KECIR Information Retrieval System for NTCIR7 IR4QA Task

Dongfeng CAI, Dongyuan LI, Yu BAI, Bo ZHOU
Knowledge Engineering Research Center,
Shenyang Institute of Aeronautical Engineering
nlpxiaobai@yahoo.com

Abstract

The KECIR group participated in the Information Retrieval for Question Answering task of the NTCIR-7 ACLIA (Advanced Cross-lingual Information Access) Task Cluster. In this paper, we describe our approach on the Simplified Chinese (CS) document collections of ACLIA for the IR4QA task. This year we compared three different approaches of standard query expansion, including Local Context Analysis (LCA) method (KECIR-CS-CS-01-T), Relevance feedback method (KECIR-CS-CS-02-DN) and using online encyclopedia method (KECIR-CS-CS-03-DN). Some famous information retrieval models, i.e., Vector Space Model, Language Model were also adopted in our study for ranking relevant documents.

Keywords: *Information Retrieval, Question Answering, Query Expansion.*

1. Introduction

Information retrieval (IR) aims at finding as many relevant documents as possible. Hence, document retrieval is the essential part of information access. It is for the first time that traditional isolated IR tasks be changed into IR for QA tasks in advanced cross-lingual information access (ACLIA) of NTCIR-7, in which IR is evaluated as a component of complex cross-lingual question answering (CCLQA)[1].

IR for QA is possibly related to all traditional IR tasks but there exists something different between them. In ad-hoc IR task, available information is poor so some ambiguities information may exists and user's intention may be not clear. Instead, IR for QA task can use not only the QA question, but also the output of question analysis, such as the result of question classification (question types), the query terms, the query focus and so on. At NTCIR-7, the QA task is moving beyond factoid CLQA and it is novel in the evaluating cross-lingual QA on complex questions for the first time. Theses changes bring out new challenges: more rigid construction of query needs deeper understanding of question, different question type may leads to different set of relevant documents and so on. For example, in IR4QA task, for biographic questions, each related document is described about the person in query, therefore, the more biographic documents exist, the more retrieval results can be got.

The KECIR group participated in the IR4QA (Information Retrieval for Question Answering) task of the NTCIR-7 ACLIA (Advanced Cross-lingual Information Access) Task Cluster. In this paper, we describe our approach on the Simplified Chinese (CS) document collections of ACLIA for the task. We examined the effect of different query generation mechanism on retrieval performance: This year we compare three different approaches to expand query terms, by using Local Context Analysis[2] method (KECIR-CS-CS-01-T); Relevance Feedback[3] method (KECIR-CS-CS-02-DN) and online encyclopedia-based method (KECIR-CS-CS-03-DN). Some classical information retrieval models, i.e., Vector Space Model, Language Model were adopted in our study for ranking relevant documents.

The remainder of this paper is organized as follows. Section 2 describes the architecture of our IR4QA system. Section 3 describes our indexing method, the Chinese Word segmentation and Named Entity Recognition approaches are also presented in this section. Section 4 describes query processing which includes original query generation, query expansion and the query length identification experiments. Section 5 describes similarity calculation in detail. Section 6 presents the experiment results of our IR4QA system at NTCIR7. In Section 7 we draw the conclusion and future works.

2. System Description

The system is composed of indexing, query processing and similarity calculation modules. The indexing module implements the pre-process of corpus and indexer. Our system is implemented based on Lucene[4] for indexing system. The query processing module aims at generating new query through query expansion techniques such as the pseudo relevance feedback, local context analysis and online encyclopedia-based method. In our system, we also adjusting the number of query terms in according with question type. For ranking retrieval results, the hybrid similarity calculation module was adopted. And the top N retrieved documents were the final results.

3. Indexing

Our retrieval system is indexed by words. Different from many western languages, there is no explicit boundary between Chinese words, therefore, the primary problem we need to solve is the word segmentation. As many other QA systems, Named Entity recognition is employed here too.

3.1 Word segmentation and Named Entity Recognition

The past evaluation has shown that retrieval performance is raised by the recognition rate of named entity. We employed a Conditional Random Field (CRF) based segmentation tools which were also used to recognized name entities like *Person, Location* and *Organization*.

3.2 Document index

Lucene[4] as an open source index toolkit, it organizes the documents by means of inverted index. In our system, the titles and contents of documents are indexed based on word respectively by using Lucene. The out-of-vocabulary (OOV) words were indexed by character. But, some invalidation document, which contains stop words like "请注意", "国际要闻目录", "发稿目录" in its title, are not indexed.

