Handling Information Access Dialogue through QA Technologies – A novel challenge for open-domain question answering –

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

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

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

Noriko Kando
National Institute of Informatics
kando@nii.ac.jp

Abstract

A novel challenge for evaluating open-domain question answering technologies is proposed. In this challenge, question answering systems are supposed to be used interactively to answer a series of related questions, whereas in the conventional setting, systems answer isolated questions one by one. Such an interaction occurs in the case of gathering information for a report on a specific topic, or when browsing information of interest to the user. In this paper, first, we explain the design of the challenge. We then discuss its reality and show how the capabilities measured by the challenge are useful and important in practical situations, and that the difficulty of the challenge is proper for evaluating the current state of open-domain question answering technologies.

1 Introduction

Open-domain question answering technologies allow users to ask a question in natural language and obtain the answer itself rather than a list of documents that contain the answer. These technologies make it possible to retrieve information itself rather than merely documents, and will lead to new styles of information access (Voorhees, 2000).

The recent research on open-domain question answering concentrates on answering factoid questions one by one in isolation from each other. Such systems that answer isolated factoid questions are the most basic level of question answering technologies, and will lead to more sophisticated technologies that can be used by professional reporters and information analysts. On some stage of that sophistication, a cub reporter writing an article on a specific topic will be able to translate the main issue addressed by his report into a set of simpler questions and then pose those questions to the question answering system (Burger et al., 2001).

In addition, there is a relation between multi-document summarization and question answering. In his lecture, Eduard Hovy mentioned that multi-document summarization may be able to be reduced into a series of question answering (Hovy, 2001). In SUMMAC, an intrinsic evaluation was conducted which measures the extent to which a summary provides answers to a set of obligatory questions on a given topic (Mani et al., 1998). Those suggest such question answering systems that can answer a series of related questions would surely be a useful aid to summarization work by human and by machine.

Against this background, question answering systems need to be able to answer a series of questions, which have a common topic and/or share a local context. In this paper, we propose a challenge to measure objectively and quantitatively such an ability of question answering systems. We call this challenge QACIAD (Question Answering Challenge for Information Access Dialogue). In this challenge, question answering systems are used interactively to participate in dialogues for accessing information. Such information access dialogue occurs such as when gathering information for a report on a specific topic, or when browsing information of interest to the user. Actually, in QACIAD, the interaction is only simulated and systems answer a series of questions in a batch mode. Although such a simulation may neglect the inherent dynamics of dialogue, it is a practical compromise for objective evaluation and, as a result, the test sets of the challenge are reusable.

Question answering systems need a wide range of abilities in order to participate in information access dialogues (Burger et al., 2001). First, the systems must respond in real time to make interaction possible. They must also properly interpret a given question within the context of a specific dialogue, and also be cooperative by adding appropriate information not mentioned explic-

itly by the user. Moreover, the systems should be able to pose a question for clarification to resolve ambiguity concerning the user's goal and intentions, and to participate in mixed initiative dialogue by making suggestions and leading the user toward solving the problem. Among these various capabilities, QACIAD focuses on the most fundamental aspect of dialogue, that is, interpreting a given question within the context of a specific dialogue. It measures context processing abilities of systems such as anaphora resolution and ellipses handling.

This paper is organized as follows. The next chapter explains the design of QACIAD. The following three chapters discuss the reality of the challenge. First, we explain the process of constructing the test set of the challenge and introduce the results of a study conducted during this process which show the validity of QACIAD. That is, QACIAD measures valid abilities needed for participating in information access dialogues. In other words, the ability measured by the challenge is crucial to the systems for realizing information access dialogues for writing reports and summaries. Second, we show the statistics of pragmatic phenomena in the constructed test set, and demonstrate that the challenge covers a wide variety of pragmatic phenomena observed in real dialogues. Third, based on a preliminary analysis of the QACIAD run, we show that the challenge has a proper difficulty for evaluating the current state of open-domain question answering technologies. In the last two chapters, we discuss problems identified while constructing the test set and conducting the run, and draw some conclusions.

