WoZ Simulation of Interactive 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

QACIAD (Question Answering Challenge for Information Access Dialogue) is an evaluation framework for measuring interactive question answering (QA) technologies. It assumes that users interactively collect information using a QA system for writing a report on a given topic and evaluates, among other things, the capabilities needed under such circumstances. This paper reports an experiment for examining the assumptions made by QACIAD. In this experiment, dialogues under the situation that QACIAD assumes are collected using WoZ (Wizard of Oz) simulating, which is frequently used for collecting dialogue data for designing speech dialogue systems, and then analyzed. The results indicate that the setting of QACIAD is real and appropriate and that one of the important capabilities for future interactive QA systems is providing cooperative and helpful responses.

1 Introduction

Open-domain question answering (QA) technologies allow users to ask a question using natural language and obtain the answer itself rather than a list of documents that contain the answer (Voorhees et al.2000). While early research in this field concentrated on answering factoid questions one by one in an isolated manner, recent research appears to be moving in several new directions. Using QA systems in an interactive environment is one of those directions. A context task was attempted in order to evaluate the systems' ability to track context for supporting interactive user sessions at TREC 2001 (Voorhees 2001). Since TREC 2004, questions in the task have been given as collections of questions related to common topics, rather than ones that are isolated and independent of each other (Voorhees 2004). It is important for researchers to recognize that such a cohesive manner is natural in QA, although the task itself is not intended for evaluating context processing abilities since, as it is given the common topic, sophisticated context processing is not needed.

Such a direction has also been envisaged as a research roadmap, in which QA systems become more sophisticated and can be used by professional reporters and information analysts (Burger et al.2001). At some stage of that sophistication, a young reporter writing an article on a specific topic will be able to translate the main issue into a set of simpler questions and pose those questions to the QA system.

Another research trend in interactive QA has been observed in several projects that are part of the ARDA AQUAINT program. These studies concern scenario-based QA, the aim of which is to handle non-factoid, explanatory, analytical questions posed by users with extensive background knowledge. Issues include managing clarification dialogues in order to disambiguate users' intentions and interests; and question decomposition to obtain simpler and more tractable questions (Small et al.2003)(Hickl et

Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006, pages 9–16,
New York City, NY, USA. June 2006. ©2006 Association for Computational Linguistics

al.2004).

The nature of questions posed by users and patterns of interaction vary depending on the users who use a QA system and on the environments in which it is used (Liddy 2002). The user may be a young reporter, a trained analyst, or a common man without special training. Questions can be answered by simple names and facts, such as those handled in early TREC conferences (Chai et al.2004), or by short passages retrieved like some systems developed in the AQUAINT program do (Small et al.2003). The situation in which QA systems are supposed to be used is an important factor of the system design and the evaluation must take such a factor into account. QACIAD (Question Answering Challenge for Information Access Dialogue) is an objective and quantitative evaluation framework to measure the abilities of QA systems used interactively to participate in dialogues for accessing information (Kato et al.2004a)(Kato et al.2006). It assumes the situation in which users interactively collect information using a QA system for writing a report on a given topic and evaluates, among other things, the capabilities needed under such circumstances, i.e. proper interpretation of questions under a given dialogue context; in other words, context processing capabilities such as anaphora resolution and ellipses handling.

We are interested in examining the assumptions made by QACIAD, and conducted an experiment, in which the dialogues under the situation QACIAD assumes were simulated using the WoZ (Wizard of Oz) technique (Fraser et al.1991) and analyzed. In WoZ simulation, which is frequently used for collecting dialogue data for designing speech dialogue systems, dialogues that become possible when a system has been developed are simulated by a human, a WoZ, who plays the role of the system, as well as a subject who is not informed that a human is behaving as the system and plays the role of its user. Analyzing the characteristics of language expressions and pragmatic devices used by users, we confirm whether QACIAD is a proper framework for evaluating QA systems used in the situation it assumes. We also examine what functions will be needed for such QA systems by analyzing intelligent behavior of the WoZs.

