LETTER

Dualized Topic-Preserving Pseudo Relevance Feedback for Question Answering

Kyoung-Soo HAN^{†a)}, *Member*

SUMMARY This study proposes an effective pseudo relevance feedback method for information retrieval in the context of question answering. The method separates two retrieval models to improve the precision of initial search and the recall of feedback search. The topic-preserving query expansion links the two models to prevent the topic shift.

key words: *pseudo relevance feedback, PRF, blind relevance feedback, query expansion, question answering*

1. Introduction

A question answering (QA) system analyzes text collections and answers natural language questions. The QA system usually consists of three phases: question analysis, information retrieval (IR), and answer extraction. In the question analysis phase, the question is analyzed to determine what type of answer is appropriate. The IR step retrieves documents or passages that may contain correct answers. In the answer extraction phase, the answer candidates are extracted from the retrieved results using various natural language processing techniques, and the answer candidates are ranked to select the final answer.

According to the analysis of the effect of each module of the QA system on the system performance, the errors of the answer type recognition module and the IR module accounted for the largest proportion of the QA system errors [1], [2]. Yao et al. [3] showed the performance of their QA system was about 2.6 times lower than that of systems which assumed perfect IR step. In other words, it is necessary to improve the performance of the IR step for a high-performance QA system.

If the IR step fails to retrieve the document containing the answer, it is impossible to extract the answer in any attempt at the answer extraction step. Therefore, high recall is important for IR in the context of QA. In traditional IR, pseudo relevance feedback (PRF), also known as blind relevance feedback, has been successfully applied to improve recall [4]–[6]. However, the PRF has not shown a successful result in the IR for QA. Monz performed the PRF experiment using the Rocchio algorithm, but the performance dropped rather [7]. Pizzato et al. applied the PRF, which extracts only person names from the top ranked documents and

expands the query, to the questions regarding person names, but it did not help to improve the performance [8]. Derczynski et al. tried term frequency based PRF, but it also did not improve performance [9]. However, they found that about 10–34% of the initial search results contain words that could improve retrieval performance, and showed promise for alternative PRF methods.

2. Dualized Topic-Preserving Pseudo Relevance Feedback

2.1 Dualized Feedback

Since PRF uses the top n results of the initial search to expand the query, the quality of the top-level result is very important. Even if the total amount of initial search results is not large, it should contain a lot of relevant documents. That is, the precision of the initial search is relatively important. It is somewhat in conflict with the features mentioned in the previous section, which suggests that the IR should have high recall in the context of QA. It is necessary to increase the precision in the initial search and to increase the recall in the feedback search. In order to reflect this fact effectively, this study proposes a method to perform PRF by separating initial retrieval model from feedback retrieval model.

For example, given the TREC question 1411 “What Spanish explorer discovered the Mississippi River?”, an initial query could be composed of a Boolean query (*spanish* $\wedge$ *explorer* $\wedge$ *discover* $\wedge$ *mississippi* $\wedge$ *river*), and then the initial search could be performed by a Boolean model. If the search result is insufficient, the search can be iteratively performed by removing a query term up to a predetermined level. The feedback query is reconstructed to consist of the terms of the question and the terms selected from the initial search results, and then the feedback search is performed through the feedback retrieval model, such as vector space model, with the feedback query.

2.2 Topic-Preserving Query Expansion

Traditional IR retrieves documents related to the topic of a user query. In the QA context, the topic of the answer document may not be directly related to that of the question. The answers to the question are often included in a very small part of the content of a document that covers a different topic. For example, document NYT19981124.0140 about

Manuscript received January 22, 2017.

Manuscript revised March 3, 2017.

Manuscript publicized March 28, 2017.

[†]The author is with Div. of Computer Engineering, Sungkyul University, Anyang, Korea.

a) E-mail: kshan@sungkyul.ac.kr

DOI: 10.1587/transinf.2017EDL8017

a pecan is considered an answer document for the TREC question 1411 because the document contains the following content:

Pecans are native to the lower Mississippi Valley but not to Georgia. The 16th-century Spanish explorer Hernando De Soto, who discovered the Mississippi River, wrote about these nuts.