2 Design of QACIAD

2.1 History

The origin of QACIAD comes from QAC1 (Question Answering Challenge), one of the tasks of the NTCIR3 workshop conducted from March 2001 through October 2002 (NTCIR, 2001). QACIAD was originally proposed in March 2001 as the third subtask of QAC1, its formal run was conducted in May 2002 (Fukumoto et al., 2001; Fukumoto et al., 2002; Fukumoto et al., 2003), and the results were reported at the NTCIR3 workshop meeting in October 2002. The current design of QACIAD reported in this paper is based on that challenge and is the result of extensive elaboration. The design of the challenge and construction of the test set were performed from January 2003 through December 2003. The formal run was conducted in December 2003, as a subtask of QAC2, which in turn is a task of the NTCIR4 workshop (NTCIR, 2003).

2.2 QAC as a common ground

QAC is a challenge for evaluating question answering technologies in Japanese. It consists of three subtasks including QACIAD, and the common scope of those subtasks covers factoid questions that have names as answers. Here, names mean not only names of proper items (named entities) including date expressions and monetary values, but also common names such as names of species and names of body parts. Although the syntactical range of the names approximately corresponds to compound nouns, some of them, such as the titles of novels and movies, deviate from that range. The underlying document set consists of two years of articles of two newspapers in QAC2, and one newspaper in QAC1. Using those documents as the data source, the systems answer various open-domain questions.

From the outset, QAC has focused on question answering technologies that can be used as components of larger intelligent systems and technologies that can handle realistic problems. It persists in requesting exact answers rather than the text snippets that contain them with the cost of avoiding handling definition questions and why questions, because such answers are crucial in order to be used as inputs to other intelligent systems such as multi-document summarization systems. Moreover, as such a situation is considered to be more realistic, the systems must collect all the possible correct answers and detect the absence of an answer. Therefore two subtasks, one of which is QACIAD, request systems to return one list of answers that contains all and only correct answers, while the other subtask requests systems to return a ranked list of possible answers as in TREC-8. In both subtasks, the presence of answers in the underlying documents is not guaranteed and the number of answers is not specified, so these subtasks are similar to the list question task in the TREC-2003 style rather than the TREC-10 style (TREC, 2003).

2.3 Information access dialogue

Considering scenes in which those question answering systems participate in a dialogue, we classified information access dialogues into the following two categories. As discussed later, dialogues in a real situation may have different features in their different portions; the classification just shows two extremes.

Gathering Type The user has a concrete objective such as writing a report and summary on a specific topic, and asks a system a series of questions all concerning that topic. The dialogue has a common global topic, and, as a result, each consecutive question shares a local context.

Browsing Type The user does not have any fixed topic of interest; the topic of interest varies as the dialogue progresses. No global topic covers a whole dialogue but each consecutive question shares a local context.

This paper proposes the design of the challenge, which can measure the abilities of question answering systems

useful in such dialogues.

2.4 The setting

QACIAD requests participant systems to return all possible answers to a series of questions, each of which is a factoid question that has names as answer. This series of questions and the answers to those questions comprise an information access dialogue. Two examples of the series of questions are shown in Figure 1, which were picked up from our test set discussed in the next chapter. Series 14 is a series of a typical gathering type, while series 22 of a typical browsing type. In QACIAD, a number of series (in the case of our test set, 36 series) are given to the system at once and systems are requested to answer those series in a batch mode. One series consists of seven questions on average. The systems must identify the type to which a series belongs, as it is not given. The systems need not identify the changes of series, as the boundary of series is given. Those, however, must not look ahead to the questions following the one currently being handled. This restriction reflects the fact that QACIAD is a simulation of interactive use of question answering systems in dialogues. This restriction, accompanied with the existence of two types of series, increases the complexity of the context processing that the systems must employ. For example, the systems need to identify that series 22 is a browsing type and the focus of the second question is Yankee stadium rather than New York Yankees without looking ahead to the following questions. Especially in Japanese, since anaphora are not realized often and the definite and indefinite are not clearly distinguished, those problems are more serious.

