Query expansion method using answer candidates and the effect of combining their results on Web question-answering

Madoka Ishioroshi **Akira Kanai** **Tatsunori Mori**
Graduate School of Environment and Information Sciences
Yokohama National University
79-7 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan
{ishioroshi,a-kanai,mori}@forest.eis.ynu.ac.jp

Abstract

In order to improve the accuracy of Web question-answering (QA) for factoid questions, in this study, we examine the effect of a combination of several QA processes. We introduce a pseudo relevance feedback method in order to obtain different document sets. This feedback method is based on a query expansion technique using answer candidates obtained by the initial QA process. The expanded queries are used in the second QA process for retrieving other document sets. The final answer candidates of QA can be refined by combining the results of the second QA process. The experimental results show that the combination of QA results obtained by using expanded queries improves the accuracy of answer candidates.

1 Introduction

Question-answering (QA) is widely regarded as an advancement in information retrieval (IR) and information extraction (IE). QA systems do not provide us with the relevant documents; instead, they directly provide answers to questions. Many recent studies focus on Web documents because the Web is an up-to-date information source. We term the QA with Web documents *Web QA* in this paper. Since it is not realistic for QA services to develop their own Web search engine, they borrow the existing commercial search engines for Web QA systems.

Note that there are *multiple different* Web search engines available for Web QA. Mori et al. (2007) proposed two methods of combining different Web search engines for factoid QA in rather straightforward ways, and reported that the accuracy of factoid QA was improved by the use of such combinations.

In contrast, there are other methods for increasing the variety of documents. For instance, the pseudo relevance feedback is one of the legitimate IR techniques to bring about such an increase.

In this paper, we will discuss the effect of combinations of several QA processes with the same motivation as Mori et al. (2007). Further, we will introduce a query expansion technique by using answer candidates obtained by using the initial QA process. We expect that the query expansion will improve the recall of the documents that contain not only the keywords in a question but also the answer candidates.

2 Related work

Several researches take advantage of the variety of descriptions obtained by using Web documents. For example, a recent version of START, which is one of the first Web-based QA systems, makes use of multiple information sources (Katz et al. (2004)). Radev et al. (2005) proposed a probabilistic approach to Web QA and used three major search engines for retrieving documents.

Although these researches utilize documents from different information sources, they do not distinguish between information sources after document retrieval. In contrast, our method, described in Section 4, exploits the data redundancy among different information sources.

From the viewpoint of query formulation, the pseudo relevance feedback obtained by query reformulation is one of the legitimate IR techniques (Manning et al. (2008)). In this paper, we propose a novel query expansion technique that adds an answer candidate obtained by the initial QA process to the original query.

3 Basic Web question-answering system

The basic web QA system used in this study is a real-time QA system based on Mori (2005). It can answer Japanese factoid questions. Because

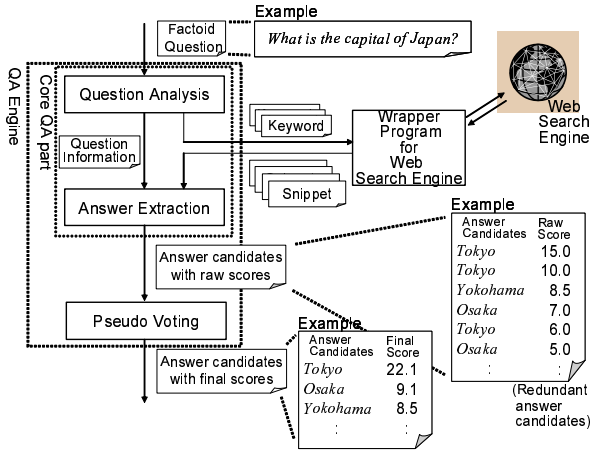

Figure 1: Basic Web QA system

downloading a couple of hundred Web documents is a time-consuming task, we use snippets to generate answer candidates. Snippets are short extractive summaries produced by a Web search engine.

As shown in Figure 1, the system comprises four processes — question analysis, wrapper program for Web search engine, answer extraction, and pseudo voting.

