An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6

Junichi Fukumoto
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 Japan
fukumoto@media.ritsumei.ac.jp

Tsuneaki Kato
University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902 Japan
kato@boz.c.u-tokyo.ac.jp

Fumito Masui
Mie University
1515 Kamihama-cho, Tsu, Mie 514-8507 Japan
masui@ai.info.mie-u.ac.jp

Tsunenori Mori
Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501 Japan
mori@forest.eis.ynu.ac.jp

Abstract

In QAC-4, we defined question answering task using any type of question, mainly focused on non-factoid questions. There are 8 participants and 14 runs from these participants. In the evaluation, four kinds of criterion were used for some portion of participants answer set. The evaluation results showed some of the participant systems could focus on the area which correct answer contents exist but have tendency to fail to extract correct answer areas. It is caused by complex question types and difficulty of correct answer scope extraction.

1 Introduction

Question Answering Challenge (QAC) was carried out from the NTCIR Workshop 3[Fukumoto et al. (2002)] and this QAC is the forth evaluation of question answering. 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. The previous three QACs are mainly focused on factoid type questions [Fukumoto et al. (2002)][Fukumoto et al. (2004)] [Fukumoto et al. (2003)] and tasks are list questions and Information Access Dialogue (IAD) questions [Kato et al. (2004b)][Kato et al. (2005)][Kato et al. (2004a)]. However, there are many types of questions such as why-type questions and how-type question. The purposes of the QAC-4 are to try question answering beyond factoid type questions that is to expand question types and to reveal how to extract answers for these type questions. Moreover, it is also necessary to establish how to evaluate such questions automatically. For evaluation of factoid type questions, we have applied matching with correct answers but this technique is not sufficient because answers for these type questions tend to be longer one. Human level evaluation is also very difficult problem. QAC-4 will welcome participants who will join evaluation (automatic or human evaluation) for such type questions. Evaluation of QAC-4 will be some sort of pilot task for the next QAC evaluation. QAC-4 will concentrate on question answering research of how to extract or generate long answers for non-factoid questions and how to evaluate such answers through discussions with participants and task organizers.

2 Task Description of QAC-4

In QAC-4, we have set two kinds of tasks: Question answering track and Evaluation track. Question an-

swering track is aimed at evaluation of question answering using non-factoid questions. We mainly focused on non-factoid type question in this track, however, there might be short answer questions in question set. In Evaluation track, we aimed at open evaluation because there are many ways of evaluation, for example, human evaluation and automatic evaluation will be possible and there are also many methods in these evaluations.

QAC-4 tasks

1. Question Answering Track
    - Question will be non-factoid type question such as why-type, definition, question which has answer consists of multiple noun phrases.
    - There will be 100 questions which are natural ones, not generated using target documents.
    - System returns a set of answers for a question.
    - Participants have to return human made answers for questions.
    - Systems answers will be used for Evaluation Track.
2. Evaluation Track
    - Participants can join evaluation of QA Track with their own evaluation method.
    - Participants will evaluate correctness and appropriateness for given questions using their own evaluation method, human evaluation or automatic evaluation.

We have set Evaluation track at QAC-4 but there is only one participation for this track from QAC task organizers. Therefore, task organizers have done human evaluation and automatic evaluation which are only evaluations in QAC-4.

Question set

We used four year (1998-2001) Mainichi Newspaper articles for target document set of QAC-4 and made 120 questions using these document sets. In question development, we showed several topics extracted from target documents and asked a question developing person make arbitrary questions toward these topics, which will be basically beyond factoid questions. We also asked her make answers for the questions. Answers are extracted from and documents and are allowed to be modified to appropriate ones. Some sentences which include answer expressions are also extracted to indicate answer scope for further research. For Formal Run, we selected 100 questions and their answers from these question sets.

Task participant list

There were 8 participants and 14 runs at QAC-4 as shown in Table 1. In this Table there are team names, system ID names of the teams and the number of runs. System ID is indicated by system ID name with Run ID in evaluation.