2 QACIAD and the previous study

QACIAD was proposed by Kato et al. as a task of QAC, which is a series of challenges for evaluating QA technologies in Japanese (Kato et al.2004b). QAC covers factoid questions in the form of complete sentences with interrogative pronouns. Any answers to those questions should be names. Here, “names” means not only names of proper items including date expressions and monetary values (called “named entities”), but also common names such as those of species and 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 newspaper articles. Being given various open-domain questions, systems are requested to extract exact answers rather than text snippets that contain the answers, and to return the answer along with the newspaper article from which it was extracted. The article should guarantee the legitimacy of the answer to a given question.

In QACIAD, which assumes interactive use of QA systems, systems are requested to answer series of related questions. The series of questions and the answers to those questions comprise an information access dialogue. All questions except the first one of each series have some anaphoric expressions, which may be zero pronouns, while each question is in the range of those handled in QAC. Although the systems are supposed to participate in dialogue interactively, the interaction is only simulated; systems answer a series of questions in batch mode. Such a simulation may neglect the inherent dynamics of dialogue, as the dialogue evolution is fixed beforehand and therefore not something that the systems can control. It is, however, a practical compromise for an objective evaluation. Since all participants must 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 task are reusable by pooling the correct answers.

Systems are requested to return one list consisting of all and only correct answers. Since the number of correct answers differs for each question and is not given, a modified F measure is used for the evaluation, which takes into account both precision and recall.

Two types of series were included in the QACIAD, which correspond to two extremes of information access dialogue: a gathering type in which 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 related to that topic; and a browsing type in which the user does not have any fixed topic of interest. Although the QACIAD assumes that users are interactively collecting information on a given topic and the gathering-type dialogue mainly occurs under such circumstances, browsing-type series are included in the task based on the observation that even when focusing on information access dialogue for writing reports, the systems must handle focus shifts appearing in browsing-type series. The systems must identify the type of series, as it is not given, although they need not identify changes of series, as the boundary is given. The systems must not look ahead to questions following the one currently being handled. This restriction reflects the fact that the QACIAD is a simulation of interactive use of QA systems in dialogues.

Examples of series of QACIAD are shown in Figure 1. The original questions are in Japanese and the figure shows their direct translations.

The evaluation of QA technologies based on QACIAD were conducted twice in QAC2 and QAC3, which are a part of the NTCIR-4 and NTCIR-5 workshops[1], respectively (Kato et al.2004b)(Kato et al.2005). It was one of the three tasks of QAC2 and the only task of QAC3. On each occasion, several novel techniques were proposed for interactive QA.

Kato et al. conducted an experiment for confirming the reality and appropriateness of QACIAD, in which subjects were presented various topics and were requested to write down series of questions in Japanese to elicit information for a report on that topic (Kato et al.2004a)(Kato et al.2006). The report was supposed to describe facts on a given topic, rather than state opinions or prospects on the topic. The questions were restricted to wh-type questions, and a natural series of questions that may contain anaphoric expressions and ellipses was constructed. Analysis of the question series collected in such a manner showed that 58% to 75% of questions for writing reports could be answered by values or names; a wide range of reference expressions is observed in questions in such a situation; and sequences of questions are sometimes very complicated and include subdialogues and focus shifts. From these observations they concluded the reality and appropriateness of the QACIAD, and validated the needs of browsing-type series in the task.

Series 30002
What genre does the "Harry Potter" series belong to?
Who is the author?
Who are the main characters in the series?
When was the first book published?
What was its title?
How many books had been published by 2001?
How many languages has it been translated into?
How many copies have been sold in Japan?

Series 30004
When did Asahi breweries Ltd. start selling their low-malt beer?
What is the brand name?
How much did it cost?
What brands of low-malt beer were already on the market at that time?
Which company had the largest share?
How much low-malt beer was sold compared to regular beer?
Which company made it originally?

Series 30024
Where was Universal Studio Japan constructed?
What is the nearest train station?
Which actor attended the ribbon-cutting ceremony on the opening day?
Which movie that he featured in was released in the New Year season of 2001?
What movie starring Kevin Costner was released in the same season?
What was the subject matter of that movie?
What role did Costner play in that movie?