The above quote is the only part of the document related to the question. The rest of the document is all about pecan farming and cooking. In this situation, the traditional PRF based on the full document will reduce IR performance by including noise terms in the query. Therefore, effective PRF method for QA should ensure that the feedback text and the terms to be used in the query expansion do not deviate from the topic of the question.

In order to expand the query while keeping its topic, this study add terms that are related to the whole query, not the individual query terms, to the query by taking the idea of local context analysis (LCA) [4]. Sentences are separated from the initially retrieved documents to form passages with l sentences. A passage retrieval is performed by applying a feedback retrieval model to these passages.

For each term t_j in the top n passages, the term is ranked according to the following association score between a question q and the term.

$$assoc(q, t_j) = \prod_{t_i \in q} (0.1 + \frac{\log(f(t_i, t_j))idf_j}{\log(n)})^{idf_i} \quad (1)$$

$$f(t_i, t_j) = \sum_{k=1}^n qf_{ik}freq_{jk} \quad (2)$$

$$qf_{ik} = \min(1, freq_{ik}) \quad (3)$$

$$idf_i = \max(1, \frac{\log_{10}N/N_i}{\epsilon}) \quad (4)$$

$$idf_j = \max(1, \frac{\log_{10}N/N_j}{\epsilon}) \quad (5)$$

where $freq_{ik}$ and $freq_{jk}$ are the number of occurrences of the query term t_i and the term t_j , respectively, in the k -th passage, N is the total number of passages, N_i and N_j are the number of passages containing t_i and t_j respectively, and ϵ is a parameter that determines the valid value of idf .

Among the top ranked terms, m terms that is not included in the query are selected as the expansion term list E . The feedback query consists of the terms from the question and E . The weight of the term t_i in the feedback query is assigned as follows.

$$w_i' = \begin{cases} \alpha w_i & t_i \in q \\ \beta b_i w_i & t_i \in E \end{cases} \quad (6)$$

$$b_i = 1 - 0.9 \times r_i/m \quad (7)$$

where w_i is the query term weight such as $tf \cdot idf$ in the feedback retrieval model, b_i is a boost factor for the query term t_i , r_i is the rank order of t_i in E , and α and β are parameters that control the effect of each component.

3. Experiments

3.1 Setup

This study used AQUAINT corpus [10] which consists of 1,033,461 English news documents drawn from three sources for the period 1998–2000 and the MIT109 test collection [11] which contains 109 questions from TREC 2002 and provides a near-exhaustive judgement of relevant documents for each question. All experiments were performed on 99 questions by removing 10 questions that do not have an answer.

The coverage and answer redundancy were used to evaluate the retrieval results [12]. The evaluation measures at rank n are defined as follows:

$$coverage(n) = \frac{|\{q \in Q | S_{q,n} \cap A_q \neq \phi\}|}{|Q|} \quad (8)$$

$$redundancy(n) = \frac{\sum_{q \in Q} |S_{q,n} \cap A_q|}{|Q|} \quad (9)$$

where Q is the question set, $|Q|$ is the number of questions in Q , A_q is the answer documents which contains answers for a question $q \in Q$, and $S_{q,n}$ is the n top ranked documents retrieved by a retrieval system given the question q . The coverage gives the proportion of the question set for which an answer can be found within the top n documents retrieved for each question. The answer redundancy gives the average number of documents within the top n ranks retrieved which contain a correct answer. The actual redundancy, which is the maximum answer redundancy that any system could achieve, was 18.6 for the 99 questions.