The process of question analysis involves receiving a question from a user and extracting several types of information including a list of keywords and the question type. The process of answer extraction involves receiving a set of snippets from the Web search engine. In this process, each morpheme is treated as an answer candidate and assigned a matching score as described below.

3.1 Raw scores for answer candidates

In the basic web QA system, a composite matching score for an answer candidate. We term this score *raw score* in this paper. As described in Mori (2005), it is a linear combination of several subscores for the answer candidate *AC* in the *i*-th retrieved sentence L_i with respect to a question sentence L_q. In order to reduce the computational cost, the A* search control is introduced in the sentence matching mechanism. With this control, the system can process the most promising candidate first, while delaying the processing of the other candidates, and perform the *n*-best search for the answer candidates.

3.2 Pseudo voting method in search scheme

Many existing QA systems exploit global information on answer candidates. In particular, redundancy is the most basic and important information. For example, there are previous studies that boost the score for answer candidates that occur multiple times in documents (Clarke et al. (2001)). This is known as the *voting method*.

In contrast, we cannot exploit the voting method directly while searching answers because the system quits the searching after *n*-best answers are found. Therefore, an approximation of the voting method, termed *pseudo voting*. In this paper, the pseudo voting score $S^v(AC, L_q)$ for the answer candidate *AC* is defined as follows:

$$S^v(AC, L_q) = (\log_{10}(freq(AC, AnsList)) + 1) \cdot \max_{L_i} S(AC, L_i, L_q) \quad (1)$$

where *AnsList* is the list of answer candidates that have been found in the *n*-best search and $freq(x, L)$ is the frequency of *x* in *L*. In this paper, we call the pseudo voting score *the final score*.

4 Feedback method based on query expansion by using answer candidates

We will introduce a feedback method that is based on a query expansion technique by using answer candidates obtained by the initial QA process. Figure 2 shows the overview of the query expansion method. The method comprises two stages. In the initial stage, the original query formulated from a question is submitted to a Web search engine. Then, the QA process is performed with the retrieved snippets in order to obtain the initial set of answer candidates. In the second stage, each of the initial answer candidates is added to the original query, and each expanded query is submitted to the Web search engine. The newly retrieved snippets are processed by the second QA process in order to search for the final answer candidates. Note that, in our current implementation, we have added only one answer candidate to the original query in each query expansion in order to maintain the number of retrieved snippets. When we increase the number of answer candidates to be added to the original query, the number of retrieved snippets decreases rapidly.

Instead of obtaining one expanded query, a series of expanded queries are generated from the original query and each of the initial top-*n* answer candidates. Each expanded query is separately submitted to a Web search engine in the second QA process in order to obtain a list of answer candidates, and then the lists are merged in the same way as that proposed by Mori et al. (2007). A

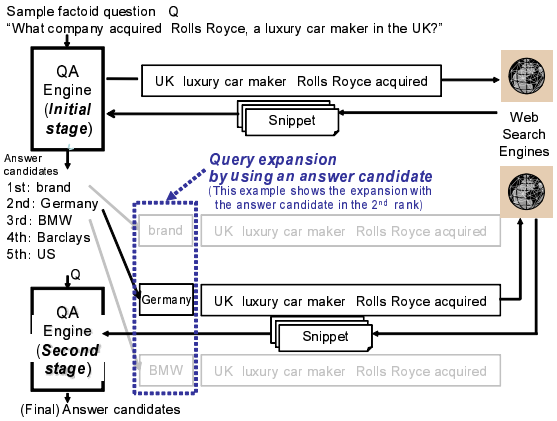

Figure 2: Overview of query expansion by using answer candidates

QA core part receives an expanded query and produces a list of answer candidates with *raw scores*. The *merger* receives the lists of answer candidates and merges the lists into one list. Then, the pseudo voting with Equation (1) is performed on the list in order to generate the list of answer candidates with *final scores*.

Figure 3 shows the outline of the combination method. We term the combination method with one search engine *Combination A*, and term the combination method with two different search engines *Combination B* in this paper.

