategy only the actual words (not the unigrams) are counted towards the document length. The values of tunable parameters k_1 and b used in our experiments are: 0.9 and 0.4 respectively.

5. Experiments

We used combinations of different segmentation approaches and query translation approaches to generate five different runs for each English-to-Chinese task. The detail of the differences between these runs is given in Table 7 and Table 8.

The 01 and 02 runs used blind translation from Google and Wikipedia with translation from the online translation service (if Wikipedia and web document based translation failed). The 03, 04, and 05 runs combine the translation results from three different sources: web documents, Wikipedia, and machine translation.

We are also interested in investigating how segmentation affects cross-lingual IR. To test this we employed two different indexing and query segmentation strategies in our runs. The 01 and 03 runs used only single Chinese character indexing and searching. The 02, 04, and 05 runs used dual indexing and searched the documents using single characters and words.

In our experiments web documents form an important data source for query expansion. As the source language of the web documents may affect the results, we created two additional runs to examine this. Run EN-CS 05 searched only in the *cn* domain; while run EN-CT 05 search only in the *tw* domain.

Overall, the 01, 02, 03, 04 runs share the same Google search results and clue text, but run 05 is different. For different experimental run groups, the statistics of the number of discovered Wikipedia pages, collected clue text and total topics in which web search failed are given in Table 3 and Table 4. From those two tables we can see that a large portion, almost 9/10, of the questions found their related Wikipedia pages. Using bilingual search on Google contributes about 1/3 of the clue text. It also can be seen that domain search is a good way to increase the chance of finding more clue text as the amount of clue text nearly triples for the EN-CS runs, and quadruples for the EN-CT runs. However, it reduces the possibility of finding the relevant Wikipedia page because of the restrictive search.

Table 3. The statistics of the EN-CS runs. #FAIL is the number of total topics for which the web search could not obtain either a Chinese Wikipedia page or Chinese clue text.

Run	# EN Wiki	# ZH Wiki	# Clue Text	# FAIL
01, 02, 03, 04	86	71	36	25
05	70	52	96	2

Table 4. The statistics of the EN-CT runs. #FAIL is the number of total topics for which the web search could not obtain either a Chinese Wikipedia page or Chinese clue text.

Run	# EN Wiki	# ZH Wiki	# Clue text	# FAIL
01, 02, 03, 04	89	65	24	27
05	57	43	93	6

6. IR4QA Results AND DISCUSSION

The official evaluation results (before bug fixes) of our submissions on the two tasks are given in Table 5 and Table 6. It can be seen that the distributions of the mean average precision (MAP), mean Q (MQ), and mean nDCG (MnDCG) are similar for both the EN-CS and EN-CT tasks.

Runs 01 and 02 that use just web documents and the Wikipedia (where possible) achieve a relatively low precision. This, however, does indicate that web documents and the Wikipedia may be good resources for query translation.

The runs using NGMI word dual indexing (runs 02, 04, and 05) bettered the equivalent character based indexing in all cases.

In both cases the runs that restricted the web search to a particular domain (run 05) performed worse than the equivalent run that did not site restrict. As can be observed from Table 3 and Table 4, the probability of finding clue text increases, and the clue text collected in Google results gets larger due to the specified Chinese domain search. However, as a side-effect, noise words increase, and the hope for finding a relevant Wikipedia page become slim. Therefore, the performance of runs with site restriction could deteriorate if the increased noise words are not taken care of.

It is also interesting to see that the QUTIS-EN-CS runs, QUTIS-EN-CS-01-T particularly, contribute the highest number of unique relevant documents in all CS submissions according to the official NTCIR assessment results[10]. This could be largely because our Wikipedia and web document based query translation and expansion approach contributes extra unexpected good query terms.

Overall, our runs did not perform well when compared to the best run submitted to the task and so care must be taken when drawing conclusions.

7. CONCLUSIONS AND FUTURE WORK

We implemented a simple strategy for query translation and expansion for cross-lingual question answering. This strategy relies on an external resource such as web documents and the Wikipedia to tackle the out-of-vocabulary problem. The performance of our system and highest number of unique relevant documents found in EN-CS experimental runs are encouraging. In future work we will use logical and contextual analysis to look for better translations of out-of-vocabulary phrases. We will also continue our work in using the Wikipedia as a method of translation and query expansion.