Figure 1: Examples of Series in QACIAD

One of the objectives of our experiment is to confirm these results in a more realistic situation. The previous experiment setting is far from the actual situations in which QA systems are used, in which subjects have to write down their questions without getting the answers. Using WoZ simulation, it is confirmed whether or not this difference affected the result. Moreover, observing the behavior of WoZs, the capabilities and functions needed for QA sys-

[1] The NTCIR Workshop is a series of evaluation workshops designed to enhance research in information access technologies including information retrieval, QA, text summarization, extraction, and so on (NTCIR 2006).

tems used in such a situation are investigated.

3 Setting

Referring to the headlines in Mainichi and Yomiuri newspapers from 2000 and 2001, we selected 101 topics, which included events, persons, and organizations. On each of those topics, a *summary* of between 800 and 1600 characters long and an *abstract* of around 100 characters long were constructed using a full text search system on the newspaper articles.[2] Four experts shared this preparation work. Twenty topics were selected from among the original 101 on the basis that enough information was gathered and compiled into the summary.[3] The topics consisted of 5 persons, 2 organizations, 7 events, 5 artifacts, and 1 syndrome, which include Mr. Colin Powell, Charles, Prince of Wales, An accident of a Russian nuclear-powerd submarine, Concorde, “Harry Potter”, and Windows XP.

The four experts who shared the preparation work acted as a WoZ in the simulated QA systems, and each expert participated in dialogues on the topic for which she/he wrote the summary and abstract, and tried to answer questions from users using the summary, a full text search system, and her/his memory.

Six subjects participated in the experiment and 10 topics were assigned to each subject. That is, each topic was handled by three subjects. The subjects behaved as users of simulated QA systems. Presented with the topic and abstract, the subjects were directed to think about questions beforehand assuming that they had to write a report on that topic, and then they participated in an information access dialogue with the “QA system.” The suggested number of questions for each topic was around 10.

The dialogues were conducted through keyboards and character displays. The subjects were told that the system could answer only simple factoid questions, and the WoZs were instructed not to answer complicated questions such as asking for reasons and opinions. No restrictions on the usage of language expressions were imposed but were left to the discretion of the WoZs, though canned phrases such as “Please wait a moment” and “Sorry, the answer could not be found” were prepared in advance. The WoZs were also instructed that they could clarify users’ questions when they were ambiguous or vague, and that their answers should be simple but cooperative and helpful responses were not forbidden.

USR: What is the title of the first book of the “Harry Potter” series?
SYS: “Harry Potter and the Sorcerer’s Stone.”
USR: How many books were sold?
SYS: The Japanese translation sold 600 thousand copies in the first half year after its publication.
USR: Who is the author?
SYS: J. K. Rowling.
USR: Who are the main characters (in that series)?
SYS: Harry Potter, Ron Weasley, Hagrid and so on.
USR: What kind of person is Harry Potter?
SYS: A child with a wizard and a witch for parents.
USR: What is the plot?
SYS: It is an adventure story involving Harry Potter and his friends who enter a school for witchcraft and wizardry.
USR: How many people went to the movie?
SYS: In the US, box-office sales of the movie “Harry Potter and the Sorcerer’s Stone” reached 188 million dollars in the first 10 days after release.

Figure 2: Example of dialogues collected

An example of the dialogues collected is shown in Figure 2. In the figure, SYS stands for utterances of the QA system simulated by a WoZ and USR represents that of the user, namely a subject. In the rest of the paper, these are referred to as system’s utterances and user’s utterances, respectively.

4 Coding and Results

Excluding meta-utterances for dialogue control such as “Please wait a moment” and “That’s all,” 620 pairs of utterances were collected, of which 22 system utterances were for clarification. Among the remaining 598 cases, the system gave some answers in 502 cases, and the other 94 utterances were negative responses: 86 utterances said that the answer could not found; 10 utterances said that the question was too complicated or that they could not answer such type of question.

4.1 Characteristics of questions and answers

The syntactic classification of user utterances and its distribution is shown in Table 1. The numbers in

[2] Those lengths of Japanese texts roughly correspond to 300 to 600 words and 40 words of English text, respectively.