5 Experimental result

In order to evaluate the effect of the feedback method, we conducted QA experiments as described below. With regard to Web search engines, we used the following Japanese Web search engines: goo (`http://www.goo.ne.jp/`), Yahoo! Japan (`http://www.yahoo.co.jp/`, Yahoo! JAPAN Developer Network). All search results were obtained on December 23, 24, and 25, 2008, with the exception of those for the preliminary experiment.

5.1 Baseline systems

As baseline systems, we prepared the following systems.

g : Baseline 1. The QA system described in Section 3 that utilizes only goo as a Web search engine.

y : Baseline 2. The same system as Baseline 1, but it uses Yahoo.

$g+y$: Baseline 3. The system proposed by Mori et al. Mori et al. (2007) that utilizes both goo and Yahoo as Web search engines. As shown in Figure 4, each list of snippets from an individual search engine is fed to a QA engine separately in order to obtain a list of answer candidates with raw scores, and the lists of answer candidates are then merged. From the merged list, the system generates a list of answer candidates with final scores by pseudo voting (Equation (1)).

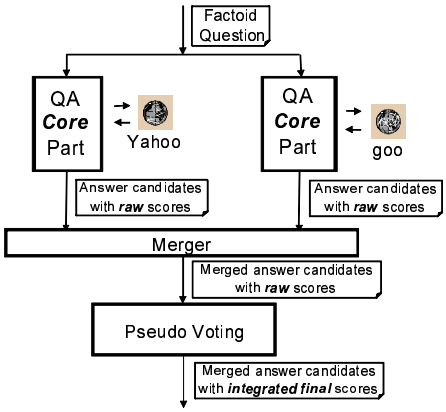

Figure 4: Baseline 3: QA system that combines answer candidates obtained with two search engines.

5.2 Question set and other experimental settings

As for the question set and the answer set, we use a subset of the question set of NTCIR-3 QAC1 (Fukumoto et al. (2002)). The *current* answers to some questions are different from the official answers. In these cases, the authors judged the answer candidates according to the current situation of the world.

In this study, when one API times out, we skip the question. Therefore, we use 166 questions out of the 200 questions in NTCIR-3 QAC1.

With regard to the parameters related to the QA engine, number of answers to be searched is 10 and number of snippets to be retrieved is 100.

Here, it should be noted that the total number of snippets for the systems of the combination methods is larger than that for systems that use only one search engine because the combination methods use several different search results for the same number of snippets to be retrieved. Identifying a fair comparison is a very difficult problem. One possible choice of settings to do so would be that the same total number of snippets is used for each experiment by adjusting number of snippets to be retrieved. However, as shown in Table 1, which is a result of the preliminary experiment about the relation between the number of snippets retrieved and the accuracy, increasing the number of documents does not necessarily improve the accuracy

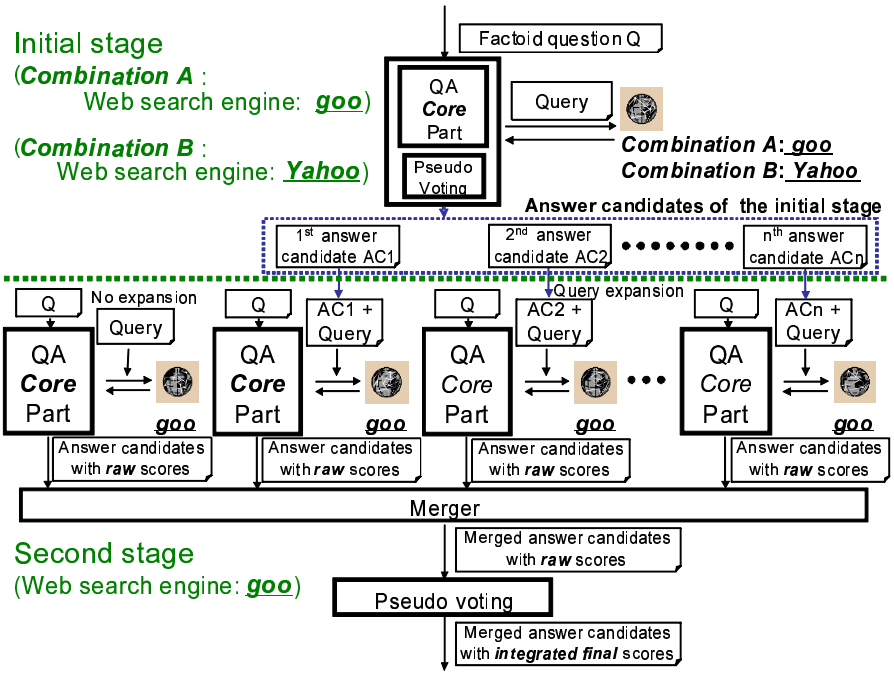

Figure 3: QA system based on a combination of query expansions

of QA. Finally, we straightforwardly adopted the parameters for each QA process.

