DLUT IR4QA system in NTCIR-8

Huang De-gen Li Ze-zhong Yang Tian
Department of Computer Science and Engineering
Dalian University of Technology
Dalian, 116023, China
huangdg@dlut.edu.cn lizezhonglaile@163.com ytx-1987@163.com

ABSTRACT
This paper describes our work on IR4QA system in NTCIR-8 that intends to evaluate which IR techniques are more useful to QA. We examine IR techniques which can find documents that contain answers to the questions. In our System, we exploit different external resource according to the type of question. In particular, we exploit Wikipedia, Google and Baidu Baike for identifying Named Entity translation, and also employ them to expand query for improving the precision of the retrieval. We use passage retrieval to improve average precision. Our experiments show that these techniques above can significantly increase retrieval precision.

Categories and Subject Descriptors
H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval – *Information filtering, Query formulation, Retrieval models, Search process, Selection process.*

General Terms
Design, Experimentation, Languages, Performance.

Keywords
Information retrieval, Question answering, Query expansion

1. INTRODUCTION
Question answering is a hot research field in Natural Language Processing, which includes many kinds of NLP and IR techniques. Usually a QA system is composed of three main parts: Question Analysis, passage/document retrieval that aims to extract documents that may contain answers, and Answer Extraction. QA research attempts to deal with a wide range of question types including: fact, list, definition, how, why etc. IR4QA is a subtask of ACLIA (Advanced Cross-lingual Information Access), which focuses on complex cross-lingual questions answering (CCLQA) problems [1][2]. In NTCIR-8, our team takes part in the IR4QA subtask, the goal is to evaluate the factors which can affect the performance of information retrieval. The goal of IR4QA is to extract documents that contain answers to the question.

There are many important factors that affect the performance of information retrieval, such as the keywords extracted for query and retrieve related documents according to the query. These two factors may lead to information lost and information overload. As a result, the retrieval system may get low rate of recall and precision. In this paper, we have experiment several techniques to help extract candidate documents which may contain answers to the questions. In particular, we use external resources such as Wikipeia, Google and Baidu Baike. And we use Indri as our retrieval tools. In this toolkit, Language models used to estimate beliefs of representation nodes. Documents are ranked through generative or distribution similarity measures. So we can choose the most relative documents to the question. This paper is mainly about the approaches and their effectiveness in IR4QA task.

2. QUESTION ANALYSIS AND KEYWORDS EXTRACTION
A typical question answering system usually takes the Question analysis as the first step of answering. Question analysis determines the type of question and extracts some other useful information for the future answering. People have done much works about this, mainly including statistical-based approaches and rule-based approaches, or both of them. We just adopt a simple rule-based method for the English question analysis. We make about 100 templates to match questions.

When the process of question analysis is undergoing, the most important informational keywords will be extracted. In other words, view the above two sub-modules as an integrity. Because the template is constituted with two kinds of elements, the first is called Constant Elements, and the other is Variable Elements, which is also called Slot directly. The Constant Elements is helpful for the classification of questions, and the words in the place of Slots often considered as keywords in the future information retrieval for the question answering.

Table 1 . Some manual templates

Question type	Template
Biology	Who is [person name]?
Relation	(What is | What's) the relation between [x] and [y]?
Definition	What is the [definition]?
List	List events related to [*].

The follow is some manually crafted templates we used in the IR4QA task, the square bracket in the patterns represents the Slot, and the content in the round bracket is regular expression. Some are lexical templates, and some patterns utilize semantic

knowledge, which is under the help of Stanford NER Parser. The templates not only match the lexical word, but also the Named Entities, such as person name, location name, organization name and date. Whether the proper nouns are extracted and translated properly, largely determines the result of the succeeding information retrieval and the final answer.

3. QUERY TRANSLATION

In the previous section, we have get the extracted English keywords as the original query, which should be translated to Chinese keywords. Then have a monolingual information retrieval, which is a typical approach of CLIR.

For the general terms, we just translate it with Google Translator. For the proper nouns, we will turn to the multilingual knowledge resource on the internet.

Wikipedia is a multilingual encyclopedia on the web and is composed and edited by volunteers all over the world (http://www.wikipedia.org). It is now the largest, most visited encyclopedia in existence. 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 [4]. Therefore, Wikipedia can be seen as a live dictionary with all kinds of languages.

Baidu Baike (http://baike.baidu.com) is also a multilingual encyclopedia as Wikipedia, which supplements the entries of that's not found in Wikipedia.

Figure 1 . Literal errors happened in questions

In the IR4QA evaluation, we find some literal errors in the original question. For example, the question of TOPIC ID="ACLIA2-CS-0035", Please list the events related to the movie "Initials D". The right spell of "Initials D" should be "Initial D", we don't modify that directly, but utilize a Levenshtein Distance to find out the most possible Wikipedia entry. In the figure 1, we can select the first entry as our destination easily. The same situation occurs in ACLIA2-CS-0009, where the "Olympics Game" should be "Olympics Games", the ACLIA2-CS-0088 have spell mistakes too.

4. QUERY EXPANSION

The query expansion in CLIR not only can expand related words as in monolingual but also can enhance query translation. Many query expansion techniques have been explored to combat errors induced by the query translation. According to the stage of the query expansion happened, roughly divided into two kinds of approaches, the pre-translation expansion and the post-translation expansion, which can be distinguished intuitively. But which plays a more important role in the information retrieval depends on the quality of query translation to a large extent [5]. The pre-translation expansion results in performance better in the case of low quality of query translation. For the question answering system, the pre-translation expansion can strengthen the effect of query translation effectively according to our experiments.