Table 1: Participants for QAC-4

Participant name	Sys.ID	Runs
Aoyama Gakuin Univ.	HARAD	1
Carnegie Mellon Univ.	LTI-J	1
Hokkaido Univ. + Mie Univ. + Otaru Univ. of Commerce	HOMIO	2
National Institute of Information and Communications Tech.	NICT	2
NTT Communication Science Laboratories	NCQAW	2
Ritsumeikan Univ.	RitsQ	1
Toyohashi Univ. of Technology	TTH	3
Yokohama National Univ.	Forst	2

There are two more participants at registration but unfortunately they could not send their results.

Evaluation schdule

Evaluation schedule is as follows:

Apr. 15, 2006	Call For Participation
May 31, 2006	Deadline of task participation
Jun. 22, 2006	Sample question set delivery
Sep. 25, 2006	Question set delivery
Oct. 20, 2006	System results due
Nov. 1, 2006	Start of Evaluation
Feb. 9, 2007	Evaluation results release

Question and Answer File Formats Description

In the following format description, unless specified others, one byte characters are used for all numbers and alphabets. A [xxx] type notation

stands for non-terminal symbols, and <CR> represents carriage return.

The Question File consists of lines with the following format.

```
[QID]: ‘‘[QUESTION]’’<CR>
```

[QID] has a form of [QuestionSetID]-[QuestionNo]-[SubQuestionNo]. [QuestionSetID] consists of four alphanumeric characters. [QuestionNo] and [SubQuestionNo] consists of five and two numeric characters, respectively. [QUESTION] is a series of two byte characters. “、” and “。” are used for punctuation marks. “？” is not used.

[Examples]
QAC4-00001-00: “世界遺産条約とはどのような条約ですか。”<CR>
QAC4-00002-00: “「琉球王国のグスク及び関連遺産群」が世界遺産に登録された理由は何ですか。”<CR>
QAC4-00003-00: “世界遺産はどのようにして決めるのですか。”<CR>

Answer File Format

The Answer File consists of lines with the following format (so called CSV format).

```
[QID], ‘‘[Answer]’’, [ArticleID],
 [MFlag]<CR>
(, ‘‘[Answer]’’, [ArticleID],
 [MFlag]<CR> )*
```

where (...)* is Kleene star, and specifies zero or more occurrences of the enclosed expression. [QID] is the same as in the question file format above. It must be unique in the file, and ordered identically with in the corresponding question file. It is allowed, however, that some of [QID]s do not list at the file. [Answer] is the answer to the question, and a series of two byte characters. [ArticleID] is the identifier of the article or one of the articles used in the process of deriving the answer. It consists of nine numbers followed with JA-. [MFlag] is “E” or “M”. It will be “E” if the answer string is a part of document [ArticleID]. It will be “M” the answer string is modified from extracted answer string from the document [ArticleID]. In the answer file, the line beginning with “#” is a comment. You may include any information, such as a support or context of your answer, as comments.

The following is an example of the answer to the question:

[Examples]
QAC4-00001-00: “世界遺産条約とはどのような条約ですか。”

There are two answers for the above question and answer file example is shown below.

QAC0-10001-00, “１９７２年の第１７回ユネスコ総会で採択された国際条約。”, JA-001207101, E <CR>
, “「世界の文化遺産及び自然遺産の保護に関する条約」”, JA-980227197, E <CR>

3 Evaluations

We have received a number of answers for Formal Run questions from 14 system runs of 8 teams. We also asked all the participants to submit correct answer set made by human and used these submitted correct answer set for development of correct answer set. We are planning to use all the submitted answer sets for evaluation, however, there are too many answers (4,499 answers for 100 questions) and it may take a lot of time to merge these answers and evaluate their appropriateness. Finally, we have developed correct answer set from the prepared correct answer set (1,171 answers for 100 questions) from question answer set and some of the submitted answer set to because of short evaluation time.

