Question Answering Challenge (QAC-1)
An Evaluation of Question Answering Tasks at the NTCIR Workshop 3

Jun'ichi FUKUMOTO
Ritsumeikan University
fukumoto@cs.ritsumei.ac.jp

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

Fumito MASUI ‡
Mie University ‡
masui@shiino.info.mie-u.ac.jp ‡

Abstract

In this paper we describe the Question Answering Challenge (QAC), a question answering task, and its first evaluation (QAC1). The project was carried out as a task of 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 QAC: Task 1, where questions required five possible answers;, Task 2, where questions had a single answer; and Task 3 ,where there was one answer to a question related to a question in Task 2. We prepared 200 questions for Task 1 and Task 2 and 40 questions for Task 3 at the Formal Run and about 900 questions for the additional run. We conducted a Dry Run and a Formal Run evaluation. There were 16 participants (two of them from among the task organizers) at the QAC1.

Keywords: *Question Answering, Information Extraction, Information Retrieval, user interaction, evaluation tools*

1 Introduction

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

The QAC was proposed at the NTCIR Workshop 2[1]. There was an organizing committee composed of 20 members and four meetings were held to discuss the evaluation method and other issues related to QA. We also created a website for the QAC at http://www.nlp.cs.ritsumei.ac.jp/qac/ in both Japanese and English and started mailing lists.

2 Question Answering task

To evaluate QA technologies, there are several technical aspects to consider for the extraction of answer expressions from knowledge sources. Question type is one aspect of the QA system evaluation. 5W1H type questions, which are question sentences using interrogative pronouns such as who, where, when, what, why and how, are typical of this type. In the QA task, the points of the question sentences are defined as a noun or noun phrase which indicates names of persons, organizations, and various artifacts and facts, such as money, size, date and so on. Moreover, information related to these can also be considered as answer candidates: for example, names of persons, their affiliations, age and status can be an answer; and for names of organizations, their annual profit, year of establishment and so on.

Another aspect to consider is how many answer expressions exist in the knowledge sources. In the TREC QA task[3] [7] [4], there is an assumption that there is only one answer for a question. However, there may be multiple answers or no answers to questions in general. This aspect makes development of a QA system difficult. If it is to be assumed that there will be multiple answers,, the system has to check all answer candidates very carefully. If there is only one answer, the system will choose the highest priority answer from the answer candidates with some priorities.

User interaction technology requires actual interaction between the computer and person. In actual QA between people, there are typically several interactions which take place in order to confirm the intention of the questions and so on. Information extraction that works in the general domain is also an important technology for achieving a genuine QA system.

Answer text retrieval is an essential technology for

a QA system. In the first stage of question answering, several target texts are retrieved for answer extraction using several key words in a given question sentence. The longer the sentence is, the more information exists for text retrieval. However, there are some cases where several meaningless key words are embedded in the question sentence.

3 Task Definition of QAC1

We will briefly describe the task definition of the QAC1. For target documents, we used Japanese newspaper articles spanning a period of two years(1998 and 1999) taken from the Mainichi Newspaper. In the QAC1, questions used for evaluation were short answer questions and the answers 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[6] and IREX[5] 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. In other words, answer expressions that did not exist in newspaper articles were permitted. Moreover, paraphrased expressions were also permitted as answer expressions. Nevertheless, accepting such answer expressions as being correct required justification on the basis of the contents of the newspaper articles.

We did not make the assumption that only one answer existed for a given question. Furthermore there were instances where there were no answer objects in documents for a given question. Also, if there were multiple answer objects in given documents, the system had to respond by giving all the possible answers in a list form.

We gave one or more follow-up questions for the first question. In Japanese, there were ellipses in the follow-up question. For example, if the first question was a question relating to a person's name, the second question was a question relating to his/her affiliations or other related personal facts.

Paraphrasing was also one aspect of the QA system evaluation. In a target text, an answer expression may have existed in another expression. In this case, the system had to recognize the paraphrased expression as the same as the original one. In other words, some expressions of a question sentence existed in paraphrased ones in a target text. To retrieve such a text to extract the answer expression, identification of the same concept in various expressions is an important aspect of the required technology.

