Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3 –

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

Jun'ichi Fukumoto
Ritsumeikan University
1-1-1 Nojihigashi, Kusatsu-shi, Shiga 525-8577, Japan
fukumoto@media.ritsumei.ac.jp

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

Abstract

We describe an overview of Question Answering Challenge (QAC) 2 Subtask 3, a novel challenge for evaluating open-domain question answering technologies, at the NTCIR Workshop 4. In QAC2 Subtask 3, 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. Reporting the results of the run conducted and techniques employed there, we then show that existing technologies have the potential to address this challenge.

Keywords: *Question Answering, Context Processing, Evaluation.*

1 Introduction

Open-domain question answering (QA) 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 [16]. Although there are some notable exceptions [12, 2, 6], the recent research on open-domain question answering concentrates on answering factoid questions one by one in isolation from each other. This type of study has been encouraged and guided by a series of TREC conferences [15].

Such systems that answer isolated factoid questions are the most basic level of QA 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 young 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 QA system [3]. Even in daily situations, questions are rarely asked in isolation, but rather in a cohesive manner that involves a sequence of related questions to meet the person's information needs.

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 [8]. 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 [10]. Those studies suggested that QA systems that can answer a series of related questions would surely be a useful aid to summarization work by humans and by machines.

Against this background, QA systems need to be able to answer a series of questions. In this paper, we describes QAC2 Subtask 3, a challenge to measure objectively and quantitatively such an ability of QA systems[1]. In Subtask 3, QA systems are used interactively

[1] QAC2 Subtask 3 is named Question Answering Challenge for

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.

2 Design of QAC2 Subtask 3

In this chapter, we explain the design of QAC2 Subtask 3, which is a challenge to measure objectively and quantitatively such abilities of QA systems that can address information access dialogues.

QA systems need a wide range of abilities in order to participate in information access dialogues [3]. 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 explicitly 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, focusing on the most fundamental aspect of dialogue, that is, interpreting a given question within the context of a specific dialogue, QAC2 Subtask 3 measures the context processing abilities of systems such as anaphora resolution and ellipsis handling.

In this challenge, QA systems are requested to answer series of related questions. This series of questions and the answers to those questions comprise an information access dialogue. Although systems are supposed to participate in dialogue interactively, the interaction is only simulated; systems answer a series of questions in a batch mode. Such a simulation may neglect the inherent dynamics of dialogue, as the evolution of dialogues is fixed beforehand and systems have no opportunity to control it. It is, however, a practical compromise for objective evaluation. Since all participants have to answer the same set of questions in the same context, the results for the same test set are comparable with each other, and the test sets of the challenge can be made reusable by pooling correct answers.

The origin of QAC2 Subtask 3 comes from QAC1, one of the tasks of the NTCIR3 workshop [4, 11]. The current design reported in this paper is its extensive elaboration.

2.1 QAC2 as a common ground

QAC2 is a challenge for evaluating QA technologies in Japanese. It consists of three subtasks, 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. Using those documents as the data source, the systems answer various open-domain questions.

From the outset, QAC has focused on QA technologies that can be used as components of larger intelligent systems and technologies that can handle realistic problems. It requests 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 Subtask 2 and 3 request systems to return one list of answers that contains all and only correct answers, while Subtask 1 requests systems to return a ranked list of possible answers as in TREC-8. In all subtasks, the presence of answers in the underlying documents is not guaranteed and the number of answers is not specified.

2.2 Information access dialogue

Considering scenes in which those QA systems participate in a dialogue, we classified information access dialogues into the following two categories.

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.

Subtask 3 was designed to measure the abilities of QA systems useful in both types of dialogue.

2.3 Characteristics of question series

Subtask 3 requests participant systems to return answers to a series of questions. This series of questions

Information Access Dialogue (QACIAD) in some papers (in [9] for example), for emphasizing its characterisitc.

and the answers to those questions comprise an information access dialogue. Three examples of the series of questions are shown in Figure 1, which were picked from our test set discussed in section 3. Series 14 and 20 are of the gathering type, while series 22 is a typical browsing type.

Precisely speaking, the series in the test set can be characterized through the pragmatic phenomena 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. *Other gathering type* series have two other 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 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 the event George Mallory was involved in.