4. Query processing

Query processing is to generate new query according to the question in topics and the result of question analysis. This includes following steps: original query generating, query extension, weight reallocating for each query term and amount adjustment of query terms.

4.1 Original query generation

Each topic is composed of two parts: question (T) and narrative (D) as illustrated in Figure 1. The workshop permits participant submitting run using one or more parts.

<TOPIC ID="ACLIA1-CS-T365">
<QUESTION LANG="CS"> 德国电信和意大利电信是什么关系？ </QUESTION>
<NARRATIVE LANG="CS"> 我想知道德国电信和意大利电信是什么关系。</NARRATIVE>
</TOPIC>

Figure 1. A sample of topic for IR4QA task

The original query is generated by three steps. The first step is word segment and entity recognition as Section 3.1 description; the second step is crucial part extraction by question templates and stopword filter, if the word in crucial part is divided in segment, it will be append to the query as a phrase; the last step is word weight calculation.

4.2 Query expansion

Three methods are used in our experiment, namely, relevance feedback, local context analysis and web expansion.

4.2.1. Relevance feedback. Relevance feedback is a widely explored technique for query expansion. It is often done using a specific measure to select terms using limited set of ranked documents of size m. One simplistic approach could use bind relevance feedback to determine candidate extension. Term frequencies can be measured given the top n documents retrieved using original query Q. After stopword and question word removal, frequent terms are appended to Q, which is then re-evaluated. The common method is Rocchio's relevance feedback, run-02 implements query extension by standard Rocchio formula.

$$Q' = \alpha Q + \beta \sum_{r=1}^{n_{rel}} \frac{D_r}{n_{rel}} - r \sum_{n=1}^{n_{norel}} \frac{D_n}{n_{norel}}$$

In this formula, α, β, γ is constant，D_r is a vector of relevant document d_r， D_n is a vector of non-relevant document d_n， n_{rel} is relevant document number，n_{norel} is non-relevant document number.

4.2.2. Local context analysis. The local context analysis (LCA) is an effective approach to expand original query terms. This approach chooses those concepts which co-occur with original query terms from top n ranked documents. In TREC, LCA has the best performance among all of traditional query expansion approach, and the approach is simple. One submission (run01) uses LCA to expand question words. The steps are described in Figure 2.

1. Retrieve top n ranked passage. A passage is a text window of fixed size (300 bytes in our experiment).
2. Concepts in the top n passage are ranked according to the formula.

$$Sim(q,c) = \prod_{k_i \in q} (\delta + \frac{\log(f(c,k_i) \times idf_c)}{\log n})^{idf_t}$$

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j}$$

$$idf_t = \max(1, \frac{\log_{10} N / np_t}{5})$$

$$idf_c = \max(1, \frac{\log_{10} N / np_c}{5})$$

where n is the number of passage, $pf_{i,j}$ is the frequency of term k_i in document i, $pf_{c,j}$ is the frequency of concept C in document j, N is the total number of passage collection, np_t is the total number of passage contained term k_t, np_c is the total number of passage contained concept C, δ =2 in our experiment.

3. Add m top ranked concepts to Q.

Figure 2. Local context analysis

The term's weight is also recalculated in new query (after expanded). For an original term, its weight will be set 2. For an new term, the weight will be $1 - 0.9 \times i / m$,

i is the position in ranking, and m is the number of expanding query terms in new query.

4.2.3. Web expansion. In our query expansion experiment, we also use the online encyclopedia as the web expansion resource.

Online encyclopedia is suitable for the keywords expansion of Biography and Definition questions. Name Entities in the passage are important terms, in addition, we extract nouns and verbs which are nearest to the original keywords(less than 5 words) as our expanding words.

For instance, the query term in "谁是本拉登?"("Who is Osama bin Laden?") is "本拉登"("Osama bin Laden"), by using online encyclopedia, we can get "本-拉登", "本・拉丹", "拉登", "宾・拉登" as the synonyms of the query term.

4.3. Query length optimization

In query expansion experiments, an interesting phenomenon is observed. For biographic question and definition question, short length query always performs well, however, long query has better performance in dealing with event question and relationship question. For example, a biographic query "农德孟" can be expanded to "农德孟 越南 越共 总书记", or "农德孟 越南 越共 总书记 中央 国会 中国". But the later generated query has more no-relevant terms. However, for a relationship query "德国电信 意大利电信", it expansion terms are "合并 政府" or "合并 政府 计划 欧洲 电话". Here the later query with five expansion terms could get more relevant documents.

It is necessary to adjust query length by question type. The number of terms is to be adjusted by experiments in training collection. In experiments, the best query length is listed in Table 1. There are four question types in our QA system, and query length is the quantity of original terms and expansion terms.

