Question Answering Challenge for Five ranked answers and List answers – Overview of NTCIR4 QAC2 Subtask 1 and 2 –

Jun'ichi Fukumoto
Ritsumeikan University
fukumoto@media.ritsumei.ac.jp

Tsuneaki Kato †
University of Tokyo†
kato@boz.c.u-tokyo.ac.jp†

Fumito Masui ‡
Mie University ‡
masui@ai.info.mie-u.ac.jp ‡

Abstract

In this paper we describe an evaluation of question answering task, Question Answering Challenge 2 (QAC2). This evaluation project was first carried out at the NTCIR Workshop 3 in October 2002. One objective of the QAC was to develop practical QA systems in a general domain by focusing on research relating to user interaction and information extraction. Our second objective was to develop an evaluation method for the question answering system and information resources for evaluation.

We defined three kinds of tasks in the QAC2 as well as QAC1: Subtask 1, where questions required five possible answers in some order, Subtask 2, where questions had one list of answers and Subtask 3, where there were a series of questions. This paper describes the evaluation overview of Subtask 1 and 2. We prepared 200 questions for Subtask 1 and Subtask 2. We conducted only Formal Run for these two subtasks. There were 18 active participants: 25 system submissions for subtask 1 from 16 participants and 14 system submissions for subtask 2 from 9 participants.

1 Introduction

The Question Answering Challenge (QAC)[1] was carried out as the first evaluation task on question answering of the NTCIR Workshop 3[1][2][3][4]. Question answering in an open domain is a task for obtaining appropriate answers to given domain independent questions written in natural language from a large corpus[5][6][7]. The purpose of the QAC was to develop practical QA systems in an open domain focusing on research of user interaction and information extraction. A further objective was to develop an evaluation method for the question answering system and information resources for evaluation.

[1]QAC home page is located in the site http://www.nlp.cs.ritsumei.ac.jp/qac/ and its mirror site http://www.ai.info.mie-u.ac.jp/qac/ .

In QAC1, we have prepared for three kinds of subtasks: five ordered answers task (Subtask 1), list task (Subtask 2) and context task (Subtask 3). In list task of QAC1, we used the same question set as Subtask 1 and average number of answers was almost one. Therefore, systems which can give only one answer for questions gave better performance in list task. So, in QAC2 we have prepared different question set from the one for Subtask 1. In context task, there was only one follow-up question for main question in QAC1 and most of participants used the same kind of system as the other tasks. In QAC2, there were seven follow-up questions for one main question.

In this paper, we will describe the evaluation overview of Subtask 1 and 2 of QAC2; task description, task participants, evaluation method, the results and so on. The details of Subtask 3 are presented in another version of report.

2 Task Design of QAC2

We will briefly describe the task definition of QAC2. We have prepared three kinds of subtasks as well as QAC1. The first one requires five ordered answers and the highest ranked answer will be scored. The second one is a list task which requires only one set of correct answers. If there are several answers in a document set, a system has to all possible answers as a answer list. If there is no answer, a system has to respond no answer. The last one is a context task based on information access dialogue. There are several questions for one or more topics. The details are presented in Overview of NTCIR4 QAC2 Subtask3.

For target documents, we used four years Japanese newspaper articles spanning a period of two years(1998 and 1999) taken from both the Mainichi Newspaper and Yomiuri Newspaper. As well as the QAC1, questions used for evaluation require short answers which were exact answers consisting of a noun or noun phrase indicating, for example, the name of a person, an organization, various artifacts or facts such as money, size, date etc. These types were basically from the Named Entity (NE) element of MUC[9] and

IREX[8] but were not limited to NE elements.

In order to get an answer, the system was able to use other information sources such as an encyclopedia, thesaurus, corpus of data and so on. However, answer expressions have to exist in newspaper articles and information of document ID is required as support information for each question. It will be justification for answer expressions being correct one on the basis of the contents of the newspaper articles. Even if answer expressions are correct, these answers with an appropriate document ID will be incorrect answers.

In expressions of question sentences, constraints on full sentence expressions are loosed in QAC2. In QAC1, a question sentence has to be complete one, that is, there is no ellipsis on tail expression of a question sentence. In QAC2, there will be ellipsis on tail expression. For example, tail expression of " (who is ...)" will be just " (who)". In this case verbal expression of question sentence will be omitted.