2.5 Evaluation measure

In QACIAD, as the systems are requested to return one list consisting all and only correct answers and the number of correct answers differs for each question[1], modified F measure is used for the evaluation, which takes account of both precision and recall. Two modifications were needed. The first is for the case where an answer list returned by a system contains the same answer more than once or answers in different expressions denoting the same item. In that case, only one answer is regarded as the correct one, and so the precision of such answer list decreases. Cases regarded as different expressions denoting the same item include a person's name with and without the position name, variations of foreign name notation, differences of monetary units used, differences of time zone referred to, and so on. The second modification is for questions with no answer. For those questions, modified F measure is 1.0 if a system returns an empty list as the answer, and is 0.0 otherwise.

[1]It is a special case that the number of answers is just one for all questions shown in Figure 1.

Series 14
When was Seiji Ozawa born?
Where was he born?
Which university did he graduate from?
Who did he study under?
Who recognized him?
Which orchestra was he conducting in 1998?
Which orchestra will he begin to conduct in 2002?

Series 22
Which stadium is home to the New York Yankees?
When was it built?
How many persons' monuments have been displayed there?
Whose monument was displayed in 1999?
When did he come to Japan on honeymoon?
Who was the bride at that time?
Who often draws pop art using her as a motif?
What company's can did he often draw also?

Figure 1: Examples of series of questions

The judgment as to whether a given answer is correct or not takes into account not only an answer itself but also the accompanying article from which the answer was extracted. When the article does not validly support the answer, that is, assessors cannot understand that the answer is the correct one for a given question by reading that article, it is regarded as incorrect even though the answer itself is correct. The correctness of an answer is determined according to the interpretation of a given question done by human assessors within the given context. The system's answers to previous questions, and its understanding of the context from which those answers were derived, are irrelevant. For example, the correct answer to the second question of series 22, namely when the Yankee stadium was built, is 1923. If the system wrongly answers the Shea stadium to the first question, and then "correctly" answers the second question 1964, the year when the Shea stadium was built, that answer to the second question is not correct. On the other hand, if the system answers 1923 to the second question with an appropriate article supporting it, that answer is correct no matter how the system answered the first question.

3 Constructing a Test Set and Usefulness of the Challenge

We collected and analyzed questions for two purposes. The first purpose was to establish a methodology for constructing a test set based on the design of QACIAD discussed in the previous chapter. The second purpose was

to confirm the reality of the challenge, that is, to determine whether it is useful for information access dialogues to use question answering systems that can answer questions that have names as answers.

3.1 Collecting questions

Questions were collected as follows. Subjects were presented various topics, which included persons, organizations, and events selected from newspaper articles, and were requested to make questions that ask for information to be used in the report on that topic. The report is supposed to describe facts on a given topic, rather than contain opinions or prospects on the topic. The questions are restricted to wh-type questions, and natural series of questions containing anaphoric expressions and so on were constructed. The topics were presented in three different ways: only by a short description of the topic, which corresponds to the title part of the TREC topic definition; with a short article or the lead of a longer article, which is representative of that topic and corresponds to the narrative part of the TREC topic definition; and with five articles concerning that topic. The number of topics was 60, selected from two years of newspaper articles. Thirty subjects participated in the experiment. Each subject made questions for ten topics for each topic presentation pattern, and was instructed to make around ten questions for each topic. It is worth noting that the questions obtained were natural in both content and expression since in this experiment the subjects did not consider whether the answers to their questions would be found in the newspapers, and some subjects did not read the articles at all.

This time, for the test set construction and preliminary analysis, 1,033 questions on 40 topics, made by three subjects for each topic with different topic presentation patterns, were used. All of the questions collected are now being analyzed extensively, especially on the differences among questions according to the topic presentation pattern.

3.2 Analysis of the questions

Our main concern here is how many of the questions collected fall into the category of questions that the current question answering systems could answer. In other words, how many of the questions can be answered by a list of names? In the case the majority of them fall into such a category, it is realistic to use question answering systems for information access dialogues and the challenge on such abilities must be useful.