Table 5: Official EN-CS IR4QA results BEFORE bug fixes

RUN ID	MAP	MQ	MnDCG
EN-CS best	0.4139	0.4499	0.6509
QUTIS-EN-CS-01-T	0.1420	0.1689	0.3527
QUTIS-EN-CS-02-T	0.1673	0.1967	0.4028
QUTIS-EN-CS-03-T	0.2504	0.2886	0.5127
QUTIS-EN-CS-04-T	**0.3198**	**0.3607**	**0.5882**
QUTIS-EN-CS-05-T	0.2752	0.3086	0.5245

Table 6: Official EN-CT IR4QA results BEFORE bug fixes

RUN ID	MAP	MQ	MnDCG
EN-CT best	0.4900	0.5263	0.7175
QUTIS-EN-CT-01-T	0.1943	0.2218	0.3997
QUTIS-EN-CT-02-T	0.2161	0.2501	0.4374
QUTIS-EN-CT-03-T	0.2656	0.2957	0.4905
QUTIS-EN-CT-04-T	**0.3231**	**0.3569**	**0.5555**
QUTIS-EN-CT-05-T	0.1040	0.1167	0.2492

8. References

[1] T. Mitamura, *et al.*, "Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access," in *Proceedings of NTCIR-7*, 2008, pp. 11-25.

[2] T. Mitamura, *et al.*, "Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access," in *Proceedings of NTCIR-8*, to appear, 2010.

[3] L. Shi, *et al.*, "RALI Experiments in IR4QA at NTCIR-7," in *NTCIR-7*, 2008, pp. 115-124.

[4] W. Luo, *et al.*, "ICT-Crossn: The System of Cross-lingual Information Retrieval of ICT in NTCIR-7," 2008, pp. 132-139.

[5] W. P. L. Robert and K. L. Kwok, "A comparison of Chinese document indexing strategies and retrieval models," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 1, pp. 225-268, 2002.

[6] A. Chen, *et al.*, "Chinese text retrieval without using a dictionary," in *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42-49.

[7] L.-X. Tang, *et al.*, "Word Segmentation for Chinese Wikipedia Using N-Gram Mutual Information," presented at the 14th Australasian Document Computing Symposium (ADCS 2009), University of New South Wales, Sydney, 2009.

[8] U.S. National Library of Medicine. *MEDLINE®/PubMed® Baseline Repository*. Available: http://mbr.nlm.nih.gov/

[9] Wikipedia, "Google Translate," ed: http://en.wikipedia.org/wiki/Google_Translate.

[10] T. Sakai, *et al.*, "Overview of NTCIR-8 ACLIA IR4QA," in *Proceedings of NTCIR-8*, to appear, 2010.

Table 7: EN-CS Runs Description

RUNID	Index Units	Translation	Query Units	IR Model
QUTIS-EN-CT-01-T	unigram	1.Web search + 2. [google translate, if 1 fails]	unigram	BM25
QUTIS-EN-CT-02-T	unigram + word	1.web search + 2. [google translate, if 1 fails]	unigram + word	BM25
QUTIS-EN-CT-03-T	unigram	web search + google translate	unigram	BM25
QUTIS-EN-CT-04-T	unigram + word	web search + google translate	unigram + word	BM25
QUTIS-EN-CT-05-T	unigram + word	web search (site: tw) + google translate	unigram + word	BM25

Table 8: EN-CT Runs Description

RUNID	Index Units	Translation	Query Units	IR Model
QUTIS-EN-CS-01-T	unigram	1.web search + 2. [google translate, if 1 fails]	unigram	BM25
QUTIS-EN-CS-02-T	unigram + word	1.web search + 2. [google translate, if 1 fails]	unigram + word	BM25
QUTIS-EN-CS-03-T	unigram	web search + google translate	unigram	BM25
QUTIS-EN-CS-04-T	unigram + word	web search + google translate	unigram + word	BM25
QUTIS-EN-CS-05-T	unigram + word	web search (site:cn) + google translate	unigram + word	BM25