On the other hand, the questions of a *browsing type* series do not have such a global topic. In some cases the referent is an item mentioned in previous questions, and in other cases it refers to the answer of the preceding question. The former is the case in the third and fourth questions in series 22. The latter is the case in the fifth, seventh and eighth.

In both series, all questions except the first one of each series have some anaphoric expressions, which may be zero pronouns. That is, anaphoric expressions are used eagerly when those are possible and no subdialogue appears in any series.

In Subtask 3, several series are given to the system at once and the systems are requested to answer those series in a batch mode. 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. However, the systems must not look ahead to the questions following the one currently being handled. This restriction reflects the fact that Subtask 3 is a simulation of interactive use of QA 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.

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 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?

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

2.4 Evaluation measure

The judgment as to whether a given answer is correct or not takes into account not only the answer itself but also the accompanying article from which the answer was extracted. If the article does not validly support the answer, that is, assessors cannot understand whether 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 sys-

tem'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 to 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. Although we know it is somewhat counterintuitive in its extreme case, it is a compromise to make the evaluation practicable to perform.

In Subtask 3, as the systems are requested to return one list consisting of all and only correct answers and the number of correct answers differs for each question, a modified *F* measure is used for the primary 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 other duplication as a wrong one. So, the precision of such an 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, the modified F measure is 1.0 if a system returns an empty list as the answer, and is 0.0 otherwise. The primary evaluation measure of this challenge is *MMF*: the mean of the modified *F* measure over all questions in a test set.

3 Constructing the Test Set

Questions for the test set were collected as follows. Subjects were presented various topics, which included persons, organizations, and events, and were requested to make questions in Japanese to elicit information for a report on that topic. The report was supposed to describe facts on a given topic, rather than contain opinions or hypotheses on the topic. The questions were restricted to wh-type questions, and a natural series of questions containing anaphoric expressions and so on were constructed.

As we were interested in the relationship between the amount of knowledge on a given topic and questions asked, 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 subjects were instructed to make questions without considering whether the answer was contained in the given articles. That is, the information given was used only to understand the topic, and then the subjects made questions to elicit the information required for their reports.

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 a series of questions including around ten questions for each.

Those questions were natural in both content and expression since in the 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.

Using the questions collected, we constructed a test set. We selected 26 from 60 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. We made each series to have around seven questions. The topics of the gathering series consisted of 5 persons, 2 organizations, 11 events, 5 artifacts, and 3 animals and fishes.

Browsing type series were constructed by using some of the remaining questions and other question collection 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. 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 [14]. 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 (5 series of the strictly gathering type among them) and 10 series of the browsing type. The average number of questions in one series was 6.92.

Table 1 shows the summary of observed pragmatic phenomena. 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

Table 1. Pragmatic phenomena observed in the test set

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

same form as generic noun phrases. Table 1 shows the occurences of such pragmatic phenomena 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. As the table indicates, a wide range of pragmatic phenomena is observed in the test set.

The numbers in parentheses show the number of cases in which the referenced item is an event. Since questions are often handled as a sequence of keywords in the current techniques of QA, it is anticipated that reference expressions referring to events that cannot be represented by one keyword are more difficult to handle than those referring to persons or organizations. This table shows our test set contains a lot of such difficult reference expressions.

Sophisticated focus tracking is indispensable to get correct answers from this test set. Systems cannot even retrieve articles containing 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 mention his move to the Vienna Philharmonic Orchestra, and only 2 also mention his birthday.

3.1 Reference set

The ability that QAC2 Subtask 3 measures is a combination of several kinds of abilities concerning question answering for handling information access dialogues. Although this may be desirable and one of the objectives, occasionally we need an isolated evaluation of context processing. This isolation cannot be achieved by introducing any evaluation measure. 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 birthday, 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.

4 The Results of the Run and the Techniques employed

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 role of the reference sets are discussed. The techniques for addressing the challenge are also examined.

4.1 Overview of the results

Figure 2 shows the *MMF* of the participant systems. The chart shows the *MMF* 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 performances shown here are not high even for the top systems, which are inadequate for practical use. However, this result shows that this challenge is not too hard, though it is challenging for existing QA technologies.