3.1 Definition

According to the above outline of task definitions, we introduced three tasks in the QAC1. The current version of the QAC task definition was presented as follows:

- Task 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 bethe score of the question. Where there are several correct answers, the system will return one of them.

  For example, questions of this task are presented by giving a question consisting of a QID (QAC1-1001-01) and a question sentence in the following way: (The English translation is shown in parentheses.)

  > QAC1-1001-01: “2000 年 10 月 1 日に合併することが決まった通信三社はどこですか。(Which three telecommunications companies decided to merge on October 1, 2000?)”

  For the question, the correct answers are “DDI”, “IDO” and “KDD”. The system has to respond by giving one of them.

  > QAC1-1002-01: “広辞苑第五版はいつ発売されましたか。(When was the fifth edition of the Kojien Japanese dictionary published?)”

  For the question QAC1-1002-01, the correct answer is “11 月 11 日 (November 11)”. If there isan expression “昨 (yesterday)” in the article dated Nov. 12, this answer will also be correct. In the QAC, a relative expression of date is permitted; however, the system has to give evidence that the answer was extracted from the article in this case.

- Task 2

  Task 2 uses the same question set as Task 1 but the evaluation method is different. 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 Task 2.

- Task 3

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

For example, questions in this task are presented in the following way: Question "QAC1-3011-02" is the follow-up question to question "QAC1-3011-01". The "-02" indicates the first follow-up question of the main question, although there is only one follow-up question in the current task definition.

> QAC1-3011-01: "久石譲が音楽を担当した宮崎駿監督の映画は何ですか。(Joe Hisaishi was a music director for which of Hayao Miyazaki's films?)"
> QAC1-3011-02: "北野武の映画は何ですか。(What is the name of the film directed by Takeshi Kitano?)"

3.2 Support information

The system is required to return support information for each answer to the questions, although it is optional. In the current definition, we assume the support information as being one document ID which will be evidence of the replied answer.

4 Question development for evaluation

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. We developed about 1200 questions of various question types that sometimes included paraphrasing. Moreover, all the task participants were required to submit about 20 questions by the time of the Formal Run. Some of them were to be used for the evaluation and others were to be open as test corrections of the QA data. The details of these questions are summarized in Table 1.

Table 1. Prepared questions for QAC1

developer	number
task organizer	1202
task participants	200
Total	1402

5 Evaluation Method

5.1 Task 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 Task 1. If n set of answers were correct, the Mean Reciprocal Rank (MRR) could be calculated as follows:

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

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

For example, the following question "QAC1-1001-01" has an answer of three companies such as DDI, IDO and KDD.

> QAC1-1001-01: "2000 年 10 月 1 日に合併することが決まった通信三社はどこですか。(Which three telecommunications companies decided to merge on October 1st, 2000?)"
> Correct answer: DDI, IDO, KDD

The three kinds of answer evaluation were as follows:

- Response1: NTT, IDO, AT&T, NII, KDD
  RR=0.5
- Response2: AT&T, BT, DDI, IDO, KDD
  RR=0.33
- Response3: DDI, AT&T, BT,NII, Docomo
  RR=1.0

The underlined answers were the correct ones. In Response 1, the system returned five answers in the above order and the second one and fifth one were correct. Therefore, the RR value of the best answer (the second one) was the score for this answer. In Response 2, the third, fourth and fifth answers were correct, and the RR value is 0.33. In Response 3, only the first answer was correct, and the RR value was therefore 1.0.

5.2 Task 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. Mean F-Measure (MF)

was used for the evaluation of Task 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.

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

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

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

For example, system responses for the same question of Task 1 "QAC1-1001-01" and their evaluations were presented as follows:

- Response1: NTT, IDO, AT&T, KDD
  P=2/4, R=2/3, F=0.57
- Response2: IDO, 日本移動通信 (IDO), KDD
  P=2/3, R=2/3, F=0.67

As well as the above examples, the underlined answers were also correct. In Response 1, the system returned four answers for the question and the second and third ones were correct. Therefore, two of the four answers were correct, indicating an accuracy of 2/4. Also, two of three correct answers were detected, indicating a recall of 2/3. In Response 2, the correct answers were the first and third ones, indicating an accuracy of 2/3. Also, two of three correct answers were detected, indicating a recall value of 2/3. In this case, however, the Response 2 is the same organization as Response 1, but the same elements were ignored when they represented the same organization.