In context task, we gave more follow-up questions for the first question than the case of QAC1 (one follow-up question). Moreover, two types of question series: gathering type and browsing type. The details are presented in another version of QAC2 overview.

The definitions of three subtasks are as follows:

- Subtask 1

  The system extracts five possible exact answers from documents in some order. The inverse number of the order, Reciprocal Rank (RR), is the score of the question. For example, if the second answer is correct, the score will be one half (1/2). The highest score will be the score of the question. Where there are several correct answers, the system will return one of them.

- Subtask 2

  The system extracts only one set of answers from the documents. If all the answers are correct, a full score will be given. If there are several answers, the system has to return all the answers. Where there are incorrect answers, penalty points will be given. The Average F-Measure (AFM) is used for the evaluation of Subtask 2. Subtask 2 uses the different question set as Subtask 1 in QAC-2.

- Subtask 3

  This task is an evaluation of a series of questions or follow-up questions. A question related to a question in Subtask 2 is given. There will be ellipses or pronominalized elements in these follow-up questions.

3 Question development method

For the QA evaluation, it was necessary to prepare a variety of questions which required elements such as a product name, the title of a novel or movie, numeric expressions and so on. In QAC2, we developed 300 questions of various question types that sometimes included paraphrasing for both Subtask 1 and 2. (Subtask 3 will be presented in another version of report.) In QAC1, subtask 1 and 2 use the same question set, but we prepared different question set for each subtask of QAC2. Therefore, most of questions for subtask 2 have multiple answers.

We gave the following instructions to make questions and their answers for question developers.

- Questions are basically made without detailed checking target documents. Question developers firstly make question sentences and then check their answers using document set. It is intended to make question sentence natural and normal one.

- A question which uses inference process to get answers will be excluded. For example, a question "How many days the Tokyo summit held" for a document "Tokyo summit was held from June 2nd to 6th." will be excluded.

- Question sentence should not include ambiguous expressions such as "famous", "pretty", "fine", "expensive" and so on. It is necessary for answer detection to make subjective judgments.

- All possible answers will be extracted from documents but the number of answers will be ten in maximum. However, it is not obligatory.

- Answer expressions are nouns, proper nouns, numeric expressions and time expressions. Answers expressions do not include Japanese particle " (of)", " (and)" and " (,)". But if such expressions are included in formal expressions, it does not the case.

4 Evaluation Method

4.1 Subtask 1

The system extracted five answers from the documents in some order. The inverse number of the order, Reciprocal Rank (RR), was the score of the question. For example, if the second answer was correct, the score was 1/2. The highest score of the five answers was the score of the question. If there were several correct answers to a question, the system might return one of them, not all of them. The Mean Reciprocal Rank (MRR) was used for the evaluation of Subtask 1. If n set of answers were correct, the Mean Reciprocal Rank (MRR) could be calculated as follows:

$$MRR = \frac{\sum_{i=1}^{n} RR_i}{Q} \quad (1)$$

$$RR_i = \frac{1}{Rank} \quad (2)$$

If a system responds no answer for no answer question, the score of such a question will be "1". If a system responds some answer for such no answer question, the score will be zero.

4.2 Subtask 2

The system extracted only one set of answers from documents. If the system's answer was correct, a score was given. If there were several answers, the system had to return all the answers. Average F-Measure (AMF) was used for the evaluation of Subtask 2. The scores were calculated in the following formula, assuming A as the number of correct answers, A_{sys} as the number of answers that the user's system output, and A_{cor} as the number of correct answers that the user's system output. Q and $Rank$ were assumed as being the number of questions and the rank of the answers respectively.

$$Recall = \frac{A_{cor}}{A} \quad (3)$$

$$Precision = \frac{A_{cor}}{A_{sys}} \quad (4)$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (5)$$

When a system gives no answer for no answer question, the score of this question was 1.0 (F-measure). On the other hand, if a system gives some answer for such No Answer questions, the score was zero.

4.3 Subtask 3

This task was an evaluation of a series of questions. The system had to return all the possible answers for a main question and its follow-up question. A score was given only for the follow-up question in the same scoring method as Subtask 2, which is MF.

