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

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

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 potential to address this challenge.

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 (Voorhees, 2000). Although there are some notable exceptions (Small et al., 2003), 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 (TREC, 2003).

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 (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 QA systems that can answer a series of questions would surely be a useful aid to summarization work.

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. In Subtask 3, QA 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.

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. Whereas QA systems need a wide range of abilities in order to participate in dialogues (Burger et al., 2001), QAC2 Subtask 3 focuses on the most fundamental aspect of dialogue, that is, interpreting a given question within the context of a specific dialogue. It measures the context processing abilities of systems such as anaphora resolution and ellipses handling. Although in this challenge, QA systems are supposed to participate in dialogue interactively, the interaction is only simulated; systems answer a series of questions in a batch mode, and so the test sets of the challenge are reusable.

The origin of QAC2 Subtask 3 comes from QAC1, one of the tasks of the NTCIR3 workshop (Fukumoto et al., 2003)(NTCIR, 2003). The current design of Subtask 3 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 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 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 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 chapter 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. 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 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. This is the case in the fifth, seventh and eighth questions in series 22.

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

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

definite and indefinite are not clearly distinguished, those problems are more serious.

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 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 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.

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 (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 (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 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. 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 1: Pragmatic phenomena observed in the test set

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

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 correspondent in the first reference set, and reflects the difficulty of context processing of the questions. The questions of the second and after of the three types of series are depicted. 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.

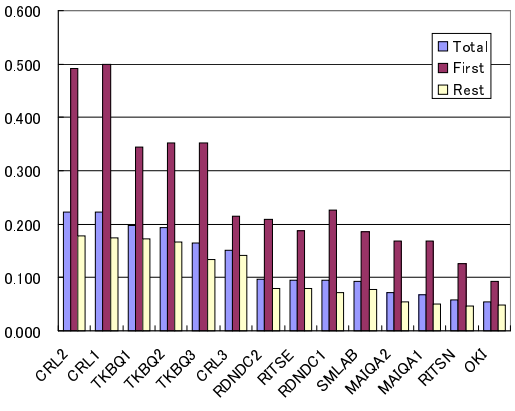

Figure 2: Evaluation by MMF

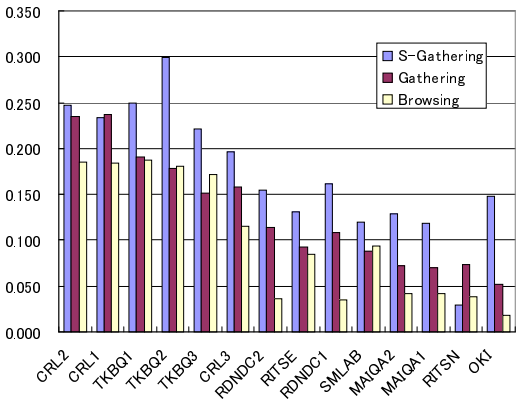

Figure 3: Differences on series types

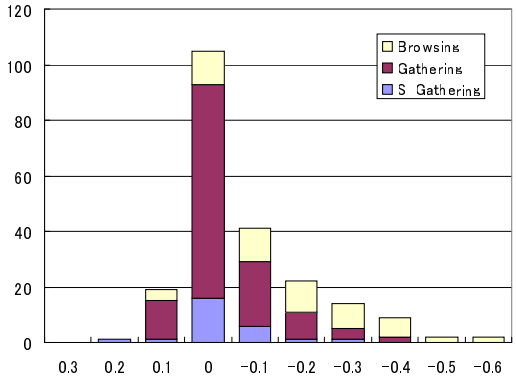

Figure 4: Difficulties of context processing

4.2 Techniques employed

As far as known from the participants' reports, techniques employed for context processing in the run are rather simple. Those, however, are some kinds of basis or seeds of techniques waiting for further developments.

In the most prevailing technique, systems do not analyze anaphoric expressions in a given question at all, but simply treat that question as a continuation of previous ones (Akiba et al., 2004; Hidaka et al., 2004; Takaki , 2004). For systems that utilize keywords extracted on the question analysis stage in subsequent document/passage retrieval and answer extraction, keywords extracted from previous questions in addition to those from the current one are took into consideration. One system, which uses a higher order features, such as word bi-grams, treats a word string made by concatenation of previous questions and the current one as input to be processed. Systems differ in which range of questions would be considered as the previous ones. Some use only the first of a series and others whole of the series up to the current one.

Another consideration is the balance between weights of keywords in previous questions and the current question. We have a system that also takes answers to previous questions into account.

We have a system employing a more sophisticated way of handling context on its question analysis stage (Fukumoto et al., 2004). It determines the referent using a shallow syntactic-semantic analysis of questions. An antecedent question is analyzed and is decomposed into the entity description, attribute description and interrogative expression, as in the simplest case, a question could be considered to be asking about some entity's some attribute. In the case where the current question has no explicit anaphoric expression, similarity of the interrogative expressions of the antecedent and the current one is used as a clue and it is determined whether the entity or attribute should be supplied to the current question. When the question contains an anaphoric expression, its semantic category is used for determining whether the referent is the entity or the attribute of the antecedent. Since one series consists of several questions, there is ambiguity on which of them could be the antecedent, which is resolved using heuristics.

QA systems could handle context in modules other than question analysis. We have a system that determines the documents from which the answers are extracted in processing the first question of a series and uses them exclusively throughout processing whole of the series (Hidaka et al., 2004). Although that technique seems rather rude, it is unique for QA technologies and its further refinement may bring a novel technique.

Each technique mentioned in this section has its pros and cons derived from its intrinsic characteristics. For example, document restriction by the first question cannot work properly for the browsing series. This time, however, it is not clear such a relationship between techniques employed and evaluation results of the run. It is probably because system performance depends on several factors such as robustness against noises 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

Tomoyoshi Akiba, Katunobu Itou and Atsushi Fujii. 2004. Question Answering using "Common Sense" and Utility Maximization Principle. *in this proceedings*.

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`.

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

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.

Jun'ichi Fukumoto, Tatsuhiro Niwa, Makoto Itoigawa and Megumi Matuda. 2004. Rits-QA: List answer detection and Context task with ellipses handling. *in this proceedings*.

Naoya Hidaka, Fumito Masui and Keiko Tosaki. 2004. MAIQA: Mie Univ. Participated System at NTCIR4 QAC2. *in this proceedings*.

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

Sharon Small, Nobuyuki Shimizu, and 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.

Toru Takaki. 2004. NTT DATA Question-Answering Experiment at the NTCIR-4 QAC2. *in this proceedings*.

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`.

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.