In Task 1 and Task 2, there was an incidence of a "No Answer" question. When there was a No Answer question and a system gave no answer, the score of this question was 1.0 (F-measure). On the other hand, if a system gave some answer for such No Answer questions, the score was zero.

5.3 Task 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 Task 2, that is MF.

5.4 Scoring Tool

We developed a scoring tool, written in Perl language, to help with the participants' evaluation. This tool can check whether the answers of a system are correct or not by comparing the correct answer and the output. The tool can show each answer evaluation and some statistics of a task.

The scoring tool was able to provide all the results of the answer checking to obtain information about whether each answer was correct or not. Figure 1 shows a part of a sample output.

```
QAC1-1020-01:  インド (India) ○, インドネシア (Indonesia) ○, タイ (Thailand) ×, 米国 (USA) ×, フランス (France) ×
QAC1-1021-01:  φ ○
```

Figure 1. Sample output of scorer (answer check)

If the answer was correct, it was marked with a circle,"○"; otherwise, it received an "×". The symbol for phi,"ϕ," was given when the system did not output any answer to a question. In that situation, a circle, "○", was given if, and only if, there was no answer in the correct answer set.

For statistical results, this tool calculates the sum of correct answers and MRR for Task 1. For Tasks 2 and 3, this tool calculates the sum of correct answers and the mean F-measure. Figure 2 shows a sample output of this tool.

Task1 Results: 35.0 marks out of 200.0 in TASK1
Average score: 0.175

Question	Answer	Output	Correct
200	272	729	38
Recall	Precision	F-measure	MRR/MF
0.139	0.521	0.759	0.175

Figure 2. Sample output of scorer (score)

The first line summarizes the results and statistics for an input result and the following lines show the details of the score. "Question" is the total number of questions in the task and "Answer" is the number of different answers to the questions. "Output" is the number of answers to the input data and "Correct" means the number of correct answers of the input data.

The details of usage of the Scoring Tool are presented in Appendix B.

6 Task Participants

In QAC1, there were fourteen active participants. Task participation of each participant is shown in Table 2. The symbol "⋆" indicates that the team submitted one result for the task. Two symbols mean two

kinds of results were submitted. The last two participants in italics are participants from the QAC task organizers and are not included in the official score results of the QAC1.

Table 2. Task participation

participants name[1]	Task		
	1	2	3
Communication Research Laboratory	⋆	⋆	⋆
Kochi Univ. of Technology		⋆	
Matsushita Electric Ind.	⋆	⋆	
NTT Corp.	⋆⋆	⋆	⋆
NTT DATA Corp.	⋆⋆	⋆	⋆
Nara Institute Science and Technology	⋆	⋆	
National Institute of Advanced Industrial Science and Technology	⋆	⋆	⋆
New York Univ.	⋆		
Oki Electric Ind.	⋆	⋆	⋆
POSTECH	⋆		
The Graduate Univ. for Advanced Studies	⋆	⋆	
Toyohashi Univ. of Technology	⋆	⋆	⋆
Univ. of Tokyo	⋆	⋆	
Yokohama National Univ.	⋆		
Mie Univ.	⋆	⋆	
Ritsumeikan Univ.	⋆	⋆	

7 Runs for Evaluation

7.1 Description of Formal Run

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

- Date of task revealed: Apr. 22, 2002 (Mon.) AM (JST)
- The result submission due: Apr. 26, 2002 (Fri.) 17:00 (JST)
- Number of questions:
  - Task 1: 200
  - Task 2: 200 (same as Task 1)
  - Task 3: 40 (follow up questions of Task 1 questions)

7.2 Additional QA runs

Task participants were required to evaluate about 900 questions in order to provide more evaluation material and develop better QA test collections. This was conducted after the Formal Run according to the following schedule.