Sentence splitting was performed using Stanford CoreNLP [13], and the experiments were carried out using a modified version of Lucene[†]. All documents and questions have stopwords removed and are stemmed by Porter stemmer. This study used the Lucene implementation for the vector space model (VSM) based on the cosine similarity as a baseline without any change. The Boolean model was modified to rank retrieval results according to idf-weighted overlap between a document d and the question q as follows [14].

$$score(d, q) = \sum_{t \in q \cap d} idf(t) \quad (10)$$

A conjunction of question terms is given as a query for the Boolean model. If the amount of search results is not sufficient, the query is broadened by dropping a term with the smallest amount of information, that is, with the largest document frequency. The query broadening and the retrieval with the broadened query are performed iteratively, and each retrieval result is appended to the previous one. The initial search retrieved up to 1000 documents.

3.2 Results

Table 1 shows the coverage and the redundancy at rank r

[†]<http://lucene.apache.org/>

Table 1 Performance of different document retrieval systems

Limit	System	Coverage		Redundancy	
5	baseline	0.3030		0.6061	
	rocchio	0.2626	(−13.3%)	0.6869	(+13.3%)
	tpqe	0.4545	▲(+50.0%)	1.0303	▲(+70.0%)
	dtpqe	0.4848	▲(+60.0%)	1.1616	▲(+91.7%)
10	baseline	0.3535		0.9899	
	rocchio	0.3434	(−2.9%)	1.2626	▲(+27.5%)
	tpqe	0.5152	▲(+45.7%)	1.5657	▲(+58.2%)
	dtpqe	0.5758	▲(+62.9%)	1.9192	▲(+93.9%)
20	baseline	0.4747		1.7475	
	rocchio	0.3939	(−17.0%)	1.9899	(+13.9%)
	tpqe	0.5758	(+21.3%)	2.4444	▲(+39.9%)
	dtpqe	0.6364	▲(+34.1%)	2.8586	▲(+63.6%)
30	baseline	0.5354		2.2929	
	rocchio	0.4141	(−22.7%)	2.3737	(+3.5%)
	tpqe	0.6061	(+13.2%)	2.9798	▲(+30.0%)
	dtpqe	0.6566	△(+22.6%)	3.5152	▲(+53.3%)
50	baseline	0.5960		3.1010	
	rocchio	0.4444	(−25.4%)	3.1010	
	tpqe	0.6162	(+3.4%)	3.6869	(+18.9%)
	dtpqe	0.6869	(+15.3%)	4.3535	▲(+40.4%)
100	baseline	0.6667		4.2323	
	rocchio	0.5556	(−16.7%)	4.2626	(+0.7%)
	tpqe	0.6465	(−3.0%)	4.6869	(+10.7%)
	dtpqe	0.7677	△(+15.1%)	5.6465	▲(+33.4%)

($r \in \{5, 10, 20, 30, 50, 100\}$) of different document retrieval systems. The rocchio represents the system that applies a document-based PRF to the baseline according to the Rocchio algorithm. The parameters were $\alpha = 1$, $\beta = 2$, and $\gamma = 0$. The tpqe is the system that applies the topic-preserving query expansion to the baseline. The dtpqe is the dualized system that initial documents are iteratively retrieved by the Boolean model and the topic-preserving query expansion based on the initial retrieval results is applied to the baseline. The parameters for both tpqe and dtpqe were $n = 100$, $\epsilon = 2$, $m = 100$, $\alpha = 1$, and $\beta = 2$. The parameters are set empirically. The statistical significance test on the difference between the performance of the retrieval systems was carried out with the randomization test [15]. The performance improvements at a confidence level of 99% are indicated with ▲, and improvements at a confidence level of 95% are indicated with △.