Figure 3 shows the difference of the performance according to the type of series: the *MMF* for the strictly gathering type, other gathering type, and browsing type. For the majority, the questions in the browsing type series are more difficult to answer, as anticipated.

Figure 4 is an example of the information obtained using the reference set. This chart is a histogram of the difference of average modified *F* measure over all participants between a question in the test set and its corresponding question in the first reference set, and

reflects the difficulty of context processing of the questions. For each of the three types of series, the numbers of questions of the second and later are depicted. We can see that many of the questions with a large difference come from the browsing type series, supporting the finding that the browsing type series are more difficult to handle.

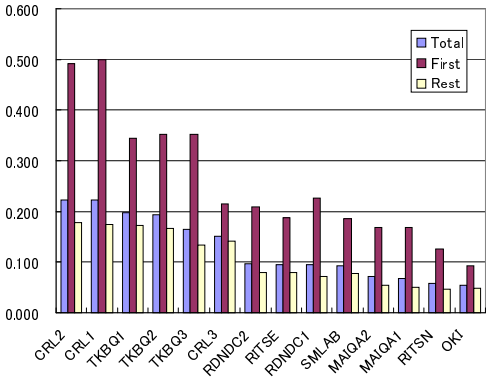

Figure 2. Evaluation by *MMF*

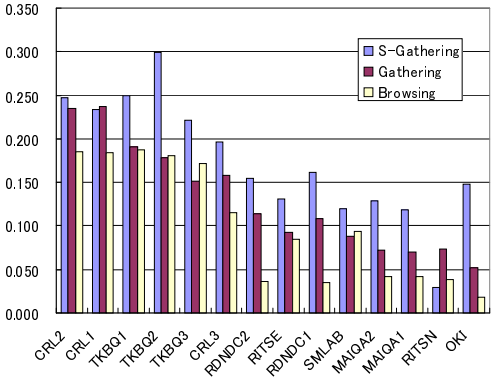

Figure 3. Differences on series types

The second reference set, accompanied with the first set, discloses whether a given question really needs context processing in order to answer it correctly. For example, considering the question "What is the name of the hybrid car that the company released in 1997?" following "What automobile company was Toyoda Shoichiro chairman of?" (Series 5), we found that the evaluation of its corresponding questions in the first and second reference set had almost the same value (0.58 and 0.56 in average *MF*), which was much higher than the evaluation of the original question (0.02). This shows that context processing is not needed for this question; rather, it makes the situation worse. It seems strange that the systems can answer the question correctly without restriction of the company, but in fact, as only one hybrid car was released in 1997, the answer is the same as the answer

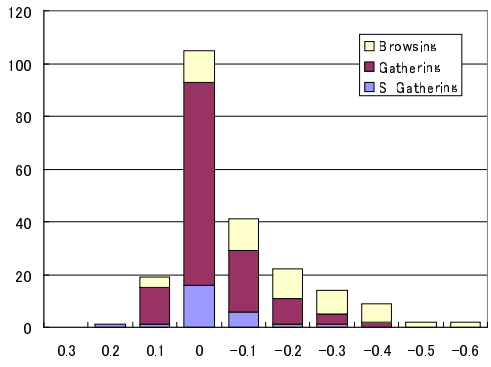

Figure 4. Difficulties of context processing

to the question "What is the name of the hybrid car released in 1997?" The question "At what altitude on Everest was he seen last?" is another example, since no one except George Mallory was seen on Everest. It is difficult to estimate for each question in advance how much dispensable context processing exists in order to obtain the correct answer, since it relates to the content of the underlying data source. Our reference sets reveal such information clearly.

4.2 Techniques employed

As far as is known from the participants' reports, techniques employed for context processing for the run are rather simple. These techniques provide a basis for further development.

In the most prevailing case, systems do not analyze referential expressions in a given question at all, but simply treat that question as a continuation of preceding questions [1, 7, 13]. Systems that use keywords extracted at the question analysis stage in the subsequent stages take keywords in the preceding questions into consideration in addition to those in the current one. In one system which uses higher order characteristics, namely word bi-grams, a concatenation of the preceding and current questions is regarded as a question to be processed. Such systems differ in which questions are considered as the preceding ones. Some use only the first of the series while others use the whole of the series up to the current one. Another consideration is the balance of weights of keywords in those questions. Our system adds the answers to the preceding question into the keywords of the current question.