- Delivery of questions: May. 13, 2002 (Mon.)
- Submission due: May. 24, 2002 (Fri.)
- Submission format: same as Formal Run Formats

The submitted results were pooled and were to be delivered after evaluation.

8 Results and Discussion

8.1 Task Analysis

In this subsection, we provide a summary of the results of the QAC formal run according to each task. Throughout this chapter, the system IDs used in the figures are names that the participants gave to their own systems; some of them have been abbreviated due to space constraints.

Task 1

Fifteen systems participated in the Task 1. The accuracy the participating systems achieved in the mean reciprocal rank (MRR) is depicted in Figure 3. The most accurate system achieved 0.61 in the MRR. This system returned correct answers in the first rank to more than half of the questions, and in up to the fifth rank for more than three fourths of the questions.

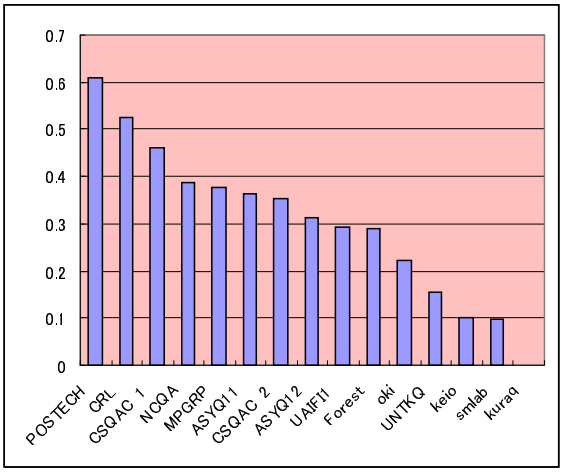

Figure 3. MRR of participant systems in Task 1

In addition to the MRR standard, we tried evaluating the systems using two other types of criteria. The first was the ratio of a system's correct answers in the first rank. The second was the ratio of a system's correct answers up to the fifth rank. Those two criteria

[1] Participant name is taken from affiliation of the first author of presented paper.

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 Task 1.

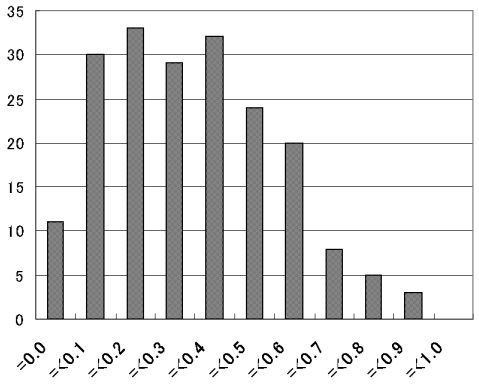

Figure 4. Average over every system's RR of the question in Task 1

Figure 4 is the histogram of the difficulty of the question set of Task 1. The difficulty of each question is calculated as the average of the reciprocal ranks all the systems achieved for that question. For 11 questions out of 195 questions (five questions with no answer were excluded), no system could return correct answers. The easiest question was `QAC1-1099-01`, which has an MRR of 0.87 and thirteen systems returned the correct answer in the first rank to this question. The distribution has a smooth curve with one peak, and there is no evidence that the level of difficulty of the questions is divided into two extremes, that is, those which are too difficult and those which are too easy. Therefore, it may be concluded that the question set used in Task 1was suitable for evaluating state of the art QA research.

Task 2

Eleven systems participated in Task 2. The accuracy the participating systems achieved in the mean F-measure (MF) is depicted in Figure 5. The most accurate system achieved 0.36 in the MF. This system always returned a list with one item, and 40% of its answers agreed with one of the correct answer items. Another system always returned a list with ten items, and 45% of its answers included at least one of the correct items, and achieved only 0.09 in the MF. The former strategy is more effective in the current question set, as more than three fourths of the questions have just one correct answer. Other systems seemed to determine the number of items included in its answer list dynamically according to a given question. We should examine several criteria for Task 2 in order to obtain a useful criterion that reflects our belief in the merits of this task.