Table 1 shows the classification of questions according to the subject asked. In the case where users ask questions to get information for a report, the number of why questions is relatively small. Moreover, there were fewer questions requesting an explanation or definition than expected, probably because questions such as "Who is Seiji Ozawa" were decomposed into relatively concrete questions such as those asking for his birthday and birth place.

Table 1: Categorization of questions by subject

Asking about	
4W (Who, When, Where, What) incl. several types of numerical values	70%
Why	4%
How, for a procedure or method	10%
Definitions, descriptions or explanations	16%

Table 2: Categorization of questions by answer type

Answered in	
Numerical values or date expressions	28%
Proper names	22%
Common names (in compound nouns)	8%
Names probably	14%
Clauses, sentences, or texts	28%

However, not all questions that were categorized as 4W questions could be answered by names. For example, whereas questions asking where, such as "Where was Shakespeare born?", could be answered by a place name, questions like "Where do lobsters like to live?" need a description and not a proper name as the answer. Table 2 shows the result of categorization according to this aspect. This categorization was conducted by inspecting questions only, and some of the questions were hard to determine decisively whether those could be answered by names or not, and so were categorized as "Names probably". For example, the question "Where does the name 'AIBO' come from?" could be answered by name if AIBO is an acronym, but there may be a long story as to its origin. Although such cases happened in other combinations of categories, those questions were categorized into a more complex category as only the border of names and descriptions are important in the current analysis.

As Table 2 shows, 58% to 72% of questions could be answered by names. The amount of those questions is almost same as the amount of 4W questions, since while some 4W questions could not be answered by names, some definition and explanation questions might be able to be answered by names. The fact that 58% to 72% of questions for writing reports could be answered by names demonstrates that question answering systems that answer these questions are useful in such situations.

In addition, the answers to 84% of those 72% questions could be found by humans from newspaper articles. This

indicates that the setting is realistic where users write reports through interacting with a question answering system that uses newspaper articles as its data source.

3.3 Constructing a test set

Using the questions collected, we constructed a test set as follows. We selected 26 from 40 topics, and chose appropriate questions and rearranged them for constructing gathering type series. Some of the questions were edited in order to resolve semantic or pragmatic ambiguities, though we tried to use the questions without modification where possible. The topics of the gathering series consisted of 5 persons, 2 organizations, 11 events, 5 artifacts, and 3 animals and fishes, among which 4 topics concerned sets of organizations and events, such as the big three companies in the beer industry, simultaneous terrorist attacks, and annual festival events.

Browsing type series were constructed by using some of the remaining questions as seeds of a sequence and by adding new questions to create a flow to/from those questions. For example, series 22 shown in Figure 1 was composed by adding the last four newly created questions to the first four questions which were collected for the Yankee stadium[2]. For such seeds, we also used the collection of questions for evaluating summarization constructed for TSC (Text Summarization Challenge), another challenge in the NTCIR workshop (TSC, 2003). Some topics used for the question collection were the same as the topics used in TSC also. We made 10 browsing series in this way.

Finally, the test set constructed this time contained 36 series and 251 questions, with 26 series of the gathering type and 10 series of the browsing type. The average number of questions in one series was 6.92.

4 Characteristics of the Test Set

This chapter describes the pragmatic characteristics of the constructed test set. Japanese has four major types of anaphoric devices: pronouns, zero pronouns, definite noun phrases, and ellipses. Zero pronouns are very common in Japanese in which pronouns are not realized on the surface. As Japanese also has a completely different determiner system from English, the difference between definite and indefinite is not apparent on the surface, and definite noun phrases usually have the same form as generic noun phrases. Table 3 shows the summary of such pragmatic phenomena observed in 215 questions obtained by removing the first one of each series from the 251 questions in the test set. The total number is more than 215 as 12 questions contain more than one phenomenon. The sixth question in series 22, "Who was the bride at that time?" is an example of such a question with multiple anaphoric expressions. The numbers in parentheses show the number of cases in which the referenced item is an event. As the table indicates, a wide range of pragmatic phenomena is observed in the test set.