In the correct answer set, each answer has its answer number. If two answers have different meaning, they have different answer numbers. If two answers have the same kind of meaning, they have the same answer number although their surface expressions are different. That is, such answers are some sort of paraphrase. Each answer has an article number which indicates where the answer is taken from. There are cases that the same answer expression exists in several articles but such multiple answers are omitted in the correct answer set. Therefore, we did not use information of article number for evaluation. This information is only for reference.

We have received 14,050 answers for 100 questions in sum and it is impossible to evaluation all the system results in a short evaluation period. We asked all the participants to reduce their number of answers. If a participant sent us the reduced answer set, we used the results for evaluation. If a participant could not send such answer set, we selected top four answers from its original submitted answer

set. We have used these resubmitted reduced answer sets of all the participants for evaluation. The sum of the reduced answers is

Human evaluation was done in the following evaluation criterion.

- Level A:

  System answer has almost the same contents as one of the correct answers and there is few information expect for the same contents in the answer. If there is an additional expression which has no effect on the contents, this case is recognized in this level.

- Level B:

  System answer includes the contents of one of the correct answers and the other information, and the main contents is not the contents of the correct answer.

- Level C:

  System answer includes some part (not all one) of the contents of the correct answers.

- Level D:

  System answer includes no information of any of the contents of the correct answers. There is a case that some surface expression of the correct answer is included in the system answer. If this expression is used for the other meaning, this case will be Level D. If this expression is used for the same meaning of a part of the correct answer, this case will be Level C.

Human evaluation was done by two assessor but they made the different parts. In this evaluation, information of article ID was not considered because there are too many expressions of one contents and it might take huge time and effort for evaluation with this information. In evaluation of factoid questions, the number of different expressions of the same object would be in reasonable level, but in non-factoid question, variation of contents expression is too much.

4 Results

In QAC-4, evaluation has been done for a pat of submitted results, not for all the data because of short of evaluation time and budget. Therefore, we could not provide any formal information of system performance comparison, which is usually provided in evaluation workshop. The tasks of this QAC-4 are very challenging, and then the status of QAC-4 is some sort of pilot task. However, we could provide human evaluation for part of system results in the above procedure.

Table 2 shows evaluation results of submitted systems. System ID shows system names of task participants and the number means system number of task participant. The participant "Forst" submitted two kinds of results and the first one is "Forst1" and the second one is "Forst2". For example, we have evaluated 317 answers of all the results submitted from the system "Forst2". There were 30, 52, 21 and 214 answers scored A, B, C and D, respectively. Two questions have no answer, and then 317 answers are from 98 questions.

Table 3 shows evaluation results according to questions. Each system submitted several answers or one answer for a question. The column of "correct Qnum" shows the number of questions in which there was at least one correct answer in evaluated answers of system results. The column "wrong Qnum" shows the number of question in which there was no correct answer. The column "including A, B, C and D" shows the number of question in which there was at least one score A, B, C and D answer, respectively. The number of answers of a question is different in this evaluation. This evaluation was done for some portion of submitted results and did not show actual system performance.

5 Discussions

As for question used for evaluation, we have prepared 100 questions in various types. There are why-type questions which require reason for a question, how-type question which require processes for a question, definition type questions which require term definition or descriptions for an inquired object and so on. In QAC-4 question, there are some other question types like questions for process, opinion, effect, situation, mechanism, problems, comments, system, attitudes, merit and so on. In these cases, expressions of question are "what kinds of opinion $\cdots$", "what kinds of merits $\cdots$", "what kinds of comments $\cdots$" and so on. In order to extract answers for these various types of questions which is very difficult task in the current level, it is necessary to recognize information type and score of this information. However, there are some systems which could provide correct answers or a part of answer expressions according to the table shown in Appendix. It means that many sys-