Table 1: Result of preliminary experiment: relation between the number of snippets retrieved and the accuracy (with goo)

# of snippets	MRR	Total # of correct answers at the n-th rank				
		1st	2nd	3rd	4th	5th
250	0.433	70	23	7	6	6
500	0.406	67	18	9	4	6
750	0.392	59	21	16	10	5

5.3 Results

The accuracy of each combination method was evaluated using the mean reciprocal rank (MRR). Reciprocal rank (RR) is the inverse of the rank of the first correct answer for each question. If no correct answer appears within the top five answer candidates, RR is 0. MRR is the average of RR over all questions.

The evaluation results are summarized in Figure 5. Because of space constraints, we have only shown the results of the system that utilizes goo in both of the initial stage and the second stage, and one that utilizes goo and Yahoo in the initial stage and the second stage, respectively[1]. Note that we employ the following notations:

[1]The system that utilizes Yahoo and goo in the initial stage and the second stage, respectively, was also effective in improving accuracy.

$S_a+S_b+\ldots$ **:** Combined QA system that uses search engines S_i $(i=a,b,\ldots)$ in the same way as Baseline $g+y$ in Section 5.1.

$S_aS_b(n)$ **:** QA system with feedback in which the search engine S_a is used in the initial stage, and the n-th answer candidate of the initial stage is utilized in the query expansion of the second stage with the search engine S_b.

$S_aS_b(1-n)$: Abbreviation of $S_aS_b(1)+\ldots+S_aS_b(n)$.

$g+gg(1-i)$ $(i=2,3,\ldots)$ and $g+yg(1-j)$ $(j=2,3,\ldots)$ correspond to Combinations A and B in Section 4, respectively.

6 Discussion

Figure 5 shows that each $gg(i)$ is not better than Baseline g. However, the combination $g+gg(1-j)$, i.e., Combination A, produces better answer candidates than g. The statistical test shows the following: "$g<g+gg(1-2), g+gg(1-3)$" and "$g \ll g+gg(1-j)$ $(j=4,\ldots,10)$[2]."

In contrast, combinations $g+gg(1-j)$ produce worse answer candidates than Baseline $g+y$. However, there is no statistically significant difference between $g+gg(1-j)$ $(j=5,\ldots,10)$ and $g+y$ at the significance level of 5%. The value of MRR for $g+gg(1-10)$, i.e., 0.503, is comparable to that for $g+y$, i.e., 0.511. This implies that we

[2]The relational symbols "$<$" and "$\ll$" represent that the systems on the righthand side are statistically superior to those of the lefthand side in terms of RR values at the significance levels of 5% and 1%, respectively.