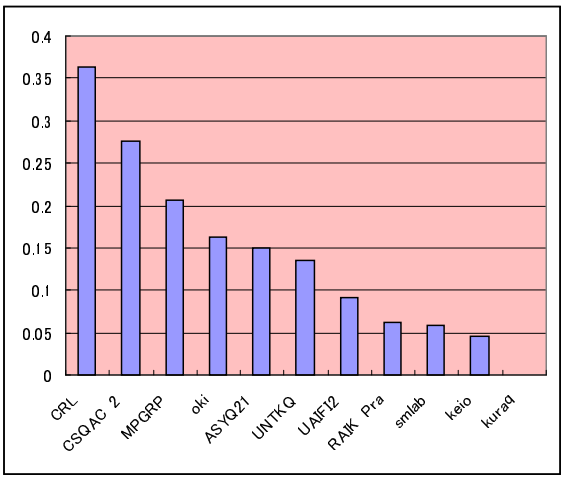

Figure 5. Mean F-measure of participant systems in Task 2

Figure 6 is the histogram of the difficulty of the question set for Task 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. Thirty questions out of 200 questions could not be answered by any system. The number of such questions was much larger than in Task 1, though the comparison may be meaningless as the criteria are different. The easiest question was `QAC1-2136-01`, the MF of which is 0.46. Since we used the same question set for both Task 1 and Task 2, we were able to discuss the characteristics of each task and the relationship between them. We were able to observe some relationships which coincided with our original expectations. Out of 11 questions that no system answered correctly in Task 1, eight questions were not answered by any system in Task 2 either. Ten of the easiest questions in Task 1 and Task 2 according to each set of criteria had six overlaps (`QAC1-XYYY-01` where `X=1` or `2`, `YYY=018, 037, 099, 114, 119, 125,` and `136`). Further examination of those relationships is needed.

Task 3

Task 3 had only six systems participating and the number of questions was just 40. We must be careful to discuss tendencies on this task in this situation. Figure 7 shows the accuracy the participating systems achieved in the MF, the same criterion as that em-

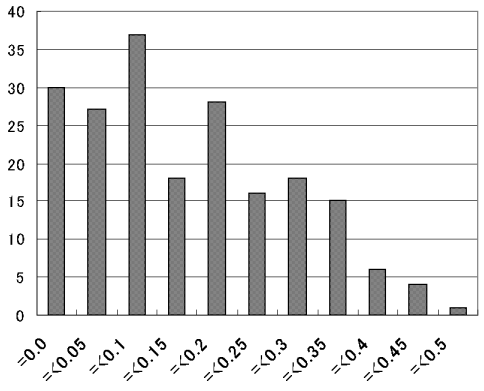

Figure 6. Average over every system's F measure of the question in Task 2

ployed in Task 2. In this task, each problem consisted of two successive questions, and the second question, which contained some anaphoric elements, was the object to be evaluated. The most accurate system achieved 0.17 in the MF. Fourteen questions out of 40, about one third, could not be answered by any system. We must examine thoroughly the characteristics of this task based on these results and call for more participants.

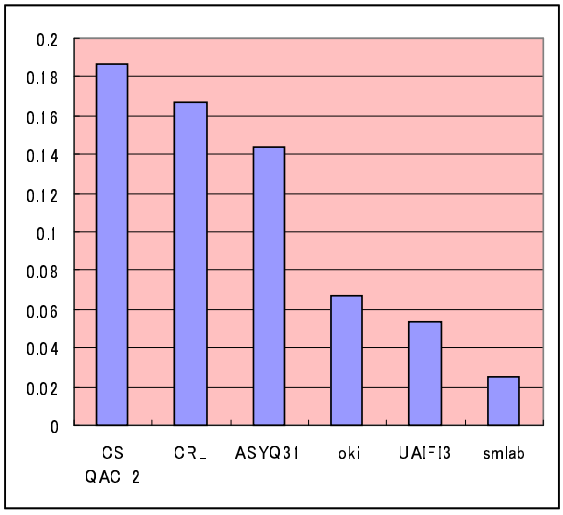

Figure 7. Mean F-measure of participant systems in Task 3

8.2 Question type and system performance

We analyzed the relationship between the types of questions used for the Formal Run evaluation and the performance of participant systems in Table 3. "Qnum" indicates the number of questions in each question type and "correct" shows the average number of correct system in Task 1 evaluation.