5 Task Participants

In QAC2, there were seventeen active participants. Task participation of each participant is shown in Table 1. The number of symbol "$\star$" indicates the number of submission for the subtask. For example, two symbols mean two kinds of results were submitted.

6 Runs for Evaluation

6.1 Description of Formal Run

We conducted the QAC Formal Run according to the following schedule and tasks.

- Date of task revealed: Dec. 3, 2003 (Wed.) 10:00 (JST)
- The result submission due: Dec. 10, 2003 (Wed.) 18:00 (JST)

Task participants are required to submit one system result within 48 hours for each subtask after getting QA data. If task participants will submit two systems for one subtask, they are required to submit the first system within 48 hours and the second one within 72 hours. If task participants will submit three systems (in case of Subtask 3), they have to submit the first system within 48 hours, the second one within 72 hours and the last one within 96 hours. The number of questions is 200 for both Subtask 1 and 2. However, there are several inappropriate questions and no answer questions in the prepared questions. Finally, we evaluated participated systems using 195 questions[2] for Subtask 1 and 199 questions[3]. The submitted results were pooled and were to be delivered after evaluation.

7 Results and Discussion

Subtask 1

There were 25 systems from 17 participants in the Subtask 1. The accuracy the participating systems achieved in the mean reciprocal rank (MRR) is depicted in Figure 1.

The most accurate system achieved 0.607 in the MRR, which is almost same score in Subtask 1 of QAC1. This system returned correct answers in the first rank to 51.3% of the questions and in up to the fifth rank to 73.8% of the questions. The average MRR of top three systems is 0.583. Among them, 44.1% of top ranked answers was correct in average and 67.4% of top five answers was correct in average.

In addition to the MRR standard, we tried evaluating the systems using two other types of criteria as well as QAC1. The first was the ratio of a systems correct answers in the first rank (Figure 2). The second was the ratio of systems' correct answers up to the fifth rank. Those two criteria showed very little difference from the evaluation using the MRR. In both cases, there were only two pairs of systems which had adjoined each other in rank in the MRR evaluation and which swapped ranks under the new criteria. This suggests that the MRR is considerably stable in measuring system accuracy for Subtask 1.

Figure 3 shows the histogram of the difficulty of the question set of Subtask 1. The difficulty of each

[2]The in appropriate questions of Subtask 1 are QAC1-10009-01, QAC1-10091-01 and QAC1-10171-01 and no answer questions are QAC1-10198-01 and QAC1-10199-01.

[3]The inappropriate question of Subtask 2 is QAC2-10169-01. for Subtask 2

Table 1. QAC2 participants

participants name	Subtask		
	1	2	3
Communication Research Laboratory	⋆⋆	⋆⋆	⋆⋆⋆
Tsukuba QA Team	⋆⋆	⋆⋆	⋆⋆⋆
NTT DATA Corp.	⋆⋆	⋆⋆	⋆⋆
Ritsumeikan Univ.	⋆⋆	⋆⋆	⋆⋆
Mie Univ.	⋆⋆	⋆	⋆⋆
Oki Electric Ind.	⋆⋆	⋆⋆	⋆
Toyohashi Univ. of Technology	⋆	⋆	⋆
NAIST	⋆	⋆	
Toshiba Corp.	⋆⋆		
Yokohama National University	⋆⋆		
Nagaoka Univ. of Technology	⋆		
Matsushita Electric Ind.	⋆		
Pohang Univ. of Science Technology	⋆		
The University of Tokyo	⋆		
Keio Univ.	⋆		
NTT Communication Science Lab.	⋆		
NYU/CRL	⋆		
IPU		⋆	

