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 retrieval model and a density-proportional based pseudo relevance feedback method were used for query expansion. To rank all documents orders, 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 NTCIQ-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].

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 [6] as retrieval model. A density-proportional based pseudo relevance feedback method was used for query expansion. To rank all documents orders, 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 [2] which were also used to recognized name entities like Person, Location and Organization.

2.2 Document index

Lemur [3] 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 [4], it was developed for NTCIR-7 CCLQA, which is based on the Question pattern library and HowNet [8].

```
<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

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 retrieve documents were the final results.

The system architecture is as follows:

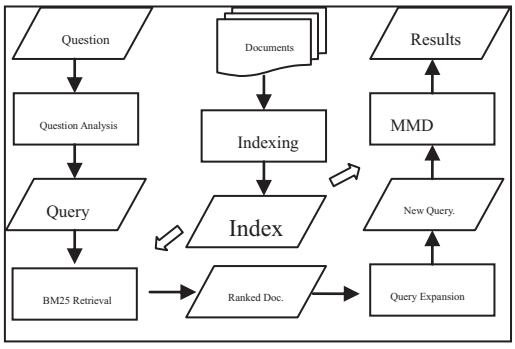

Figure2. System architecture

4. Query expansion

The local context analysis (LCA) [7] 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.

A new method namely density-proportional

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

2. Statistic the time of word w_j co-occurs with original query term q_i in a window, a window contains 10 words.

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

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

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

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

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

Where $I(q_i, w_j)_{l \in \{1,2,.....L_Q\}}$ is the relevant degree of q_i with w_i, l is the number of original query term q_i in $Q = \{q_1, q_2, ...q_m\}$, which co-occurs with w_i, L_Q is the length of Q, N_{w_j} is the number of w_j in all the documents, $|C|$ is the number of all the documents.

4. These words whose values of l are the same will be put into the same aggregate.

5. Filtrate the person names, empty words, stop words and some words divided by mistakes.

6. Add m top ranked words to Q.

Figure3. DP steps

based pseudo relevance feedback method (DP) is used in our two submissions (run01 and run02). The steps are described in Figure 3.

There is partial result of density-proportional based pseudo relevance feedback in Figure4.

何大一（何大一是谁？）	艾伦戴蒙德|0.321928|1 鸡尾酒|0.0924462|1 剪断|0.0671143|1 医学博士|0.0331669|1 发明人|0.0310269|1 国家卫生部|0.0310269|1 美籍|0.0302825|1 美籍华人|0.0297474|1 开创者|0.0280145|1
SARS（什么是SARS病毒？）	药验|0.994158|2 灭活疫苗|0.388507|2 灭活|0.219212|2 克治|0.19152|2 漱口|0.159527|2 冠状病毒|0.151693|2 果子狸|0.138132|2 球蛋白|0.119591|2 溯源|0.11379|2 中间宿主|0.106291|2 接触传染|0.0667222|2 医大|0.0623692|2
刘国梁（刘国梁和中国男子乒乓球队是什么关系？）	男队|0.0346149|4 世乒赛|0.0114719|4 主教练|0.0109093|4 乒坛|0.00997275|4 团体赛|0.0054645|4 出征|0.00362991|4 男单|0.00346223|4 教练|0.0033319|4 乒乓球|0.00315089|4
陈信安（谁是陈信安？）	弹跳|0.0161197|1 对垒|0.00672585|1 国王队|0.00194828|1 决赛圈|0.00177343|1 夏季|0.000993202|1 篮球|0.000966563|1 晋级|0.000581356|1 男篮|0.000516372|1 前锋|0.000372982|1
北方领土（日本的北方领土在哪？）	**竹岛|0.818563|2** 四岛|0.378914|2 外务省|0.307445|2 南千岛群岛|0.235511|2 外相|0.229104|2 固有|0.226444|2 首相|0.160984|2 冲绳|0.11391|2 归还|0.085211|2 东京|0.069988|2 大臣|0.0675209|2 理事长|0.0453959|2 日本政府|0.0440085|2 缔结|0.0416699|2 领土|0.0377489|2

Figure4.Partial result of density-proportional based pseudo relevance feedback

For example, the location question “日本的北方领土在哪？”, its query terms are “日本 北方领土”,in “竹岛|0.818563|2”,“竹岛”is the expansion term of“北方领土”, “0.818563” is the relevant degree of “竹岛”and “北方领土”, “2” denotes that “竹岛”co-occurs with “北方领土” and “日本”,

The KECIR group participated in the IR4QA (Information Retrieval for Question Answering) task of the NTCIR-7 ACLIA Task Cluster[4]. 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.

Table 1.The query length

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

We remain five expansion terms in the query“日本的北方领土在哪？”. Query terms become “日本 北方领土 竹岛 四岛 外务省 南千岛群岛 外相”.