Table 3: Pragmatic phenomena observed in the test set

Type	Occurence
Pronouns	76 (21)
Zero pronouns	134 (33)
Definite noun phrases	11 (4)
Ellipses	7

Precisely speaking, the series in the test set can be characterized through the pragmatic phenomena that they contain. *Gathering type* series consist of questions that have a common referent in a broad sense, which is a global topic mentioned in the first question of the series. *Strictly gathering type* series can be distinguished as a special case of gathering type series. In those series, all questions refer exactly to the same item mentioned in the first question and do not have any other anaphoric expression. In other words, questions about the common topic introduced by the first question comprise a whole sequence. Series 14 in Figure 1 is an example of the strictly gathering type and all questions can be interpreted by supplying Seiji Ozawa, who is introduced in the first question. The test set has 5 series of the strictly gathering type. Other gathering type series have other two types of questions. The first type of questions not only has a reference to the global topic but also refers to other items or has an ellipsis. The second type of questions has a reference to a complex item, such as an event that contains the global topic as its component. Series 20 shown in Figure 2 is such a series. The third question refers not only to the global topic, George Mallory, in this case, but also to his famous phrase. The sixth one refers to an event George Mallory was concerned in.

On the other hand, the questions of a *browsing type* series do not have such a global topic. Sometimes the referent is the answer of the immediately preceding question, such as the fifth, seventh and eighth questions in series 22 in Figure 1. No series, however, consists solely of questions that have only a reference to the answer to the immediately previous questions. All series contain references to the answers to non-immediately previous questions or items mentioned in the previous questions, or more than one pragmatic phenomenon. In series 22, the third, fourth and sixth questions belong to such a case.

In both types, therefore, the shifting pattern of the focus is not simple, and so a sophisticated way is needed to track it. Such focus tracking is indispensable to get correct answers. Systems cannot even retrieve articles con-

[2]The question focus of the first one was changed.

Series 20
In which country was George Mallory born?
What was his famous phrase?
When did he say it?
How old was he when he started climbing mountains?
On which expedition did he go missing near the top of Everest?
When did it happen?
At what altitude on Everest was he seen last?
Who found his body?

Figure 2: Another example of series of questions

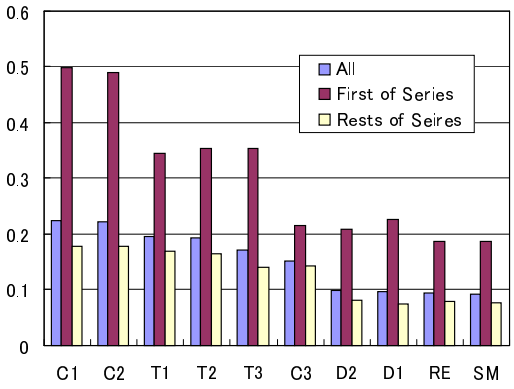

Figure 3: Evaluation by mean F measure

taining the answer just by accumulating keywords. This is clear for the browsing type, as an article is unlikely to mention both the New York Yankees and Campbell soup. In the gathering type, since the topics mentioned in relatively many articles were chosen, it is not easy to locate the answer to a question from those articles retrieved using that topic as the keyword. For example, there are 155 articles mentioning Seiji Ozawa in our document sets, of which 22 articles mention his move to the Vienna Philharmonic Orchestra, and only two articles also mention his birthday. An extensive, quantitative analysis is now in progress.

5 Difficulty of the Challenge and the Current State of Technologies

Seven teams and fourteen systems participated in the run using the test set mentioned in the previous chapter conducted in December 2003. In this chapter, based on a preliminary analysis of the run, the difficulty of the challenge and the current state of technologies for addressing the challenge are discussed. The techniques employed in the participant systems have not yet been published, but will be published by the NTCIR workshop 4 meeting at the latest.