Another system employs a more sophisticated way of handling context in its question analysis stage [5]. It determines referents using a shallow syntactic-semantic analysis of questions. An antecedent question is analyzed and decomposed into the entity description, attribute description and interrogative ex-

pression. For example, the entity description, attribute description and interrogative expression of the question "Who is the president of the United States?" are "the United States", "the president" and "who", respectively. When the current question has no apparent reference expression, that is, when it has zero anaphora or is an elliptical fragment, the similarity of interrogative expressions of the antecedent and the current one is used as a clue to determine whether the entity or the attribute should be supplied as the referent. The difference between "Who is ϕ of France?" and "How large is ϕ land area?" following the above question is handled in this manner. When the question contains a reference expression, its semantic category is used for that decision. As one series consists of several questions, there is ambiguity as to which of them could be the antecedent, which is resolved using a heuristics.

QA systems could handle context in modules other than question analysis. A system determines the documents from which the answers are extracted while processing the first question of a series and uses them exclusively while processing the whole of the series [7]. Although this technique seems rather crude, it is unique for question answering and its refinement may lead to a novel technique.

Each technique mentioned in this section has its advantages and disadvantages derived from its intrinsic characteristics. For example, document restriction by the first question cannot work properly for the browsing series. In our experience of the run, however, it is not clear that such a relationship exists between techniques employed and evaluation results of the run. This is probably because system performances depend on several factors including robustness against noise such as existence of spurious keywords.

5 Conclusion

A novel challenge, QAC2 Subtask 3, was proposed for evaluating the abilities for handling information access dialogues through open-domain QA technologies. QA systems with such abilities measured by this challenge are expected to be useful for making reports and summaries. 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. Many techniques have proposed for addressing this challenge, which make difficulties of the challenge reasonable. We, nevertheless, believe QA technologies still have ample room to develop and accomplish a better result on the challenge.

References

[1] Tomoyoshi Akiba, Katunobu Itou and Atsushi Fujii. 2004. Question Answering using "Common Sense" and Utility Maximization Principle. *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.297 - 303. *Revised and reprinted in this proceedings*.

[2] Joyce Y. Chai and Rong Jin. 2004. Discource Structure for Context Question Answering. *Proceedings of HLT-NAACL2004 Workshop on Pragmatics of Question Answering*, pp. 23-30.

[3] John Burger, Claire Cardie, 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`.

[4] 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.

[5] Jun'ichi Fukumoto, Tatsuhiro Niwa, Makoto Itoigawa and Megumi Matuda. 2004. Rits-QA: List answer detection and Context task with ellipses handling. *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.310 - 314. *Revised and reprinted in this proceedings.*

[6] Sanda Harabagiu, Dan Moldovan, Marius Paşca, et al. 2001. Answering complex, list and context questions with LCC's Question-Answering Server. *Proceedings of TREC 2001*.

[7] Naoya Hidaka, Fumito Masui and Keiko Tosaki. 2004. MAIQA: Mie Univ. Participated System at NTCIR4 QAC2. *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.315 - 319. *Revised and reprinted in this proceedings.*

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

[9] Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui and Noriko Kando. 2004. Handling Information Access Dialogue through QA Technologies – A novel challenge for open-domain question answering –. *Proceedings of HLT-NAACL2004 Workshop on Pragmatics of Question Answering*, pp. 70-77.

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

[11] NTCIR (NII-NACSIS Test Collection for IR Systems) Project Home Page. 2003. `http://research.nii.ac.jp/ntcir/index-en.html`.

[12] Sharon Small, Nobuyuki Shimizu, et al. 2003. HITIQA: A Data Driven Approach to Interactive Question Answering: A Preliminary Report. *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 94-104.

[13] Toru Takaki. 2004. NTT DATA Question-Answering Experiment at the NTCIR-4 QAC2. *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.402 - 405. *Revised and reprinted in this proceedings*.

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

[15] TREC Home Page. 2003. `http://trec.nist.gov/`.

[16] 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.