As shown in Table 1, document-based PRF (rocchio) cannot improve the coverage, but can improve the redundancy insignificantly except in one case. The passage-based topic-preserving PRF (tpqe) improves the performance significantly, and the dualized topic-preserving PRF (dtpqe) was the most effective. Overall, the more search results you see, the lower the performance improvement. This is particularly even worse in coverage. The proposed dualized topic-preserving PRF method improves the redundancy statistically significantly even for the 100th search result. The higher redundancy means that the answer extraction module is more likely to extract the answer. On the other hand, the proposed method steadily improves baseline performance in terms of coverage, but it is not statistically significant for more than 30 search results. However, since the answer extraction module in QA usually uses only some top ranked

Table 2 Number of questions that the performance differs from the baseline

Limit	System	Coverage		Redundancy	
		+Q	−Q	+Q	−Q
5	rocchio	3	7	10	9
	tpqe	23	8	32	13
	dtpqe	30	12	39	15
10	rocchio	4	5	16	6
	tpqe	24	8	36	15
	dtpqe	30	8	45	11
20	rocchio	3	11	18	16
	tpqe	23	13	38	19
	dtpqe	26	10	48	16
30	rocchio	3	15	18	16
	tpqe	20	13	39	25
	dtpqe	22	10	49	17
50	rocchio	2	17	13	27
	tpqe	16	14	35	25
	dtpqe	20	11	45	21
100	rocchio	2	13	11	29
	tpqe	13	15	33	32
	dtpqe	19	9	42	24

documents from the search results, the results of this study seem to be meaningful enough.

In order to examine the effect of the proposed method on each question, we analyzed the questions that showed a change in performance compared to the baseline when the proposed method was applied. In Table 2, +Q refers to the number of questions whose performance is higher than the baseline, and −Q refers to the number of questions whose performance is lower than the baseline. Among the 99 questions, the numbers not shown in the table are the number of questions whose performance was the same. There were many cases where document-based PRF could not answer the questions that baseline could answer, but it was extremely rare for the document-based PRF to answer questions that baseline could not answer. The proposed dualized PRF method may be most effective when the answer extraction module uses the top 10–20 documents. Based on the top 10 documents, the dualized PRF had a higher coverage than the baseline for 30.3% of the 99 questions and had a higher redundancy than the baseline for 43.4% of the questions.

3.3 Analysis

In order to investigate the actual operation results of the proposed method, this study examines TREC question 1411 as an example. The TREC answer to the question is “Hernando de Soto”. The baseline ranked the first correct answer document at 263. Some of the top documents, except for duplicate ones, are shown in Table 3. These are documents that cover only some of the words in the question.

When the proposed PRF method is applied to the question 1411, the expanded query terms were as follows.

hernando, soto, de, french, 1541, priest, 16th, gold, 1542, nativ, nut, shell, date, new, cree, di, cathol, down, found, convert

Table 3 Top ranked documents by the baseline for question 1411

Rank	Docno	Part of the document
1	APW19980623.0283	Hijacker of Spanish domestic flight gives himself up, national Spanish radio says.
5	XIE19981003.0029	An estimated 155,000 gallons of oil leaked from an American exploration underwater pipeline in the Gulf of Mexico, the Associated Press reported Friday.
6	APW19981106.0439	Spanish government seeks extradition of Pinochet from Britain.

Table 4 Top ranked documents by the proposed PRF method for question 1411. The answer documents are marked with *.

Rank	Docno	Part of the document
1	APW19990507.0059*	In 1541, Spanish explorer Hernando de Soto reached the Mississippi River.
3	APW19990520.0132	In 1542, Spanish explorer Hernando de Soto died while searching for gold along the Mississippi River.
4	NYT19981124.0140*	The 16th-century Spanish explorer Hernando De Soto, who discovered the Mississippi River, wrote about these nuts.

The term list already contains the answer words.

Table 4 shows the ranked documents, except for duplicate ones, in the search results when the feedback search is performed with the expanded query. Three answer documents were included in the top five documents.