We made some experiments on traditional retrieval models; we selected the top 100 documents as resource of comparison and computed the mean recall and mean precision of them. Table 2 is a list of comparison results.

Table 2.Mean recall

Name of retrieve model	Use initial query terms	Use Query expansion
KL-diverse	0.536631	0.558433
Okapi BM25	0.532504	0.554608
VSM	0.528042	0.543024

Table 3.Mean precision

Name of retrieve model	Use initial query terms	Use Query expansion
KL-diverse	0.24842	0.255890
Okapi BM25	0.24771	0.255479
VSM	0.23307	0.253014

From Table2 and Table3, we can find that query expansion can help to improving the effect of retrieval system.

5. MMD retrieval model

Statistical Language Modeling(SLM) tries to build a mathematic Modeling for nature language processing by using statistics and probability, as to find out the rules of human languages and methods which to solve the specifically questions of nature language processing. In 1998, Ponte and Croft used SLM to solve the questions of information retriever firstly and got a very good result in their paper. [6].

Its basic idea is to estimate a Language Modeling for a document, and computer the probability of the Language Modeling creates a query. Than to rank documents according to their probabilities.[4]. The approach of formulary is as follows:

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

Where $c(q_i, Q)$ is the number of word q_i which appears in query terms. $P(q_i \mid D)$ can be calculated by many smoothing methods.

However, SLM is based on suppose that query terms are all unattached. We think that the distance among query terms in a document is shorter; the value of relevant degree between query and the document is bigger. So we put forward a new retrieval method namely Minimal Mean Distance (MMD). The steps are described in Figure 5.

1 Build index, find out all the documents which contain query terms, and record the locations and time of every query term in these documents.

2 Computer the minimal mean distance of query term q_i among other query terms in a document.

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

Where $\min L(q_i, q_j)$ is the minimal distance q_i between q_j, $q_{i\_\min L}$ is the minimal mean distance of query term q_i among other query terms, n_{term} is the number of query terms' categories in a document.

3 Join MMD with SLM, we can get a new retrieval formulary as follows:

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

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

$$f(q_{i\_\min L}) = \begin{cases} \frac{1}{q_{i\_\min L}+1} & (\text{document D contains } q_i) \\ \frac{1}{len_D+1} & (\text{document D does not contain } q_i) \end{cases}$$

Where len_D is the length of a document, if the document does not contain q_i, $q_{i\_\min L} = len_D \cdot l_Q$ is the number of query terms' categories in the query.

Figure 5 MMD steps

6. Evaluation

6.1 Evaluation results

We all submited five runs to NTCIR-8, the description of five runs in Table4, the results of five runs in Table4.

In addition to measuring the effectiveness of ranked retrieval, organizer also examine the coverage of relevant documents and the number of unique relevant documents for each a run, the results of five runs in Table 6 and Table 7.

We can get conclusions as follows:

(1) In table 5(Mean effectiveness over 73 topics), the results of KECIR-CS-CS-03 and KECIR-CS-CS-02 are better than KECIR-CS-CS-01's, KECIR-CS-CS-01 isn't the best run. The results of KECIR-CS-CS-01, KECIR-CS-CS-02 and KECIR-CS-CS-03 are better than KECIR-CS-CS-04

(2) In table 6(Coverage of relevant documents summed across 73 CS topics), the results of KECIR-CS-CS-01, KECIR-CS-CS-02 and KECIR-CS-CS-03 are better than KECIR-CS-CS-04 and KECIR-CS-CS-05. KECIR-CS-CS-03 is the best

run.

Table 4. Submitting run description

Run name	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 tookit of lucene, SVM as retrieval model.
KECIR-CS-CS-05	Use the tookit of lucene, SVM as retrieval model.

Table 5. Mean effectiveness over 73 topics

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

Table6. Coverage of relevant documents

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

Table 7. 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**

(3) In table 7(Unique relevant documents number), KECIR-CS-CS-01 and KECIR-CS-CS-02 are the best.

6.2 Question analysis

(1) The effect of Chinese word segmentation.

The result of Chinese word segmentation affect the effect of retrieve straightly, for example, in the question“ACLIA2-CS-0006 巴厘岛爆炸和本拉丹的关系？（What is the relationship between Bali bombings and Bin Laden?）”, after the management of Chinese word segmentation, the query terms are “巴厘岛 爆炸 本拉丹”, there are many documents which have not query term“爆炸案”but“爆炸”, if we use the strategy of query term matching, these documents which don’t contain “爆炸案”can’t be retrieved, so the query term“爆炸案” should be divided into“爆炸”and “案”. In the question “ACLIA2-CS-0051 何谓‘丁克族’？(What does "DINK" represent?)”, after management of Chinese word segmentation, “何谓‘丁克族’”became“丁克族”, many documents have words“丁克 家族，丁克一族，丁克 家庭”, so these relevant documents can’t be retrieve by the method of query term matching, “丁克族”should be divided into“丁克”and“族”. Another question “ACLIA2-CS-0090马六甲海峡和日本人之间有何关系？(What is the relationship between the Straits of Malacca and the Japanese people?)”, the query terms are “马六甲海峡 日本人”, but some documents contain “马六甲海峡 日本”, if“日本人”didn’t be divided into “日本” and “ 人”, we can’t retrieve these relevant documents naturally.