Table 2: Evaluation results of submitted system answers

system ID	all answers	A	B	C	D	no answer
Forst1	591	45	104	34	408	0
Forst2	317	30	52	21	214	2
HOMIO1	100	5	4	7	84	0
HOMIO2	100	3	7	4	86	0
LTI-J	377	24	30	13	310	1
NCQAW1	330	37	15	6	272	32
NCQAW2	323	31	11	4	277	32
NICT1	345	25	65	14	241	0
NICT2	363	6	119	24	214	0
HARAD	204	21	7	7	169	38
RitsQ	286	31	6	14	235	15
TTH1	353	34	36	24	259	0
TTH2	394	22	42	24	306	0
TTH3	354	30	43	26	255	0
sum	4236	344	541	222	3330	120
average	302.6	24.6	38.6	15.9	237.9	8.6

Table 3: Evaluation results according to questions

systemID	answered Qnum	correct Qnum	wrong Qnum	including A	including B	including C	including D	no answer
Forst1	100	73	27	31	54	26	89	0
Forst2	98	50	48	19	29	17	85	2
HOMIO1	100	16	84	5	4	7	84	0
HOMIO2	100	14	86	3	7	4	86	0
LTI-J	99	42	57	19	23	10	96	1
NCQAW1	68	29	39	21	10	6	65	32
NCQAW2	68	20	48	15	8	4	65	32
NICT1	100	57	43	18	42	10	89	0
NICT2	100	77	23	5	67	19	90	0
HARAD	62	19	43	15	5	7	59	38
RitsQ	85	30	55	22	6	10	81	15
TTH1	100	56	44	29	28	19	95	0
TTH2	100	57	43	21	34	21	98	0
TTH3	100	57	43	28	33	21	95	0

tems are able to provide some extracted passages which include answer expression but fail to extract only answer expression.

In human evaluation, we have classified system answers into four types: a correct answer, an answer including correct contents, an answer including a part of correct answer, wrong answer based on correct answer set. This correct answer set was developed from prepared question and answer set plus participant submitted human made answer set. Some of human answer sets sent from some participants consist of more than hundred of answers for one question. In the other cases, the number of human answers submitted from participants were also too much. Then, we chose some of them, not all of them. However, this information will be helpful in order to make correct answer set more complete and wide extensive. In the previous QACs, we have used document ID for evaluation and prepared paraphrased expressions which appeared in documents. But, we could not prepare alternative expressions of answers not so much because variation of answer expressions were too much and scope of these expressions are also very difficult to define.

As for the number of answers for a question, some systems submitted a number of answers as correct answer set and the length of answer expressions is too long. We could not predict this phenomenon before starting evaluation of this task; therefore, we could not evaluate all the submitted results in a short evaluation period. However, we will open all the submitted results and human answers provided from all the participants. We hope it will help research on question answering and its evaluation research in future. Moreover, it is necessary to restrict the size of answer expressions but appropriate size of answer expressions are various in questions. However, answer score might not exceed more than one paragraph except for some special cases. It will be necessary to control size of answer in question answering task in reasonable task setting.

6 Conclusion

In QAC-4, we defined question answering task using any type of question, mainly focused on non-factoid questions. We have prepared 100 questions and answer set for these questions which was developed from prepared question answer set plus some of participants submitted answer sets. In the evaluation, four kinds of criterion were used for some portion of participants answer set. The evaluation results showed some of the participant systems could focus on the area which correct answer contents exist but have tendency to fail to extract correct answer areas. It is caused by complex question types and difficulty of correct answer scope extraction.

In this evaluation, we could developed question answer set with human evaluation, sample answers and some system answers. We hope test set and evaluation materials will be helpful for further progress of question answer researches.

