KECIR：An Information Retrieval System for IR4QA Task

Dongfeng Cai, Shengqiao Kang, Yu Bai, Peiyan Wang
Knowledge Engineering Research Center, Shenyang Institute of Aeronautical Engineering
nlpxiaobai@yahoo.com

Abstract

This paper describes our work on the subtask of simplified Chinese monolingual information retrieval for question answering system at ntcir-8. We use the lemur toolkit to build index in unit of Chinese word. OKAPI BM25 as a retrieval model and a density-proportional based pseudo relevance feedback method were used for query expansion. To rank all documents, Statistical language modeling and Minimal Mean Distance (MMD) calculating method were employed. Evaluation at NTCIR-8 shows that the best T-run from our team in terms of Mean nDCG is 0.5981, 0.3411 in Mean AP and 0.3749 in Mean Q.

Keywords: Information Retrieval, density-proportional, pseudo relevance feedback, Minimal Mean Distance

1. Introduction

Information retrieval (IR) aims at finding as many relevant documents as possible. Hence, document retrieval is the essential part of information access. As in NTCIR-7, the traditional isolated IR tasks were changed into IR for QA tasks in advanced cross-lingual information access (ACLIA), IR4QA at NTCIR-8 evaluates cross-language IR using English topics and targeting documents in Simplified Chinese (CS), Traditional Chinese (CT) or Japanese (JA). The corresponding monolingual IR subtasks are also within the scope [1,2].

In this paper, we describe our work on the subtask of simplified Chinese monolingual information retrieval for question answering system at NTCIR-8. We use the lemur toolkit to build index in unit of Chinese word, and use OKAPI BM25 as a retrieval model. A density-proportional based pseudo relevance feedback method was used for query expansion. To rank all documents, Statistical language modeling and Minimal Mean Distance calculating method were employed. Evaluation at NTCIR-8 shows that the best T-run from our team in terms of Mean nDCG is 0.5981, 0.3411 in Mean AP and 0.3749 in Mean Q.

The rest of the paper is organized as follows: In section 2, we present related studies. Section 3 provides an overview of the system architecture. In section 4 and 5, we introduce the query expansion method and document ranking strategy respectively. Section 6 gives the evaluation results and error analysis. Finally, the conclusion is given in section 7.

2. Related Studies

2.1 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 [3] which were also used to recognized name entities like Person, Location and Organization.

2.2 Document index

Lemur [4] as an open source index toolkit, it organizes the documents by means of inverted index. In our system, the contents of documents and the contents of paragraph are indexed based on word by using Lemur respectively. some invalidation document, which contains stop words like “国际要闻目录”, “发稿目录” in its title, are not indexed.

2.3 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. For CS-CS Subtask, only Chinese QUESTION field is used.

The original query is generated by the Question Analysis model [5], it was developed for CCLQA Task, which is based on the Question pattern library and HowNet.

3. 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 Lemur for indexing system. A density-proportional based pseudo relevance feedback method was used for query expansion. Statistical language modeling and Minimal Mean Distance (MMD) calculating method were employed in second retrieval. And the top 1000 documents were returned as the final results.

The system architecture is shown in figure2.

```
<TOPIC ID="ACLIA2-CS-0001">
    <QUESTION LANG="EN"><![CDATA[Who is the best actor in the 76th Oscar's?]]></QUESTION>
    <QUESTION LANG="CS"><![CDATA[第76界奥斯卡最佳男主角是谁？]]></QUESTION>
    <NARRATIVE LANG="EN"><![CDATA[The user would like to know the name of the best actor in the 76th Oscar's.]]></NARRATIVE>
    <NARRATIVE LANG="CS"><![CDATA[使用者想知道第76届奥斯卡最佳男主角是谁。]]></NARRATIVE>
</TOPIC>
```

Figure1. A sample of topic for IR4QA task

4. Query expansion

In NTCIR-7 IR4QA Task, our Experimental results show the validity of the local context analysis (LCA) to expand original query terms [6]. This approach chooses those concepts which co-occur with original query terms from top n ranked documents. Based on this, we propose a new method namely density-proportional based pseudo relevance feedback method (DP). Details are described following:

step1. Use OKAPI BM25 as retrieval model, select top 100 ranked documents as relevant documents.

step2. Statistic the times of word w_j co-occurs with original query term q_i in a window, which contains 10 words around.

$$C(q_i, w_j) = \sum_{k=1}^{n} C(q_i, w_j)_k$$

Where n is the number of relevant documents, n =100, $C(q_i, w_j)$ is the times of w_j co-occurs with q_i in relevant documents.

step3. The relevant degree of q_i with w_j can be calculated according to the formula as follows:

$$I(q_i, w_j)_{l \in \{1,2,\ldots,L_Q\}} = \frac{\log(C(q_i, w_j)/N_{w_j} + 1)}{Max_{w_1, w_2, w_3 .. w_m \ldots}(\log(C(q_i, w_m)/N_{w_m} + 1)}$$

$$N_{wj} = \sum_{r=1}^{|C|} C(q_i, w_j)_r$$

Where $I(q_i,w_j)_l$ is the relevant degree of q_i with w_i, l is the number of original query term q_i in $Q = \{q_1, q_2 \ldots . q_m\}$, which co-occurs with w_i, L_Q is the length of Q, N_{wj} is the number of w_j in all the documents，$|C|$ is the number of all the documents.

step4. Put words which got the same value of l will be into the same set.

step5. Filtrate stop words and low-weighted words.

step6. Add top m ranked words to Q. m is as setting in Table1.