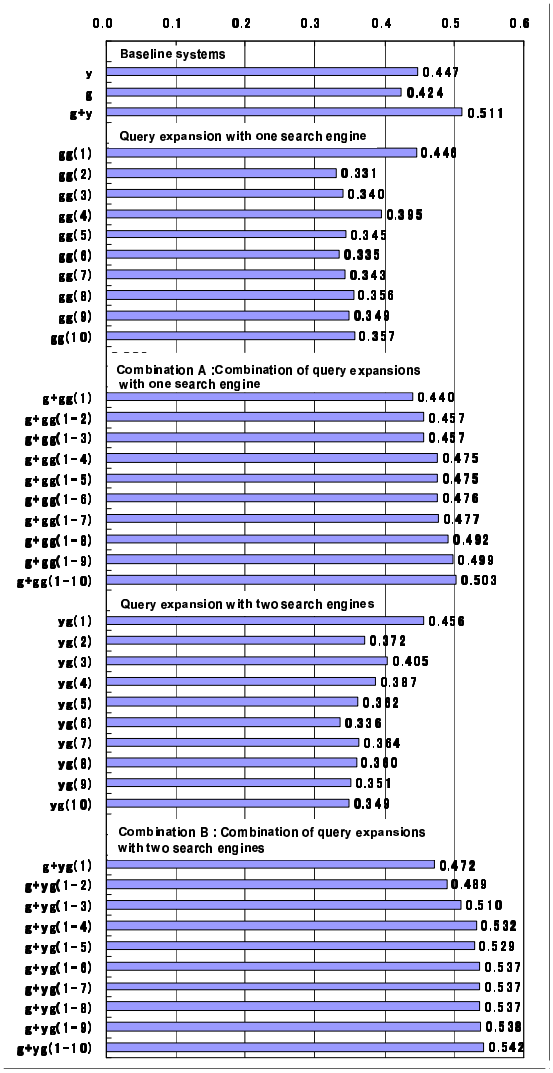

Figure 5: MRR of QA for each setting

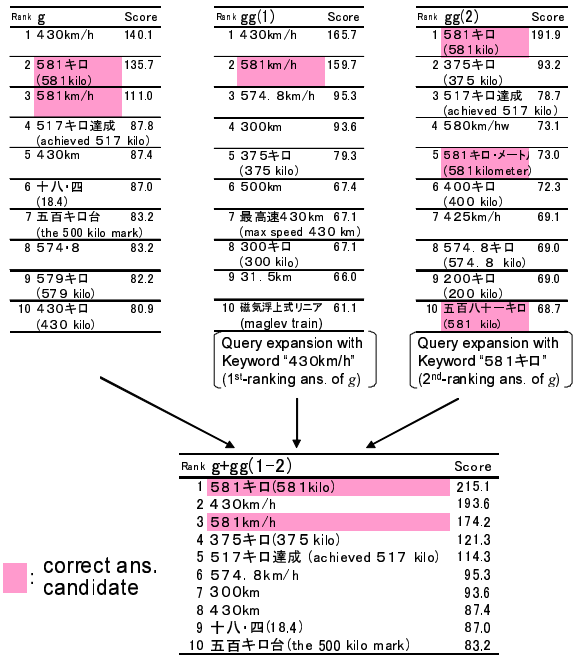

Figure 6: Successful example of the feedback method (Answer candidates for Question QAC1-1126-01, "リニアモーターカーの走行試験で出た最高速はどのくらいでしたか。" ("How much was the maximum speed by the test run of the linear motor car?"))

can achieve the accuracy comparable to that obtained using multiple Web search engines, even if we utilize only one Web search engine.

Each $yg(i)$ $(i = 2,\ldots,10)$ is not better than Baselines g or y. However, the combination $g+yg(1-j)$, i.e., Combination B, produces better answer candidates than Baseline g or y. The statistical test shows the following: "$y < g+yg(1-3)$," "$y \ll g+yg(1-j)$ $(j = 4,\ldots,10)$," "$g < g+yg(1)$," and "$g \ll g+yg(1-j)$ $(j = 2,\ldots,10)$." With regard to the comparison with Baseline $g+y$, $g+yg(1-j)$ $(j = 4,\ldots,10)$ outperforms the baseline in terms of MRR. However, there is no statistically significant difference between $g+yg(1-j)$ $(j = 1,\ldots,10)$ and $g+y$ at the significance level of 5%.

Figure 6 shows a successful example of the feedback method. In this example, the first-ranking answer candidate in the result of g is not correct, although the second and third places are correct. On the other hand, in the case of $gg(2)$, in which the second-ranking answer candidate of g is used for query expansion, the correct answer candidate "581 キロ (581 kilo)" appears in the first place. In addition, the correct answer candidate has a higher raw score than other candidates. As a result, the correct answer candidate is appropriately boosted in Combination $g+gg(1-2)$ by the pseudo voting with Equation (1).