[3] We prepared 101 topics to meet the additional needs of other experiments. It is absolutely not the case that only 20 out of 101 were acceptable for our experiments.

Table 1: Syntactic classification of user utterances

Syntactic form	
Wh-type Question	87.7% (544)
Yes-no Question	9.5% (59)
Imperative (Information request)	2.6% (16)
Declarative (Answer to clarification)	0.2% (1)

Table 2: Categorization of user utterances by subject

Asking about	
Who, Where, What	32.5% (201)
When	16.3% (101)
How much/many (for several types of numerical values)	16.8% (104)
Why	6.5% (40)
How (for procedures or situations)	17.0% (105)
Definitions, Descriptions, Explanations	10.8% (67)
Other (Multiple Whs)	0.2% (1)

Table 3: Categorization of user utterances by answer type

Answered in	
Numerical values	14.3% (72)
Date expressions	16.7% (84)
Proper names	22.1% (111)
Common names	8.8% (44)
Compound nouns except names	4.2% (21)
Noun phrases	6.2% (31)
Clauses, sentences, or texts	27.7% (139)

Table 4: Pragmatic phenomena observed

Type	
No reference expression	203
Pronouns	14
Zero pronouns	317
Definite noun phrases	104
Ellipses	1

parentheses are numbers of occurrences. In spite of the direction of using wh-type questions, more than 10% of utterances are yes-no questions and imperatives for requesting information. Most of the user responses to clarification questions from the system are rephrasing of the question concerned; only one response has a declarative form. Examples of rephrasing will be shown in section 4.3.

The classification of user questions and requests according to the subject asked or requested is shown in Table 2; the classification of system answers according to their syntactic and semantic categorization is shown in Table 3. In Table 2, the classification of yes-no questions was estimated based on the information provided in the helpful responses to those. The classification in Table 3 was conducted based on the syntactic and semantic form of the exact part of the answer itself rather than on whole utterances of the system. For example, the categorization of the system utterance "He was born on April 5, 1935," which is the answer to "When was Mr. Colin Powell born?" is not a sentence but a date expression.

4.2 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 apparent 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 4 shows a summary of such pragmatic phenomena observed. The total number is more than 620 as some utterances contain more than one anaphoric expression. "How many crew members were in *the submarine* when *the accident* happened?" is an example of such a question with multiple anaphoric expressions.

Among 203 questions with no reference expression, 60 questions (30%) are the first one of a series of utterances on the same topic. The others are divided into two cases. In the first and major case, the current foci appear literally rather than in pronouns or definite noun phrases. The second case, which is relatively minor, is the beginning of and the resume from a subdialogue. For example, when a user himself resumed talking about Charles, Prince of Wales after a few questions about his children, the user used a literal expression like "Who is the long-term partner of *Charles, Prince of Wales*?"; after talking about the details of the accident, a user asked “What is the major characteristic of *Concorde*?”

Out of 437 reference expressions excluding ellipses in Table 4, 94 expressions (22%) refer to items other than the global topic. Moreover, 76 expressions of those (81%) are a sole reference expression in the utterance. The existence of questions without referring to the global topic, accompanied with the existence of some questions with no reference expression, means that the focus shifts in those dialogues, and then the local focus can differ from the global topic.

4.3 Clarifications

As mentioned above, only 22 clarification utterances were observed in the experiment. They are categorized into those asking about an ambiguous referent (13 utterances) and those asking about a vague question subject (9 utterances). Examples of each case are shown below.

Clarification asking about an ambiguous referent

USR: When was he [Mr. Colin Powell] installed as head of security of the United States?
SYS: December 6, 2000.
USR: Who was the president at that time?
SYS: Mr. George W. Bush.
USR: Which political party did he belong to?
SYS: *Whose political party are you asking about?*
USR: Which political party did Mr. Colin Powell belong to?

Clarification asking about a vague question subject

USR: Who developed Windows XP?
SYS: *Are you asking for a company name?*
USR: What company developed Windows XP?