A sample of density-proportional based pseudo relevance feedback results is shown in Table1.

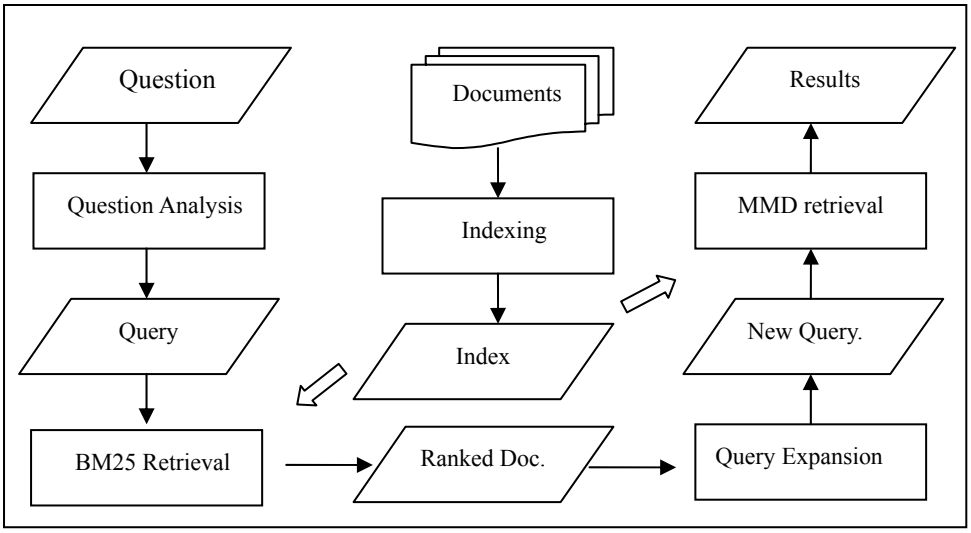

Figure2. The architecture of KECIR IR System

Table1. A Sample of density-proportional based pseudo relevance feedback

Sample word	Expanded word	relevant	co-occurs
北方领土 In Question：“日本的北方领土在哪？”	竹岛	0.8186	2
	四岛	0.3790	2
	外务省	0.3074	2
	南千岛群岛	0.2355	2
	外相	0.2291	2
	固有	0.2264	2
	首相	0.1610	2
	…	…	…

“北方领土” is one of key terms of the location type question “日本的北方领土在哪？”, as in table1, we list some of the expansions term of “北方领土”, the relevant in here represents the relevant degree between expanded word and “北方领土”, co-occurs denotes that how many terms in query key terms co-occur with the expanded one.

The expanding length of a query is dependent on the type of the question. For the complex type like “biography”, “definition”, “relationship” and “event”, we abide by the rules which were used in KECIR IR4QA System at NTCIR7, and for the factoid ones the expanding length will be +5. the rules are shown in table2.

Table 2. The expanding length of a query

Question type	Query length
biography	+3
definition	+3
relationship	+4
event	+6
Why	+5
person	+5
date	+5
Organization	+5
location	+5

The KECIR group participated in the IR4QA (Information Retrieval for Question Answering) task of the NTCIR-7 ACLIA Task Cluster [6]. In this evaluation, we used the same method to adjust query length by type as last, there are five new types of queries in NTCIR-8, we remain five expansion terms in these queries. The query length is listed in Table 1.

By using different retrieval model, the comparison experiments between using expansion and without expansion are shown in table3 and table4.

Table 3.Mean recall of top 100

model	without expansion	with expansion
KL-diverse	0.536631	0.558433
BM25	0.532504	0.554608
VSM	0.528042	0.543024

Table 4.Mean precision of top 100

model	without expansion	with expansion
KL-diverse	0.24842	0.255890
BM25	0.24771	0.255479
VSM	0.23307	0.253014

5. MMD retrieval model

Ponte and Croft used Statistical Language Modeling (SLM) to solve the questions of information retriever firstly and got an effective result [7]. However, SLM is based on Independence assumption; this limits its real application. We consider the relevant degree between query and the document is depending on the distance among query terms in a document. The smaller distance among query terms in a document, the higher value of relevant degree between query and the document. Based on this, we propose a retrieval method namely Minimal Mean Distance (MMD). The steps are described following:

step1. To find out all the documents which contain query terms, get locations and frequency of every query term in these documents.

step2. For a query term q_i , compute its minimal mean distance with other query terms.

$$q_{i\_\min L} = \frac{\sum_{j\neq i}^{n_{term}} \min L(q_i, q_j)}{C_{n_{term}}^2}$$

Where min $L(q_i, q_j)$ is the minimal distance between q_i and q_j, n_{term} is the number of unique query terms in a document.