Figure 7 shows an example of failure of the feedback method. In this example, the correct answer candidate "トリノ (Torino)" appears at higher ranks in the results of g, $yg(1)$ and $yg(2)$. However, the incorrect answer candidate "バンクーバー (Vancouver)" has the highest raw score in the result of $yg(2)$. As a result, the incorrect answer candidate is given a higher score than the correct answer candidate in Combination $g+gg(1-2)$.

As shown in the successful example, when the query expansion is performed by using a correct

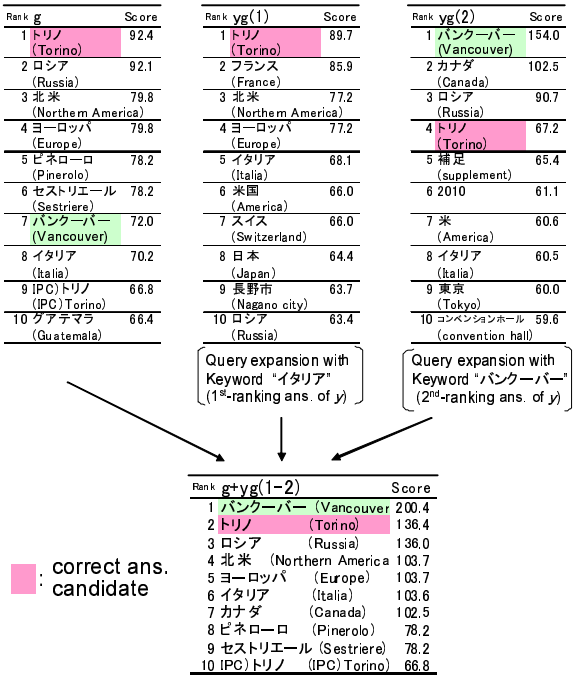

Figure 7: Failed example of the feedback method (Answer candidates for Question QAC1-1200-01, "２００６年の冬季五輪の開催地はどこですか。" ("Where is the venue of the Olympic Winter Games in 2006?"))

answer candidate, the candidate tends to be given much higher raw score than that without the query expansion. On the other hand, when an incorrect answer candidate is used for the query expansion, the gain of the score by the query expansion is smaller than that by the query expansion with a correct answer candidate. Moreover, in this case, new correct answer candidates may be found because snippets of different contexts are obtained from Web search engines.

However, as shown in the failed example, there is a possibility that incorrect answer candidates are wrongly boosted. In order to improve it, we need more sophisticated voting methods than the pseudo voting with Equation (1).

7 Conclusion

We examined the effect of the query expansion technique by using answer candidates obtained by the initial QA process, and their combinations. The experimental result showed that the introduction of the query expansion itself did not improve the accuracy of QA, but QA systems based on a combination of query expansions outperform baselines.

In this paper, we investigated only one feedback method and its combination. The development of more effective feedback and combination methods will be included in our future work.

References

Charles L.A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. 2001. Exploiting redundancy in question answering. In *Proceedings of SIGIR '01: the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 358–365.

Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. 2002. Question Answering Challenge (QAC-1) — Question answering evaluation at NTCIR Workshop 3 —. In *Working Notes of the Third NTCIR Workshop meeting – Part IV: Question Answering Challenge (QAC1)*, pages 1–6.

B. Katz, M. Bilotti, S. Felshin, A. Fernandes, W. Hildebrandt, R. Katzir, J. Lin, D. Loreto, G. Marton, F. Mora, and O. Uzuner. 2004. Answering multiple questions on a topic from heterogeneous resources. In *Proceedings of TREC 2004*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tatsunori Mori. 2005. Japanese question-answering system using A* search and its improvement. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(3):280–304. URL `http://portal.acm.org/TALIP/`.

Tatsunori Mori, Akira Kanai, Madoka Ishioroshi, and Mitsuru Sato. 2007. Effect of combining different Web search engines on Web question-answering. In *Proceedings of the 10th Conference of Pacific Association for Computational Linguistics (PACLING 2007)*, pages 325–332. URL `http://mandrake.csse.unimelb.edu.au/pacling2007/files/final/2%6/26_Paper_meta.pdf`.

Dragomir R. Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal. 2005. Probabilistic question answering on the web. *Journal of the American Society for Information Science and Technology*, 56(3).