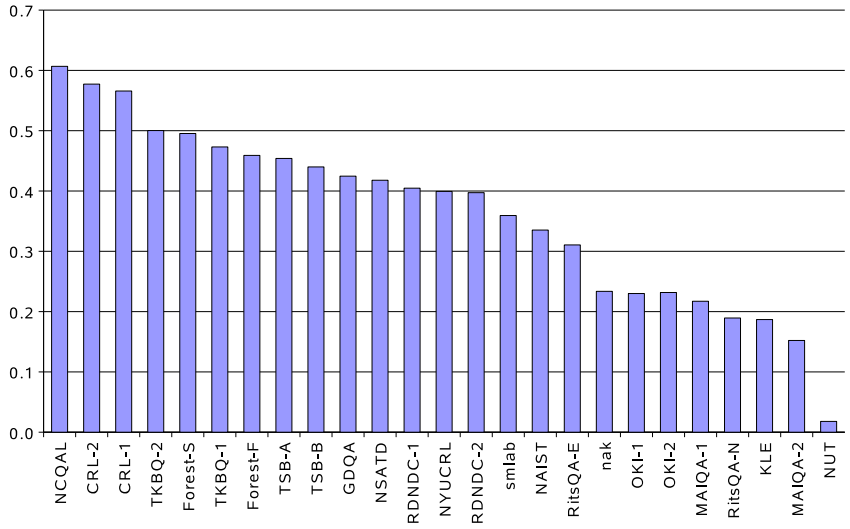

Figure 1. MRR of participant systems in Subtask 1

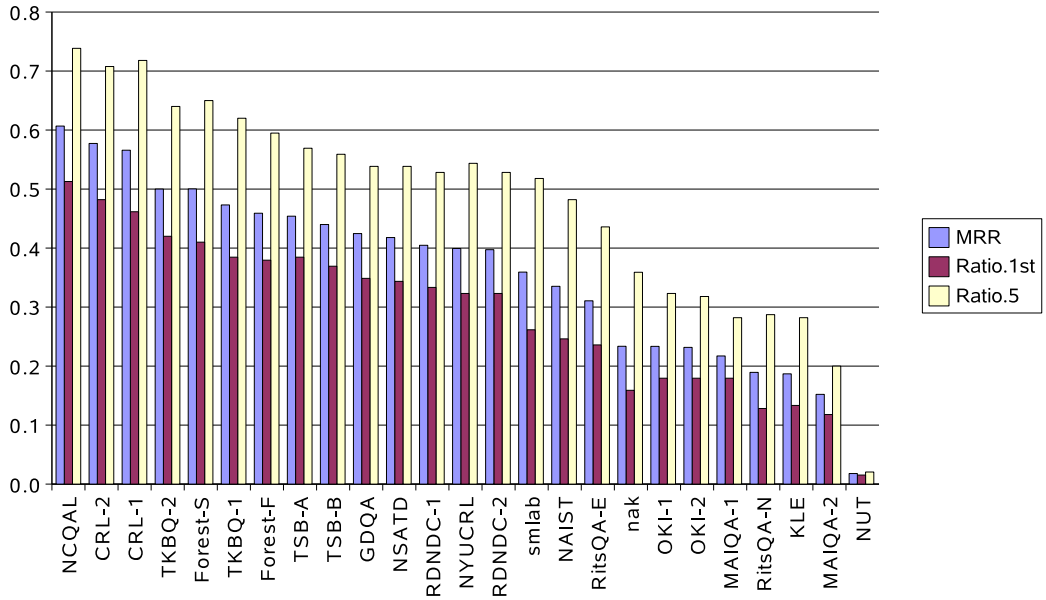

Figure 2. MRR, correct ratio of 1st ranked answer and among 5th ranked ones

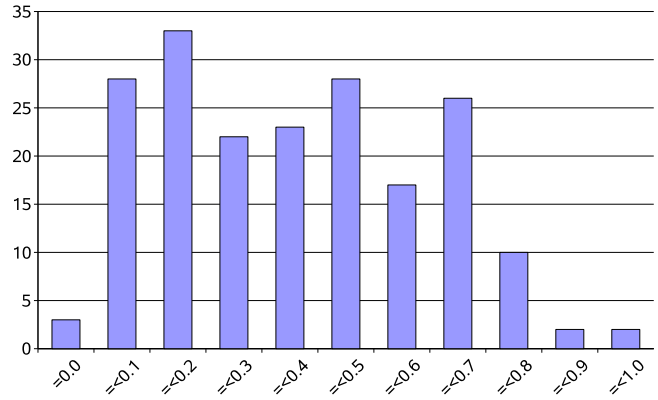

Figure 3. Difficulty of questions in Subtask 1

question is calculated as the average of the reciprocal ranks all the systems achieved for that question.