4.4 Cooperative and helpful responses

Rather than simple and literal answers to the users' questions and requests, cooperative and helpful responses, in which, for example, information not requested directly by a user was additionally provided (Kaplan 1983), were observed frequently. The results of analyzing such cooperative and helpful behavior in the system responses are shown in Table 5. Questions asking reasons, procedures, situations, definitions, descriptions or explanations, which are categorized into the last three categories except others in Table 2, were excluded from this analysis, since for those questions it is difficult to judge objectively whether some information is additional or not and, in addition, most of those are out of the scope of QACIAD, which is our main interest. Table 5 shows the number of occurrences and its percentage among positive responses. Sometimes, more than one behavior was observed in one utterance. 148 utterances out of 334 positive responses (44%) contained some helpful behavior. The definitions and manually constructed examples of cooperative and helpful behavior listed in Table 5 are shown below.

Table 5: Cooperative and helpful behavior observed

Type of behavior	
Providing additional information	52 (15.6%)
Providing detailed information	38 (11.4%)
Elaboration of subject	38 (11.4%)
Providing alternative information	47 (14.1%)
Exemplification	28 (8.4%)
Providing speculations	3 (0.9%)

Providing additional information Information not requested directly from a user is additionally provided.

USR: What year was Mr. Colin Powell born?
SYS: He was born in 1935 in New York.

Providing detailed information More detailed information than that requested directly by a user is provided.

USR: What year was Mr. Colin Powell born?
SYS: He was born on April 5, 1935.

Elaboration of subject A relatively vague subject of a question is elaborated and made concrete. In the following example, a question with a vague expression, “how well” is answered by a value as a concrete statistic.

USR: How well did the movie do?
SYS: In the US, box-office sales reached 188 million dollars in the first 10 days after release.

Providing alternative information Rather than exact information a user requested, alternative related information is provided. An alternative may be proposed by relaxing the condition imposed by the user. In the following example, attendance figures were altered to box-office sales.

USR: How many people went to the movie?

SYS: In the US, box-office sales reached 188 million dollars in the first 10 days after release.

Exemplification Instead of a comprehensive list or abstracted general description, some major instances are listed.

USR: What kinds of character goods are on sale?

SYS: Stuffed owls, uniforms for schools of witchcraft and wizardry, and so on.

Providing speculation Rather than fact, speculations and opinions are provided, which can be considered a special case of providing alternative information. Although there are few cases of such behavior in Table 5, there were many such answers in response to questions asking for reasons and situations.

USR: What exploded?

SYS: It is surmised that torpedoes in the shed at the fore exploded.

Detecting and correcting erroneous user assumptions is a famous helpful behavior discussed in the literature (Pollack 1989) (Kaplan 1983). Only two occurrences of such behavior were observed, one of which is shown below.

USR: How long did it take from the explosion until the shipwreck?

SYS: The explosion happened while the submarine was going down, and ...

5 Discussion

The analysis of dialogue collected by the WoZ simulation showed very similar results to those obtained in a previous study, in which sequences of questions were written down by subjects without knowing the answers to questions. That is, as shown in Table 2, when users asked questions to get information for a report, the number of why-questions was relatively small. Moreover, there were fewer questions requesting an explanation or definition than expected, probably because definition questions such as “Who is Mr. Colin Powell?” were decomposed into relatively concrete questions such as those asking for his birthday and birthplace. The remainder (65%) could be answered in values and names. Table 3 indicates that 62% of the questions in our experiments were answered by values or names. If compound nouns describing events or situations, which are usually distinguished from names, are considered to be in the range of answers, the percentage of answerable questions reaches 68%. From these results, the setting of QACIAD looks realistic where users write reports interacting with a QA system handling factoid questions that have values and names as answers.

A wide range of reference expressions is observed in information access dialogues for writing reports. Moreover, our study confirmed that those sequences of questions were sometimes very complicated and included subdialogues and focus shifts. It is expected that using an interactive QA system that can manage those pragmatic phenomena will enable fluent information access dialogue for writing reports. In this sense, the objective of QACIAD is appropriate.