(2) The effect of expression of query topic.

As we know, one query topic can be expressed by multifarious manners, at the same time; we can

describe one concept by different words. We can understand the query topic, but computer catches on the meaning of query only by word matching. So the result of retrieval is decided by the query terms of question. For example, the question"阿拉法特何时过世?", query terms are "阿拉法特 过世",
"过世" has the same meaning as "去世" and "辞世". If we don't use some methods to expand "过世" into "去世", "辞世", many relevant documents which only contain "去世" or "辞世" may be ignored. For the question "ACLIA2-CS-0093长江三峡工程的贡献是什么？(What is the contributions of Three Gorges Dam?)", if there wasn't description of question topic, we can't know which benefit of Three Gorges Dam the querist want to know, about environment or economy? In fact, many relevant documents don't contain the query term "贡献", the contributing query term only has"长江三峡工程", if we retrieve documents by "长江三峡工程" and "贡献", computer must consider some documents which contain "长江三峡工程" and "贡献" at one time are relevant, so it will effect the recall and precision of retrieval result.

(3) The effect of query expansion.

Run "KECIR-CS-CS-02" used density-proportional based pseudo relevance feedback method to expand query words, we can find that, in mean effectiveness over 73 topics, its result is better than KECIR-CS-CS-04 and KECIR-CS-CS-05, KECIR-CS-CS-04 and KECIR-CS-CS-05 didn't use any query expansion method. But when density-proportional based pseudo relevance feedback method and MMD as retrieval model were used in KECIR-CS-CS-01, the result isn't the best, it is show that MMD retrieval model depended on query terms seriously, the tiny change of the query terms' number will due to the obvious difference of retrieve result. The precision and the best number of the expansion terms should be studied in future.

(4) The disadvantage of Statistical Language Modeling.

Statistical Language Modeling (SLM) has got very good results in many experiments, it supposes that all the query terms are unattached, the similar degree of the query between a document is determined by the frequency of each query term arises in a document. We made an experiment in question "ACLIA2-CS-0001 第76界奥斯卡最佳男主角是谁？(Who is the best actor in the 76th Oscar?)", Table 8 is some information about the top 5 documents which were retrieved by SLM.

Table8. The frequency of query terms

Sequence number of document	The frequencies of query terms arise in a document			
	76届	奥斯卡	最佳	男主角
XIN_CMN_20040228.0059	3	17	17	8
XIN_CMN_20040127.0190	1	6	21	7
XIN_CMN_20040301.0082	2	7	9	1
XIN_CMN_20040301.0064	1	1	13	4
XIN_CMN_20040301.0023	1	3	2	2

In Table 8, XIN_CMN_20040228.0059 and XIN_CMN_20040301.0023 contain answers, we can find that XIN_CMN_20040127.0190, XIN_CMN_20040301.0082 and XIN_CMN_20040301.0064 don't have the answer of ACLIA2-CS-0001, but the compositors of them are before XIN_CMN_20040301.0023', in fact, the frequency of query terms isn't the only factor to evaluate the probability of a document contains answers. If we can use the relations among query terms, the retrieval results may be better. MMD takes into account the factor of distance among query terms in a document, we think the distance is shorter; the document is more similar to the query.

7. Conclusion and future work

It's our second time to participate information retrieval task at NTCIR. It is obvious that the traditional retrieve models are not fully suitable for the IR4QA test; they are still the most straightforward scheme to deal with the problem.

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. 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 joined them together, we didn't get the expectable results.

In future, we will keep on studying the method of query expansion to solve the problem of choosing suitable query length by type, and how to computer the similarity of query and a document by joining query expansion and MMD method.

8. Reference

1. Tetsuya Sakai, etc. Overview of NTCIR-8 ACLIA IR4QA. 2010.

2 Teruko Mitamura, etc. Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access. 2010.

3. Bo ZHOU, The Recognition Of Chinese Organizations And Locations' names By CRF.

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

5. Lemur, http:// www.lemurproject.org/.

4. Dongfeng CAI, Dongyuan LI, Yu BAI, Bo ZHOU.KECIR Information Retrieval System for NTCIR-7 IR4QA Task. 2008.

5. 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.

6 Karen Spärck Jones, Steve Walker, and Stephen E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2). Information Processing and Management, 36(6):779-840. 2000.

7 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.

8 Hownet, http://www.keenage.com/.