There was only one question out of 197 which no system could return correct answers although there was 11 questions out of 195 questions (five questions with no answer were excluded) in QAC1. That is QAC2-10016-01. Therefore, if we could merge all the systems in some way, this system will return correct answers for most of all questions used for QAC2.

The easiest question was QAC2-10127-01 and its MRR was 0.97. 24 systems out of 25 systems returned the correct answer and 23 systems were in the first rank to this question.

The distribution has a smooth curve with one peak. Compared with the results of QAC1, the number of difficult questions decreased. Difficult level of question set was almost same as the previous evaluation, QAC1. Also, MRR of top level systems got better than QAC1. Therefore, performance of QA system has progress in average.

Subtask 2

Fourteen systems from nine organizations participated in Subtask 2. (13 systems from 13 organization in QAC1.) The accuracy the participating systems achieved in the mean F-measure (MF) is depicted in Figure 4.

The most accurate system achieved 0.321 in the MF and returned 117 correct answers. The second and third systems were 0.3179 and 0.3176, respectively, and their performance were almost the same as the top ranked system.

Figure 5 shows MF, Precision and Recall of participant systems.

In the first and fourth systems, their precisions were 0.458 and 0.471, and recalls were 0.291 and 0.252, therefore, the strategy of these systems is precision oriented. On the other hand, the precision and recall of the 7th system were 0.202 and 0.391, respectively. So, this system is recall oriented system. In the other systems, there is less characteristics on precision and recall, therefore, they were balanced systems.

Figure 6 is the histogram of the difficulty of the question set for Subtask 2. The difficulty of each question in this case was calculated as the average of the F-measures everything the system achieved for that question.

One of the easiest questions was QAC2-20046-01, the MF of which is 0.71. All the systems returned correct answers in this question and two systems returned complete answers, F=1.0. The other easiest questions were QAC2-20030-01(MF=0.587), QAC2-20159-01(MF=0.512) and QAC2-20077-01(MF=0.422). All of them were questions on person name and their question expressions were " (who is ...)", which is very simple type of question.

8 Conclusion

We have given an overview of the Question Answering Challenge (QAC2). We defined three kinds of QA tasks, which utilized newspaper articles covering a period of two years, and an evaluation method for the tasks. We also reported the results of these tasks in terms of statistical results based on MRR and MF and discussed the level of difficulty the questions for each task from the point of view of the average of the systems' performance.

Acknowledgements

We would like to express our thanks to all of the task participants and members of the organizing committee. We would also like to say thank you to the staff of the NII for their support and for providing us with the opportunity to do this kind of evaluation.

References

[1] J. Fukumoto and T. Kato. An overview of question and answering challenge (QAC) of the next NTCIR workshop. In *Proceedings of the Second NTCIR Workshop Meeting*, pages 375–377, 2001.

[2] J. Fukumoto, T. Kato, and F. Masui. Question and 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 (QAC-1)*, pages 1–10, 2002.

[3] J. Fukumoto, T. Kato, and F. Masui. Question and answering challenge (QAC-1) : Question Answering Evaluation at NTCIR workshop 3. In *AAAI 2003 Spring Symposium New Direction in Question Answering*, pages 122–133, 2003.

[4] NTCIR (NII-NACSIS Test Collection for IR Systems) Project. http://research.nii.ac.jp/ntcir/index-en.html.

[5] TREC Home Page, 2003.

[6] J. Burger, C. Cardie et.al. Issues, tasks and program structures to roadmap research in question & answering (q&a), 2001. NIST DUC Vision and Roadmap Documents, http://www-nlpir.nist.gov /projects/duc/roadmapping.html.

[7] E.M.Voorhees and D.M.Tice. Building a question answering test collection. In *Proceedings of SIGIR2000*, pages 200–207, 2000.

[8] Information retrieval and extraction exercise (IREX). http://cs.nyu.edu/cs/projects/proteus/irex/.

[9] Proceedings of 7th message understanding conference (MUC-7), darpa, 1998.