References

J. Fukumoto, T. Kato, and F. 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*, pages 1–10.

J. Fukumoto, T. Kato, and F. Masui. 2003. Question Answering Challenge (QAC-1) an evaluation of question answering tasks at the NTCIR Workshop 3. In *Proc. of AAAI Spring Symposium on New Directions in Question Answering*, pages 122–133.

J. Fukumoto, T. Kato, and F. Masui. 2004. Question Answering Challenge for five ranked answers and list answers - Overview of NTCIR4 QAC2 Subtask 1 and 2-. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 283–290.

T. Kato, J. Fukumoto, and F. Masui. 2004a. Handling information access dialogue through QA technologies - a novel challenge for open-domain question answering. In *Proc. of the Workshop on Pragmatics of Question Answering at HLT-NAACL*, pages 70–77.

T. Kato, J. Fukumoto, and F. Masui. 2004b. Question Answering Challenge for information access dialogue :- Overview of NTCIR4 QAC2 Subtask 3-. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 291–297.

T. Kato, J. Fukumoto, and F. Masui. 2005. An overview of NTCIR-5 QAC3. In *Proc. of the Fifth NTCIR Workshop Meeting*, pages 361–372.

Appendix A. Number of answered systems per question

QID	answered	correct	wrong	include A	include B	include C	include D	no answer
QAC4-00001-00	14	12	3	7	6	4	9	0
QAC4-00002-00	12	6	6	3	1	3	12	2
QAC4-00003-00	12	7	3	0	2	6	10	2
QAC4-00004-00	12	6	5	2	4	0	12	2
QAC4-00005-00	14	6	5	3	2	2	12	0
QAC4-00006-00	14	7	8	3	4	3	14	0
QAC4-00007-00	14	11	3	5	7	0	12	0
QAC4-00008-00	14	3	6	2	1	0	14	0
QAC4-00009-00	12	1	4	1	0	0	12	2
QAC4-00010-00	14	6	5	4	3	0	13	0
QAC4-00011-00	12	2	5	0	2	0	12	2
QAC4-00012-00	14	0	7	0	0	0	14	0
QAC4-00013-00	12	6	9	4	2	4	10	2
QAC4-00014-00	12	6	1	4	2	0	10	2
QAC4-00015-00	14	7	8	3	4	0	14	0
QAC4-00016-00	13	12	2	7	0	7	10	1
QAC4-00017-00	13	2	9	2	0	0	13	1
QAC4-00018-00	14	12	2	6	5	4	12	0
QAC4-00019-00	11	3	2	2	0	1	10	3
QAC4-00020-00	13	8	7	4	5	1	12	1
QAC4-00021-00	12	7	7	6	4	0	12	2
QAC4-00022-00	13	12	3	7	7	1	8	1
QAC4-00023-00	13	8	5	1	7	0	13	1
QAC4-00024-00	13	10	4	1	7	9	12	1
QAC4-00025-00	12	6	3	1	3	4	10	2
QAC4-00026-00	14	8	3	1	7	0	14	0
QAC4-00027-00	13	7	4	1	5	4	11	1
QAC4-00028-00	12	11	3	4	9	5	11	2
QAC4-00029-00	12	6	7	4	3	0	12	2
QAC4-00030-00	13	7	4	2	7	1	12	1
QAC4-00031-00	14	6	4	0	5	1	14	0
QAC4-00032-00	14	5	6	3	2	3	14	0
QAC4-00033-00	11	8	5	4	7	1	9	3
QAC4-00034-00	14	9	6	5	4	3	10	0
QAC4-00035-00	14	7	4	3	5	0	14	0
QAC4-00036-00	13	5	5	2	1	5	13	1
QAC4-00037-00	14	5	6	1	4	0	14	0
QAC4-00038-00	13	6	7	5	1	3	12	1
QAC4-00039-00	13	8	4	1	0	8	13	1
QAC4-00040-00	13	4	4	0	2	3	13	1
QAC4-00041-00	14	6	7	4	4	1	14	0
QAC4-00042-00	13	6	6	5	1	2	12	1
QAC4-00043-00	14	6	8	5	2	2	13	0
QAC4-00044-00	11	6	6	2	5	2	11	3
QAC4-00045-00	13	2	6	0	1	2	13	1
QAC4-00046-00	11	7	10	6	1	1	11	3
QAC4-00047-00	14	1	8	1	0	0	14	0
QAC4-00048-00	11	5	3	2	5	0	11	3
QAC4-00049-00	11	7	4	2	7	2	8	3
QAC4-00050-00	12	4	6	1	4	0	12	2
QAC4-00051-00	13	8	3	2	3	4	11	1
QAC4-00052-00	13	3	6	0	3	0	13	1
QAC4-00053-00	12	5	6	3	3	0	12	2
QAC4-00054-00	13	1	7	0	1	0	13	1