Step3. Compute document correlativity.

$$P(Q \mid D) = \prod_{q_i=1}^{l_Q} P(q_i \mid D) * e^{f(q_{i\_\min L})-1}$$

Here, l_Q is the number of unique query terms in a query.

$$P(q_i \mid D) = \begin{cases} \dfrac{w(q_i, D)+1}{L_D + |V|} \end{cases}$$

$$f(q_{i\_\min L}) = \begin{cases} \dfrac{1}{q_{i\_\min L}+1} & (q_i \in \mathrm{D}) \\ \dfrac{1}{len_D+1} & (q_i \notin \mathrm{D}) \end{cases}$$

Where len_D is the length of a document, if the document does not contain q_i, then let

$$q_{i\_min\ L} = len_D$$

6. Evaluation

At NTCIR8 IR4QA Task, five runs were submitted. To evaluate system responses, Official evaluation matrix [1] was used.

Table5. The description of submitted runs

Runs	Description
KECIR-CS-CS-01	Query expansion uses density-proportional based pseudo relevance feedback method, MMD as retrieval model.
KECIR-CS-CS-02	Query expansion uses density-proportional based pseudo relevance feedback method, BM25 as retrieval model .
KECIR-CS-CS-03	**MMD as retrieval model.**
KECIR-CS-CS-04	Use paragraph retrieval strategy and the lucene tookits[8], SVM as retrieval model.
KECIR-CS-CS-05	Use the lucene tookits, SVM as retrieval model.

Table6. Evaluate res. Of each runs

Run name	Mean AP	Mean Q	Mean nDCG
KECIR-CS-CS-01	0.3154	0.3546	0.5870
KECIR-CS-CS-02	0.3265	0.3635	0.5941
KECIR-CS-CS-03	**0.3411**	**0.3749**	**0.5981**
KECIR-CS-CS-04	0.2833	0.3193	0.5322
KECIR-CS-CS-05	0.2782	0.3177	0.5586

Run-02 used density-proportional based pseudo relevance feedback method to expand query words, the results shows that in mean effectiveness over 73 topics is better than Run-04 and Run-05, But when density-proportional based pseudo relevance feedback method combined with MMD as used in Run-01, the result is not as good as we expected. While the run-3 which with MMD only gets the best effort in this task.

Table7. Coverage of relevant documents

Runs	Coverage
KECIR-CS-CS-01	4455
KECIR-CS-CS-02	4434
KECIR-CS-CS-03	**4620**
KECIR-CS-CS-05	3653
KECIR-CS-CS-04	3645

Table8. Unique relevant documents

Run name	Unique relevant
KECIR-CS-CS-01	**17**
KECIR-CS-CS-02	**17**
KECIR-CS-CS-04	14
KECIR-CS-CS-05	5
KECIR-CS-CS-03	**0**

7. Conclusion and future work

Our experiments focus on the query expansion method and similarity calculation to improve the IR system performance and for improving the performance of QA system. A density-proportional based pseudo relevance feedback method was used for query expansion. Minimal Mean Distance calculating method was employed in retrieval model. In our experiments, the two methods have got better results than traditional retrieve models separately. But when combined them together, the evaluate result is not as good as we expected. We will conduct the further research to these questions.

8. Reference

[1]. Tetsuya Sakai, etc. Overview of NTCIR-8 ACLIA IR4QA. The proceedings of the 8th NTCIR workshop meeting. 2010.

[2]. Teruko Mitamura, etc. Overview of the NTCIR-8 ACLIA Tasks. The proceedings of the 8th NTCIR workshop meeting. 2010.

[3]. Zhou Bo, Cai Dongfeng. Recognition of Chinese organization name based on conditional random fields. Journal of Shenyang Institute of Aeronautical Engineering. 2009.

[4]. Lemur, http:// www.lemurproject.org/.

[5]. Yu Bai, Li Guo, Lei Liu, etc. KECIR Question Answering System at NTCIR7 CCLQA. The proceedings of the 7th NTCIR workshop meeting. 2008.

[6]. Dongfeng CAI, Dongyuan LI, Yu BAI, Bo ZHOU.KECIR Information Retrieval System for NTCIR-7 IR4QA Task. The proceedings of the 7th NTCIR workshop meeting. 2008.

[7]. J.M.Ponte , W.B.Croft..A language modeling approach to information retrieval1. The 21st Annual Int'l ACM SIGIR Conf.Research and Development in Information Retrieval. Melbourne. 1998

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