It could be concluded from these results that the reality and appropriateness of QACIAD was reconfirmed in a more realistic situation. And yet suspicion remains that even in our WoZ simulation, the subjects were not motivated appropriately, as suggested by the lack of dynamic dialogue development in the example shown in Figure 2. Especially, the users often gave up too easily when they did not obtain answers to prepared questions.[4] The truth, however, may be that in the environment of gathering information for writing reports, dynamic dialogue development is limited compared to the case when trained analysts use QA systems for problem solving. If so, research on this type of QA systems represents a proper milestone toward interactive QA systems in a broad sense.

Another finding of our experiment is the importance of cooperative and helpful responses. Nearly half of WoZ utterances were not simple literal responses but included some cooperative and helpful behavior. This situation contrasts with a relatively small number of clarification dialogues. The importance of this behavior, which was emphasized in research on dialogues systems in the 80s and 90s, was reconfirmed in the latest research, although question-answering technologies were redefined in the late 90s. Some behavior such as providing alternative information could be viewed as a second-best

[4] It is understandable, however, that there were few rephrasing attempts since users were informed that paraphrasing such as “What is the population of the US?” to “How many people are living in the US?” are usually in vain.

strategy of resource-bounded human WoZs. Even so, it is impossible to eliminate completely the need for such a strategy by improving core QA technologies. In addition, intrinsic cooperative and helpful behavior such as providing additional information was also often observed. These facts, accompanied by the fact that such dialogues are perceived as fluent and felicitous, suggest that the capability to behave cooperatively and helpfully is essential for interactive QA technologies.

6 Conclusion

Through WoZ simulation, the capabilities and functions needed for interactive QA systems used as a participant in information access dialogues for writing reports were examined. The results are compatible with those of previous research, and reconfirmed the reality and appropriateness of QACIAD. A new finding of our experiment is the importance of cooperative and helpful behavior of QA systems, which was frequently observed in utterances of the WoZs who simulated interactive QA systems. Designing such cooperative functions is indispensable. While this fact is well known in the context of past research on dialogue systems, it has been reconfirmed in the context of the latest interactive QA technologies.

References

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.

John Burger, Claire Cardie, Vinay Chaudhri, 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`.

Norma M. Fraser and G. Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, Vol 5, No.1, pp. 81-99.

Andrew Hickl, Johm Lehmann, John Williams, and Sanda Harabagiu. 2004. Experiments with Interactive Question Answering in Complex Scenarios. *Proceedings of HLT-NAACL2004 Workshop on Pragmatics of Question Answering*, pp. 60-69.

Joerrold Kaplan. 1983. Cooperative Responses from a Portable Natural Language Database Query System. Michael Brady and Robert C. Berwick eds. *Computational Models of Discourse*, pp. 167–208, The MIT Press.

Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui and Noriko Kando. 2004a. 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.

Tsuneaki Kato, Jun'ici Fukumoto and Fumito Masui. 2004b. Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3 –. *Proceedings of NTCIR-4 Workshop Meeting*.

Tsuneaki Kato, Jun'ici Fukumoto and Fumito Masui. 2005. An Overview of NTCIR-5 QAC3. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 361–372.

Tsuneaki Kato, Jun'ici Fukumoto, Fumito Masui and Noriko Kando. 2006. Are Open-domain Question Answering Technologies Useful for Information Access Dialogues? – An empirical study and a proposal of a novel challenge – *ACL Trans. of Asian Language Information Processing*, In Printing.

Elizabeth D. Liddy. 2002. Why are People Asking these Questions? : A Call for Bringing *Situation* into Question-Answering System Evaluation. *LREC Workshop Proceedings on Question Answering · Strategy and Resources*, pp. 5-8.

NTCIR Project Home Page. 2006. `http://research.nii.ac.jp/˜ntcadm/index-en.html`

Martha E. Pollack. 1989. Plans as Complex Mental Attitudes. Philip R. Cohen, Jerry Morgan and Martha E. Pollack eds. *Intentions in Communication*, pp. 77–103, The MIT Press.

Sharon Small, Nobuyuki Shimizu, Tomek Strzalkowski, and Liu Ting 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.

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.

Ellen M. Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. *Proceedings of TREC 2001*.

Ellen M. Voorhees. 2004. Overview of the TREC 2004 Question Answering Track. *Proceedings of TREC 2004*.