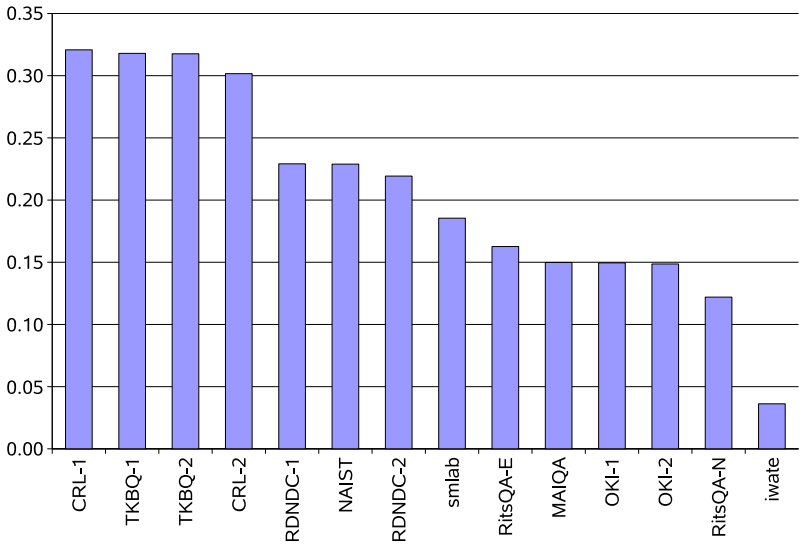

Figure 4. Average F-measure of participant systems in Subtask 2

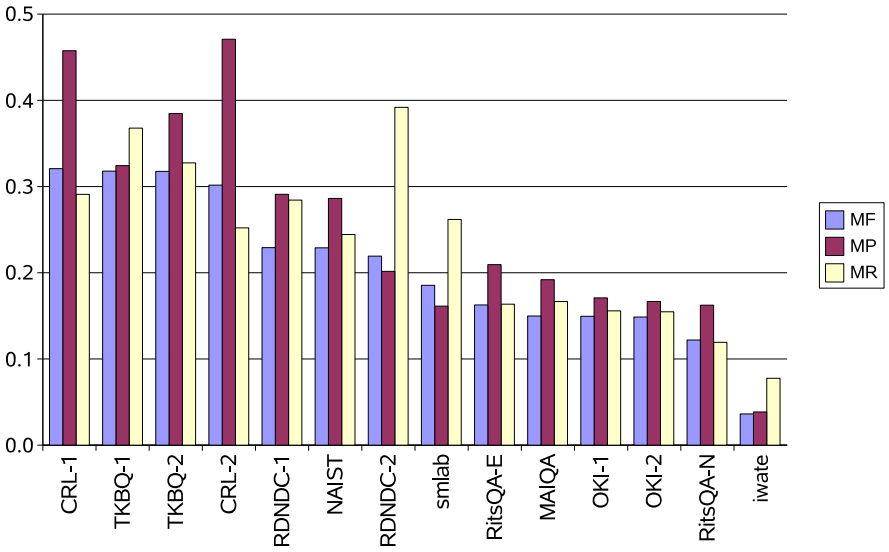

Figure 5. MF, Precision and Recall of participant systems

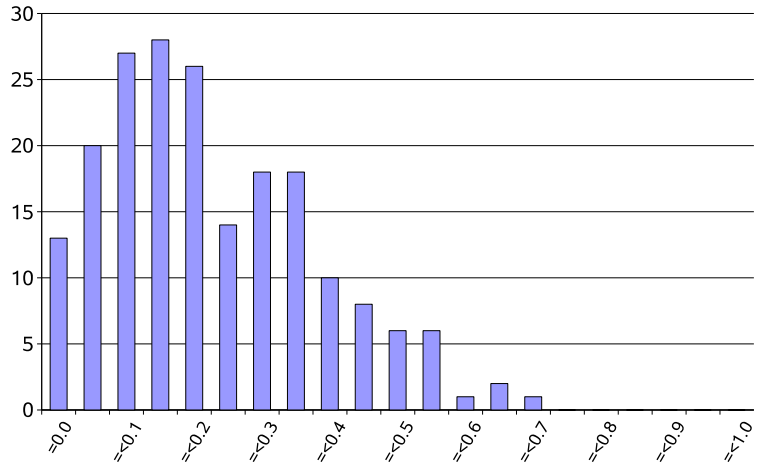

Figure 6. Average over every system's F measure of the question in Subtask 2