For the pre-translation expansion, we designed a simple but effective method. We submit the English query words to Google Search, the returned snippets are seen as relevant documents. We have an approximate Pseudo Relevance Feedback method, but our objective that expansion isn't English unigram word; we select the bigrams as our destination. That's because normal unigram words, will continue to lead to translation ambiguities. A pre-translation expansion with much ambiguities or noises won't be of help to the final retrieval result. While a bigram can be viewed as a unigram together with its brief context, thus have few translation ambiguities. For example, the question "What is the relationship between Garfield and Fox?", we can get the bigram expansion "Century Fox", from which the translator can identify the "Fox" is not an animal. In NTCIR-7, a question "What's the relationship between Jordan and basketball", the translation "Jordan" may be a big ambiguity for the computer, but we can get "Michael Jordan" to replace "Jordan" though pre-translation expansion, thus, the expanded bigrams can also revise the original query, which is often in an abbreviated form.

We utilize the result of question analysis, which classify the questions into nine kinds, including Biography, Place, Date, Event and Why questions etc. We notice that for different kinds of questions, we should adopt different expansion approaches, and utilize different resources. Not more kinds of approaches and resources to utilize means better performance, improper query expansion often bring more noises and aggravate the result of IR.

For the Biography question, we just have a post-translation query expansion. The source we utilized is Wikipedia, if not succeed, then turn to Baidu Baike, if failed continuously, we will have to submit them to Google Search. If we find the entries in the Wikipedia or Baidu Baike, we just get the first paragraph, and use some templates such as "出生(*)", "称为(*)", "籍贯(*)" to extract relevant query words. Else, we just use a modified tf-idf method to get the relevant words.

The Definition question and the Location question are processed in the same way with Biography question. Of course, they have different templates to extract related words. There are many similarities among the three kinds of questions, the query words are some Named Entities or single words, which called simple questions.

For the Why question, List question and other question, we have a pre-translation and post-translation expansion together. Because these questions is different from the above three kinds of question, whose keywords can be more than one word, and can not get a

much satisfied quality of translation, so pre-translation expansion is integrated to cope with the limitations.

For the Relationship question, it often has a form of "What is the relationship between X and Y ? ". When we determine the candidate extensions, we adopt a modified tf-idf method. In the method, the tf is replaced by a product between tf_x and tf_y .The tf_x represent the term frequency near of X, and tf_y means near of Y.

5. DOCUMNET RETRIEVAL

5.1 Index

Documents should be indexed first to have document retrieval. We use Indri to create an index, and the index unit is words segmented by Chinese Word Segment system of DLUT which is based on CRF [6], and makes high performance in segment words.

5.2 Structural Query Language

Some details also should be noticed. For example, when we retrieval a western person, his Chinese name should be divided into several parts by the dot in the name. For the name "本．拉登" ("bin Laden"), the "拉登" ("Laden") often appears in Chinese news separately. So we should give it a appropriate query weight. The place name such as "台北市" ("Taipei City") often appears in the form of "台北" ("Taipei") without of place suffix, so a #syn(台北市, 台北) should be added.

5.3 Passage Retrieval

In our system, we use passage as retrieval unit, since document is not a proper granularity for information retrieval oriented to question answering. We separate the document into several passages. Each passage retrieved with score assigned by Indri. Then we assigned each document with its best passage score. Documents will be sorted according to their scores, the more high score means more relevant document.

5.4 Hybrid Method

Firstly, we have text retrieval as the above procedures. At last, we have a hybrid retrieval. We first retrieval as normal with web expansion, in the second step, we have a retrieval without expansion at all. Then we combined the two results. The steps mentioned above take the first third of the previous results into the final results, Then, alternately fetch the results into the final results, every time 5 for latter but 1 for previous. Experiments show a better performance of this hybrid approach, a reasonable explanation is that, the earlier part of result that with expansions is retrieved by both queries and expansions, so they are the most related documents. The latter part may be caused either by queries or expansions, so may even less related as the earlier part of result that without expansions at all.

6. EXPERIMENTS AND RESULTS

The experiments are run on Xinhua News of People's Republic of China, which is simplified Chinese style. We have submitted 2 results (the second and the third is same), DLUT-EN-CS-01-T has a normal information retrieval and the other in a hybrid way.

Table 2. Evaluation Result

	Mean AP	Mean Q	Mean nDCG
DLUT-EN-CS-01-T	0.3817	0.4052	0.6394
DLUT-EN-CS-02-T	0.3882	0.4114	0.6461

7. REFERENCES

[1] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Ji, D., Chen, K.-H., Nyberg, E. Overview of the NTCIR-7 ACLIA IR4QA Task, *Proceedings of NTCIR-7*, to appear, 2008.

[2] Sakai, T., Kando, Hideki Shima, Noriko Kando. Overview of the NTCIR-8 ACLIA IR4QA Task, Proceedings of NTCIR-8, to appear, 2010..

[3] Lixin Shi, Jian-Yun Nie and Guihong Cao. RALI Experiments in IR4QA at NTCIR-7 . NTCIR-7, pp.115-124, 2008.

[4] Chih-Chuan Hsu, Yu-Te Li, You-Wei Chen and Shih-Hung Wu. Query Expansion via Link Analysis of Wikipedia for CLIR. NTCIR-7, pp.125-131, 2008.

[5] Paul McNamee, mes Mayfield. Comparing cross-language query expansion techniques by degrading translation resources, pp.159-166,2002

[6] Luo Yan-yan, Huang De-gen. Chinese Word Segmentation Based on the Marginal Probabilities Generated by CRFs. Journal of Chinese Information Processing. 2009.