On the other hand, the proposed method did not find any answer document in 50 search results for the TREC question 1417 “Who was the first person to run the mile in less than four minutes?”. The initial Boolean model ranked the first and only answer document to 14th place, but the answer document included in the initial search results was too few to extract a lot of appropriate terms. The terms added to the query were as follows.

two, hour, per, three, year, time, 2, said, 1, ha, 125, long, 5, 10, more, kilomet, 99th, later, millen, go

A number of sports related documents were retrieved by the expanded query. Three answer documents were included between top 50 and top 100 of the retrieval results. If there is no named entity, such as person name or organization name, in the question, the terms tend to be dispersed, and it seems difficult to focus on the topic of the query.

4. Conclusion

This study suggested an effective PRF method for IR in the context of QA. The proposed PRF method divided the initial retrieval model and the feedback retrieval model. The initial retrieval is aimed at a search with an early high precision, while the feedback retrieval aims at a search with a high recall. The topic-preserving query expansion links between the two models. In the previous studies, PRF did not help to improve IR performance in the context of QA, but this study

suggested a way to improve the performance through PRF. It can be an interesting topic in future studies to apply the proposed method to general IR research. The Boolean model that broadens the query by dropping the word with the highest document frequency was used for the initial retrieval, but further research is needed on how to perform accurate initial retrieval for QA more effectively.

References

- [1] D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu, “Performance issues and error analysis in an open-domain question answering system,” *ACM Trans. Information Systems (TOIS)*, vol.21, no.2, pp.133–154, 2003.
- [2] S. Abney, M. Collins, and A. Singhal, “Answer extraction,” *Proc. sixth conference on Applied natural language processing*, pp.296–301, Association for Computational Linguistics, 2000.
- [3] X. Yao, B. Van Durme, and P. Clark, “Automatic coupling of answer extraction and information retrieval,” *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, pp.159–165, 2013.
- [4] J. Xu and W.B. Croft, “Query expansion using local and global document analysis,” *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’96*, New York, NY, USA, pp.4–11, ACM, 1996.
- [5] M. Mitra, A. Singhal, and C. Buckley, “Improving automatic query expansion,” *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’98*, New York, NY, USA, pp.206–214, ACM, 1998.
- [6] S.E. Robertson and S. Walker, “Okapi/keenbow at trec-8,” *Proc. Eighth Text REtrieval Conference (TREC-8)*, pp.151–162, 1999.
- [7] C. Monz, *From document retrieval to question answering*, Ph.D. dissertation, Institute for Logic, Language and Computation, University of Amsterdam, 2003.
- [8] L.A. Pizzato, D. Molla, and C. Paris, “Pseudo relevance feedback using named entities for question answering,” *Proc. Australasian Language Technology Workshop 2006*, Sydney, Australia, pp.83–90, Nov. 2006.
- [9] L. Derczynski, J. Wang, R. Gaizauskas, and M.A. Greenwood, “A data driven approach to query expansion in question answering,” *Coling 2008: Proc. 2nd workshop on Information Retrieval for Question Answering*, Manchester, UK, pp.34–41, Coling 2008 Organizing Committee, Aug. 2008.
- [10] D. Graff, “The acquaint corpus of english news text,” *Linguistic Data Consortium (LDC)*, 2002.
- [11] J. Lin and B. Katz, “Building a reusable test collection for question answering,” *J. American Society for Information Science and Technology*, vol.57, no.7, pp.851–861, 2006.
- [12] I. Roberts and R. Gaizauskas, “Evaluating passage retrieval approaches for question answering,” *European Conference on Information Retrieval*, pp.72–84, Springer, 2004.
- [13] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” *Association for Computational Linguistics (ACL) System Demonstrations*, pp.55–60, 2014.
- [14] H. Saggion, R. Gaizauskas, M. Hepple, I. Roberts, and M.A. Greenwood, “Exploring the performance of boolean retrieval strategies for open domain question answering,” *Proc. IR4QA Workshop at SIGIR*, 2004.
- [15] M.D. Smucker, J. Allan, and B. Carterette, “A comparison of statistical significance tests for information retrieval evaluation,” *Proc. Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM ’07*, New York, NY, USA, pp.623–632, ACM, 2007.