Table 1. The best query length

Question type	Query length
biography	+ 3
definition	+ 3
relationship	+ 4
event	+ 6

In definition questions, the query which is appended to fifteen terms has a little improvement than the one with three expansion terms, but time costs much, so we remain three expansion terms in query. And the expansion length of event question is six in our system. Section 6.1 formulates the experiment in detail.

5. Similarity Calculation

5.1 Vector Space Model

Vector Space Model (VSM) is a popular model in recent years in Information Retrieval field. It is simple and effective. The open source tool Lucene mainly ranks documents by VSM. The approach of similarity calculation is as follows:

$$sim(d_j, q) = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

where $w_{i,j}$ is term weight in document d_j, which is computed by $tf * idf$; $w_{i,q}$ is term weight in query q, which is a weight after query expansion.

5.2 Language Model

Language model has been applied to the problem of ad-hoc retrieval successfully. Query-likelihood is used in our experiment. It estimates a language model for each document, and then ranks documents by the probability of generating query in this language model.

There are two kinds of query-likelihood model, Multivariate Bernoulli model[5] and Multinomial model[6]. Multinomial model treat the query Q as a sequence of independent words, rather than a binary vector, leading to a multinomial view of model D:

$$P(Q \mid D) = \prod_{w} p(w \mid D)^{c(w,Q)}$$

Here $c(w,Q)$ is the number of times the word w occurs in the query. The probabilities $p(w \mid D)$ are taken to be smoothed relative frequencies of word form document D itself. Our experiments uses Dirichlet smoothing method[7].

$$p(w \mid D) = \frac{c(w,D) + \mu p(w \mid C)}{\mid D \mid + \mu}$$

where $\mid D \mid$ is document length, $c(w,D)$ is term w frequency in document D, $p(w \mid C)$ is term's probability in document collection, μ is a smoothing parameter.

5.3 Hybrid Model

Hybrid model is the combination of VSM and language model. The idea is as voting as answer extraction in QA system. The document which occurs in two lists (VSM ranked list and Language Model ranked list) is put in the top, that is, the final similarity of document is a linear combination of the vector-space-model document score and the language-model document score.

$$sim(d,q) = (1-\alpha) \cdot sim_{vsm}(d,q) + \alpha \cdot sim_{LM}(d,q)$$

where α is parameter to adjust the model's weight.

6. Experiments and Evaluation

In experiments, the training data is composed of 40 questions derived from the EPAN interface provided by workshop, and the number of every type questions is 10. For each question, we also construct the answer sets manually.

6.1 Query length optimization experiment

If query length is different by question type, every part will achieve its best performance, and then the whole IR system will be the best. Query length is the sum of the original query length and expanding query length. Because original length is certain, query length is changed into the expanding query length. In this experiment, we determine the expanding length.

a) In biographic question, question focus on one person, and the document containing the person is regarded as relevant document. Here, query expansion aims at finding synonym, that is, those terms represent the same person, but name is different. For example, the original query "农德孟" is expanded as "农德孟 越南 越共 总书记", "越南 总书记" or "越共 总书记" is another call of "农德孟". From the result of query length shown in Table 2, the new query added to three terms is good.

Table 2. Expanding length of biographic question

Expanding length	Mean AP	Mean Q	Mean nDCG
0	0.593	0.589	0.687
1	0.621	0.611	0.720
2	0.647	0.635	0.770
3	**0.654**	**0.644**	**0.782**
4	0.645	0.639	0.780
5	0.640	0.638	0.780

b) In definition question, finding synonym is the goal of query expansion. The short query length is helpful for getting better performance, and the good points appear in 3 and 15. Mean Average Precision (MAP) is 0.572 at 3, and MAP is 0.582 at 15. For instance, the query "温室效应", three expanding terms is as follows: "地球 气体 二氧化碳", the query with fifteen expanding term is "地球 气体 二氧化碳 美国 研究 法国 全球 排放 发现 国家 气候 科学家 造成 气温 减少". Considering searching time and almost the same performance, finally, we choose three terms as expanding key. The experiment result is shown in Table 3.

Table 3. Expanding length of definition question

Expanding length	Mean AP	Mean Q	Mean nDCG
0	0.549	0.535	0.684
2	0.571	0.557	0.706
3	**0.572**	**0.559**	**0.709**
4	0.559	0.551	0.701
14	0.571	0.558	0.701
15	0.582	0.568	0.717
16	0.578	0.565	0.710

c) For relationship question, short length of query is not good enough, such as the topic "ACLIA1-CS-T365". When expanding length is two, the query is "德国电信 意大利电信 合并 政府". When expanding length is four, query is "德国电信 意大利电信 合并 政府 计划 欧洲". In this instance, four expanding length is a better strategy. The result is shown in Table 4.