QID	answered	correct	wrong	include A	include B	include C	include D	no answer
QAC4-00055-00	12	9	5	5	4	1	11	2
QAC4-00056-00	13	4	5	1	4	0	10	1
QAC4-00057-00	11	4	4	1	3	0	11	3
QAC4-00058-00	11	1	4	0	1	0	11	3
QAC4-00059-00	13	10	8	9	6	2	7	1
QAC4-00060-00	12	4	7	2	2	0	12	2
QAC4-00061-00	14	6	8	3	4	2	13	0
QAC4-00062-00	14	10	2	4	5	3	14	0
QAC4-00063-00	12	0	5	0	0	0	12	2
QAC4-00064-00	13	1	6	0	1	0	13	1
QAC4-00065-00	12	10	3	4	4	6	11	2
QAC4-00066-00	13	7	3	1	4	3	9	1
QAC4-00067-00	12	1	5	0	1	0	12	2
QAC4-00068-00	13	2	7	0	2	0	13	1
QAC4-00069-00	11	3	4	1	2	0	9	3
QAC4-00070-00	14	3	7	2	1	0	12	0
QAC4-00071-00	14	5	9	4	1	2	14	0
QAC4-00072-00	13	0	7	0	0	0	13	1
QAC4-00073-00	14	2	6	1	1	0	14	0
QAC4-00074-00	11	10	2	2	10	4	7	3
QAC4-00075-00	12	9	2	0	9	2	10	2
QAC4-00076-00	13	7	3	3	3	3	12	1
QAC4-00077-00	12	7	5	3	6	1	11	2
QAC4-00078-00	11	9	3	1	9	7	8	3
QAC4-00079-00	14	6	6	1	5	0	13	0
QAC4-00080-00	13	9	9	6	3	3	13	1
QAC4-00081-00	14	0	7	0	0	0	14	0
QAC4-00082-00	14	3	8	1	0	2	14	0
QAC4-00083-00	11	5	6	1	4	0	10	3
QAC4-00084-00	14	10	6	9	9	1	14	0
QAC4-00085-00	14	13	4	6	7	11	9	0
QAC4-00086-00	12	4	7	2	2	1	10	2
QAC4-00087-00	14	3	6	2	0	1	14	0
QAC4-00088-00	12	4	5	1	0	3	12	2
QAC4-00089-00	13	9	8	7	4	1	13	1
QAC4-00090-00	13	4	9	2	3	1	13	1
QAC4-00091-00	11	7	3	1	7	1	6	3
QAC4-00092-00	13	9	3	2	8	1	12	1
QAC4-00093-00	14	2	6	0	1	1	14	0
QAC4-00094-00	11	7	5	2	6	4	11	3
QAC4-00095-00	13	7	5	1	2	4	13	1
QAC4-00096-00	13	4	4	0	3	1	11	1
QAC4-00097-00	13	9	4	6	6	2	11	1
QAC4-00098-00	14	9	3	2	6	4	12	0
QAC4-00099-00	14	8	8	3	7	1	14	0
QAC4-00100-00	11	6	5	2	4	0	9	3