Table 3. System performance analysis for each question type in Task 1

question type	Qnum	correct
artifact name	51	5.4
person name	44	6.8
numeric value	21	5.4
location name	15	6.1
date	14	7.6
country name	13	8.7
company name	12	6.3
other name	12	2.8
organization name	7	5.9
distance	3	7
money	3	7.7
time	3	3
quantity	1	3
percentage	1	6

8.3 System features

It should be emphasized that several architectures or techniques have been tried and employed in the participant systems, though it is not possible to discuss here the relations between those attempts and the achieved system performance shown in the previous subsection. For the answer extraction, which extracts answer candidates from retrieved texts or passages, methods using numerical measures are still predominant, in which text is treated as a sequence of words and the distance between keywords and answer candidates characterized by an NE tagger plays an important role. Some promising attempts can be found, however, such as those based on the matching of syntactic or semantic structures or logical forms. Although meticulously hand-crafted knowledge was still invaluable, machine learning techniques were employed for acquiring several kinds of knowledge of the systems let alone for NE tagging. On the other hand, many systems also use existing tools for their morphological analysis and document retrieval. It can perhaps be said that the infrastructure has been put in place for researchers who want to take up the challenge of question answering research. A matter also worth special mention is that, in addition to system developments, many related activities were also undertaken including the proposal of methods of error analysis, construction of a corpus of questions, and various efforts to answer the challenges of speech driven question answering.

9 Conclusion

We have given an overview of the Question Answering Challenge (QAC1). 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.

We are planning to conduct the second evaluation of the QAC as the QAC2 at the NTCIR Workshop 3 scheduled in May 2004. We will continue our analyses of the results from various aspects and develop better task definitions for the QAC2.

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. 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.

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

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

[6] Proceedings of 7th Message Understanding Conference (MUC-7), DARPA, 1998.

[7] http://trec.nist.gov/.

Appendix A: Data Formats

The documents used in Dry Run and Formal Run, are Mainichi Newspaper (1998-1999). The document set in the CD-ROM has to be converted using the program, mai2sgml[2], and the output of this conversion is the standard document files for the dry and formal run. Any information not included in this output, such as keywords attached to each article, is considered additional/extra knowledge, which does not included in original newspaper articles.

Format 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.

Question File Format

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

QAC0-10001-00: ”大学審議会の会長は誰ですか。”<CR>
QAC0-10002-00: ”タージ・マハールはどこにありますか。”<CR>
QAC0-10003-00: ”千葉県の県庁所在地は何市ですか。”<CR>
QAC0-30001-01: ”９８年のアカデミー賞作品賞を受賞した作品は。”<CR>
QAC0-30001-02: ”何という名前の男優が主演したのですか。”<CR>

Answer File Format

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

[QID](, ”[Answer]”, [ArticleID], [HTFlag], [Offset])*<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. The value of the <DOCNO>tag is used for the identifier, which consists of nine numeric characters.

[HTFlag] is “H” or “T”. It will be “H” if the part of article used for deriving the answer and specified in [ArticleID] is the headline, which is the part tagged with <HEADLINE>, “T” if it is the text, which is the part tagged with <TEXT>. This is optional, and when omitted, it should be the empty string , that is, two delimiters, i.e. commas, appear consecutively.

[Offset] is the position of the part used for deriving the answer in the headline or body of text. That position should be represented using number of characters from the beginning of the headline or text. The head of them is represented as zero. A space placed at the beginning of paragraphs is included into characters, while carriage return is not included. This is optional, and when omitted, it should be the empty string , that is, two delimiters, i.e. commas, appear consecutively. “The part of article used for deriving the answer” in the above explanation is typically the portion of the articles where your system extracted the answer from. It does not mean that systems should extract the answer from articles. If your system does not use such extraction for deriving answers, please give us the most relevant position to judge the correctness of your answer. If you can’t specify that anyway, you may omit [HTFlag] and [Offset].