Table 4. Expanding length of relationship question

Expanding length	Mean AP	Mean Q	Mean nDCG
0	0.527	0.516	0.707
2	0.547	0.537	0.734
3	0.562	0.558	0.746
4	**0.572**	**0.569**	**0.760**
5	0.563	0.556	0.750
6	0.550	0.544	0.736

d) For event question, with the increasing of expanding length, the system performance is raised until expanding length is fifteen. However, time costs much more. So we get top six words as expanding terms, and the performance is still acceptable. The result is shown in Table 5.

Table 5. Expanding length of event question

Expanding length	Mean AP	Mean Q	Mean nDCG
0	0.247	0.248	0.427
4	0.310	0.309	0.495
5	0.321	0.319	0.506
6	**0.325**	**0.323**	**0.512**
7	0.328	0.328	0.518
8	0.330	0.330	0.520
14	0.335	0.333	0.527
15	0.336	0.334	0.528
16	0.333	0.332	0.524

6.2 Hybrid model experiment

There is a list with unique model in Table 6. From it, Language model is better than VSM, and hybrid model is better than Language model.

Table 6. Model performance

Model	MAP
VSM	0.392
Language model	0.461
Hybrid model	0.502

In merging step, merging factor α has great effects on the performance. The experiment aims at finding the best α from 0.2 to 0.8. Table 7 shows the change of MAP. The best value is 0.5, its MAP reaches 0.441.

Table 7. Determine merging factor

α	0.2	0.3	0.4	0.5	0.6	0.7	0.8
MAP	.362	.381	.410	**.441**	.430	.421	.418

6.3 Submission and analysis

Three runs were officially submitted to NTCIR-7 for evaluation including one T run (01); two D runs (02 and 03). The setting is shown in Table 8. The results is shown in Table 9, along with the mean AP, mean Q and mean nDCG scores.

Run01 could provide acceptable performance for QA system. Run02 is nearly same to run01. Run03 is not good enough compared with run01, because expansion

words list of run 03 is produced from online encyclopedia, which resembles a dictionary. The dictionary has lots of words that can't match the word of document collection, especially for the relationship question and event question.

Table 8. Submitting run description

ID	RUN	Description
1	01-T	Query expansion uses LCA. Only title used in query. Similarity calculation uses hybrid model
2	02-DN	Query expansion uses Rocchio formula. Title and narrative field is used in query. Similarity calculation uses hybrid model.
3	03-DN	Query expansion uses Baidu online encyclopedia. Title and narrative field is used in query. Similarity calculation uses hybrid model.

Table 9. Official real-qrels evaluation results

RUN	Mean AP	Mean Q	Mean nDCG
KECIR-CS-CS-01	0.5013	0.4842	0.6562
KECIR-CS-CS-02	0.4806	0.4645	0.6306
KECIR-CS-CS-03	0.4429	0.4292	0.6011

Three runs have low performance in Mean nDCG metric, because those runs didn't submit enough relevant documents. Run01 gets top 200 ranked documents as relevance; run02 and run03 only contain 100 documents. In future, we plan to re-evaluate these methods with 1000 ranked documents.

7. Conclusions and future work

It's our first time to participate information retrieval task at NTCIR. Our experiments focus on the query expansion method and similarity calculation to improve the IR system performance and whole QA system performance. The essential goal is finding a suitable retrieval approach for QA. The experiments compare three different query expansion techniques, Relevance Feedback, Local Context Analysis and online encyclopedia with words grained level. For each question type, query length should be different. The experimental results show that relevance feedback and local context analysis method have the nearly same performance, and choosing suitable query length by question type is feasible.

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, In Proceedings of NTCIR-7, 2008.

[2] Jinxi Xu, Bruce W Croft. Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th Annual ACM-SIGIR Conference, pages 4-11, 1996.

[3] Chris Buckley, Gerard Salton, James Allan. The effect of adding relevance information. In Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 292-300, 1994.

[4] Apache Lucene, http://lucene.apache.org/.

[5] J.Ponte and W.B. Croft. A language modeling approach to information retrieval. In proceedings on the 21st annual international ACM SIGIR conference, pages 275-281, 1998.

[6] D.Miller, T. Leek, and R. Schwartz. A hidden markov model information retrieval system. In proceedings on the 22nd annual international ACM SIGIR conference, pages 214-221, 1999.

[7] C.Zhai and Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of SIGIR'01, pages 334-342, Sept 2001.