Figure 3 shows the mean modified F measures of the top 10 participant systems. The chart shows the mean modified F measure of three categories: all of the test set questions, the questions of the first of each series, and questions of the second and after. As anticipated, it is more difficult to answer correctly the questions other than the first question of each series. This indicates that more sophisticated context processing is needed.

The mean modified F measure is not high even for the top systems. This is probably because of not only the difficulties of context processing but also the difficulties of returning the list of all and only correct answers. It is difficult to achieve high recall since some of the questions have many correct answers, such as asking for all of the countries and regions which participated in the 1998 football world cup held in France. The modified F measure is only 0.33 if a system returns a list of five items including the only correct answer, as the precision is 0.2 in that case. In order to remove the effects of such difficulties on answering lists approximately, the number of questions to which the system gives at least one of the correct answers was calculated. The result is shown in Figure 4. The rank of the systems somewhat changes by this approximation, as some systems benefit from this approximation and others do not. Based on this criterion, the best system answered correctly 45% of the questions, which is inadequate for practical use. However, the result shows that this challenge is not too hard and desperate, though it is challenging for existing question answering technologies.

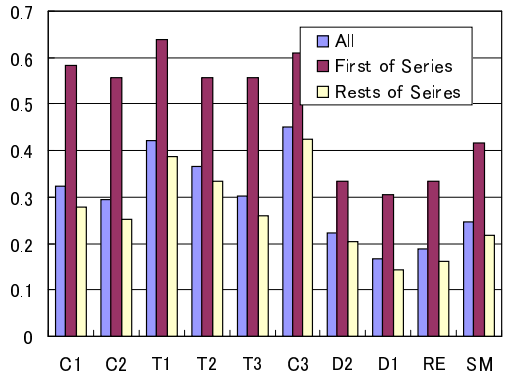

Figure 4: Another evaluation

The mean modified F measures for the strictly gathering type, other gathering type, and browsing type are shown in Figure 5. For the majority, the questions in the

browsing type series are more difficult to answer, as anticipated.

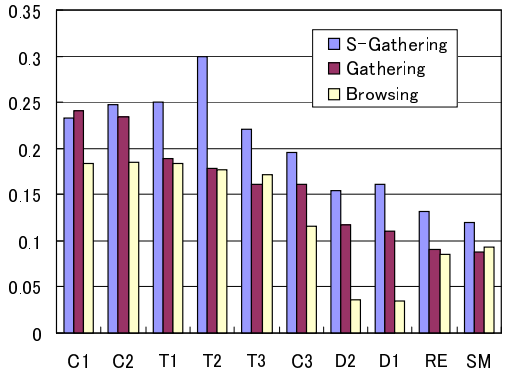

Figure 5: Differences on series types

6 Discussion

With existing technologies, which still have room for study for answering ordinary questions without pragmatic processing and particularly remain inadequate for answering list questions correctly, QACIAD cannot easily independently evaluate the context processing abilities from other general abilities concerning question answering. The ability that QACIAD measures is a combination of several kinds of abilities concerning question answering for handling information access dialogues. Although this may be desirable and an objective of QACIAD, sometime we need an isolated evaluation of context processing. In order to fulfill this need, we devised two types of accompanying test sets for reference. The first reference test set consists of isolated questions, that is, not in series, obtained from questions of the original test set by manually resolving all anaphoric expressions including zero anaphora. The second reference test set consists of isolated questions obtained from questions of the original test set by mechanically removing anaphoric expressions. Though most of the questions in the second test set are semantically under-specified, such as asking a birthday without specifying whose one, all the questions are syntactically well formed in the case of Japanese. The first reference test set measures the ceiling of the context processing in a given original test set, while the second measures the floor. These are only for reference, since there are several ways of resolving anaphora and context processing sometimes makes thing worse. Nevertheless, the reference test sets should be useful for analyzing the characteristics of technologies used by the participant systems. We are now analyzing the results of the run on those reference test sets for our current test set, and will present the results in due course.