For each question, the quad-gram of ”[Answer]”, [ArticleID], [HTFlag], and [Offset] is repeated more than zero times. In task one, the order of this quad-grams represents the order of the confidence. That is, the most confident

[2]We can obtain this program “mai2sgml” from the URL address, http://lr-www.pi.titech.ac.jp/tsc/tsctools/index-jp.html.

answer candidate should be placed first. The number of candidates is up to five in the dry run. In task two and three, as the answer is a set, the elements of the answer are listed in an arbitrary order.

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.

Examples

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

QAC0-10001-00: "大学審議会の会長は誰ですか。"

It is postulated that the answer is derived using the article shown the below. Three answer candidates are listed. Although all the [ArticleID] are identical in this example, it is not the case in general.

QAC0-10001-00, "石川忠雄", 980701002, T, 24, "町村信孝", 980701002, T, 42, "大学審提言", 980701002, H, 0<CR>

<DOC>
<DOCNO>980701002</DOCNO>
<SECTION> 1 面 </SECTION>
<AE> 無 </AE>
<WORDS>713</WORDS>
<HEADLINE> 大学審提言「勉強させる大学」にーー卒業へ評価厳格化 </HEADLINE>
<TEXT>
「２１世紀の大学像」を検討している大学審議会（石川忠雄会長）は３０日、中間まとめを町村信孝文相に提出した。単位まとめ取り防止や厳格な成績評価で「勉強しなくても（以下略）
</TEXT>
</DOC>

Appendix B: Usage of Scoring Tool

This tool can be used on command line. As an input argument, a filename of the user's system output should be given. In addition, some expressions such as below options can be used.

–answer / **-a filename** Specifies the filename of the correct answer set. The character code in the file needs to be same as the one used in the user's system output.

–help / **-h** Shows help.

–version / **-v** Shows version of the program.

–task / **-t number** Selects tasks. A number, 1, 2, or 3 follows this option. 1, 2, 3 are for TASK1, TASK2, TASK3, respectively.

–extract / **-e number** Shows the inner data. A number, 1, 2, 3, or 4, follows this option.

The number "1" shows information on each question including, question ID, the total number of answers, the number of different answers, answer number, answer, article ID.

ex.1

```
QAC1-1084-01    15    9
1    法隆寺        990131022
2    東大寺        980521199
2    東大寺        981126218
3    薬師寺        981126218
3    薬師寺        981230150
4    興福寺        981126218
⋮    ⋮                ⋮
```

The number "2" shows the answers that the user's system output including, question ID, the number of answers, answer number, answer, article ID.

ex.2

```
QAC1-1084-01    6
0    法隆寺        990131022
0    法隆寺        990131023
1    東京タワー    980521199
2    東大寺        981126218
3    バーミヤン    981126218
4    薬師寺        981230150
```

The number "3" shows information in detail on correct answers that user's system output. The information includes the correct answers and the answer numbers that correspond to the answer numbers in the correct answer set. The symbol '-' that precedes the answer number means that the answer is correct, but the article ID might not be correct.

ex.3

```
法隆寺    |    1
法隆寺    |   -1
東大寺    |    2
薬師寺    |    3
  :       :    :
```

The number "4" shows the score given to each question in Task2. The option is valid for only Task2. Question ID, the number of correct answers, the number of answers that user's system output, the number of correct answers that user's system output and F-measure score for each answers that user's system output.

ex.4

```
QAC1-2146-01:   1   5   1   0.333333
QAC1-2147-01:   1   1   1   1.000000
QAC1-2148-01:   2   5   0   0.000000
QAC1-2149-01:   3   1   1   0.500000
      :         :   :   :       :
```

5 shows the result of answer checking. Question ID, question, list of correct answer and whole answers that user's system output. The correct answer that user's system output are marked with asterisk. The option is valid for only Task1.

ex.5

```
QAC1-1046-01  “奈良の世界遺産にはどのような
ものがありますか”
CORRECT ANSWER:  薬師寺 東大寺 法隆寺
 平常宮跡 興福寺 春日大社 春日山原始林
 唐招提寺 元興寺
法隆寺 *
東京タワー
東大寺 *
バーミヤン
薬師寺 *
```