As described above, we believe that the task setting in QACIAD is real, even though this is not clear from the evaluation method. There are two major problems. The first concerns the F measure. First, the F measure cannot be calculated until the number of correct answers is fixed, which means the value of the F measure changes when a new correct answer is found. This makes the evaluation cumbersome. Especially in question answering, as the number of correct answers is usually relatively small, the recall rate sometimes falls to half if a minor alternative answer is found to a question that had been assumed to have only one correct answer. Even worse, some questions have more than one way of enumerating correct answers. For example, to a question asking for the sites of a ski jump competition, a system may answer six city names, and another system may answer three country names. Neither are wrong. A system could even answer four city names and one country name. We need a principle for handling such cases. In TREC-2003 this problem were cleverly avoided by carefully checking the question[3].

The second and more serious problem comes from handling dialogues. As mentioned above, whether an answer is correct or not is determined by human interpretation of a given question within the given context and is not affected by a system's interpretation and the answers it returned to the previous questions. Many feel that this evaluation criterion is somewhat peculiar. As mentioned in the example in chapter 2.5, in series 22, the answer to the second question, 1923, is considered correct even if the system wrongly answered the Shea stadium to the first question. This is not completely absurd because that system may manage the context intensionally, in which case the system may interpret the second question as "When was the home to the New York Yankees built?" It is doubtful, however, whether such a "correct" answer has any value in practice. This problem shows the importance of cooperative response. It may be effective to change the style of answering from a current list of answers to answers with additional information. In this example, it would be better to answer "The Yankee stadium was built in 1923", and the correctness of answers should be judged by including this additional information. The difficult and remaining problem is to formalize this type of cooperative response to a sufficient level for use in objective evaluations like QACIAD.

7 Conclusion

A novel challenge, QACIAD (Question Answering Challenge for Information Access Dialogues), was proposed for evaluating the abilities for handling information access dialogues through open-domain question answer-

[3]Personal communication with Dr. Ellen Voorhees.

ing technologies. Question answering systems with such abilities measured by this challenge are expected to be useful for making reports and summaries. The proposed challenge has reasonable difficulties with existing technologies. Our proposal also has several important ideas, including the distinction of series of questions into gathering type and browsing type series, and the introduction of reference test sets for extracting and evaluating the context processing abilities of the systems.

Acknowledgments

The authors would like to thank all participants in NTCIR4 Workshop QAC2 task for their valuable comments on the task design and intensive works for addressing the task. This research was supported in part by the joint research grant of National Institute of Informatics.

References

Eduard Hovy. 2001. http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/isi_hovy_duc.pdf.

John Burger, Claire Cardie, and et al. 2001. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) http://www-nlpir.nist.gov/projrcts/duc/roadmpping.html.

NTCIR3 Workshop publication Home Page. 2001. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html.

NTCIR4 Workshop Home Page. 2003. http://research.nii.ac.jp/ntcir/workshop/work-en.html.

Jun'ichi Fukumoto and Tsuneaki Kato. 2001. An Overview of Question and Answering Challenge (QAC) of the next NTCIR Workshop. *The Second NTCIR Workshop Meeting*.

Jun'ichi Fukumoto, Tsuneaki Kato and Fumito Masui. 2002. Question Answering Challenge(QAC-1) Question answering evaluation at NTCIR workshop 3 *NTCIR workshop 3 Meeting Overview*, pp. 77 - 86.

Jun'ichi Fukumoto, Tsuneaki Kato and Fumito Masui. 2003. Question Answering Challenge(QAC-1) An Evaluation of question answering tasks at the NTCIR workshop 3 *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 122-133.

Inderjeet Mani, David House, and et al. 1998. The TIPSER SUMMAC text summarization evaluation final report. Technical Report MTR98W0000138, The MITRE Corporation.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a Question Answering Test Collection *the Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200 - 207.

TREC Home Page. 2003. http://trec.nist.gov/.

Text Summarization Challenge Home Page. 2003. http://lr-www.pi.titech.ac.jp/tsc/index-en.html.