An Overview of NTCIR-5 QAC3

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

This paper provides an overview of NTCIR-5 QAC3 (Question Answering Challenge 3). QAC is a series of challenges for evaluating question answering technologies in Japanese. QAC3 follows the same course as QAC based on the success of the previous two workshops, with its task limited to that corresponding to QAC2 Subtask 3 aiming at the convergence of research resources for novel subjects. This task assumes interactive use of QA systems and evaluates, among other things, the abilities needed under such circumstances, i.e. proper interpretation of questions under a given dialogue context; in other words, context processing abilities such as anaphora resolution and ellipses handling (hereafter we refer to the task as the IAD task, where IAD stands for Information Access Dialogue, and to the whole workshop as QAC3). The IAD task in QAC3 is based on QAC2 Subtask 3 with several improvements, including elaboration of the scope of questions and answers and introduction of multi-grade evaluation and the concept of a correct answer set. In addition, a new WoZ method was devised and applied in the QAC3 test set construction. QAC3 had as many participants as QAC2 Subtask 3, and new trials and advances in existing methods were observed from the submission results.

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. These technologies make it possible to retrieve actual information rather than merely documents, and will lead to new styles of information access [18].

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. At TREC, which provides encouragement and guidance for QA research, a context task was attempted in order to evaluate the systems' ability to track context for supporting interactive user sessions at TREC 2001 [19]. 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 [20]. It is important for the researchers to recognize that such a cohesive manner is natural in QA, although the task itself is not for evaluating context processing abilities since, as it is given what is 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 [2]. 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 is observed in several projects that are part of the ARDA AQUAINT program [1][16][9][4]. These studies con-

cern 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.

QAC is a series of challenges for evaluating QA technologies for Japanese factoid questions, started in NTCIR-3 as QAC1 [3]. QAC3 follows the same course as QAC based on the success of the previous two workshops. From the outset, QAC has emphasized the importance of handling realistic problems, and one of the subtasks was to include interactive use of QA systems in its scope. The subtask was then extensively elaborated and established as an IAD task that can evaluate QA systems as a participant of information access dialogue [5]. The IAD task assumes the situation in which users interactively collect information using a QA system for writing a report on a given topic, while the range of individual questions remains in factoid questions. Our empirical studies show that this situation setting is realistic and restriction does not cause any significant problems [7]. The IAD task is the only QAC3 subtask conducted in NTCIR-5.

In this paper, the IAD task in QAC3 is defined first. Its improvements are discussed in detail, such as elaboration of the scope of questions and answers and introduction of multi-grade evaluation and the concept of a correct answer set. Next, the procedure and process of QAC3 are described. The new WoZ method, which was devised and applied in the QAC3 test set construction, is explained along with the characteristics of the test set constructed. Then, analyses of the test set, submission results, and the whole workshop follow. QAC3 had as many participants as QAC2 Subtask 3, and new trials and advances in existing methods were observed from the submission results.

2 Definition of the task

2.1 Basics of the IAD task

QAC is a series of challenges for evaluating QA technologies in Japanese. It covers factoid questions in the form of complete sentences with interrogative pronouns. Any answers to those questions should be names. Here, names mean not only names of proper items (named entities) including date expressions and monetary values, but also common names such as for 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. 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. The underlying document set consists of two years of articles from two newspapers, which, in QAC3, consisted of Mainichi and Yomiuri newspapers from 2000 and 2001. Using those documents as the data source, the systems answer various open-domain questions.

The IAD task assumes interactive use of QA systems and evaluates, among other things, the abilities needed under such circumstances, i.e. proper interpretation of questions under a given dialogue context, in other words, context processing abilities such as anaphora resolution and ellipses handling. In the IAD task, QA 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. 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 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.

We believe that there are two extremes of information access dialogues: 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, which therefore varies as the dialogue progresses. Accordingly, two types of series corresponding to these extremes were included in the IAD task. As mentioned, the IAD task assumes that users are interactively collecting information on a given topic. While mainly the gathering-type dialogue occurs under such circumstances, some focus shifts are also observed. This is why we included browsing-type series in our task; even when we are 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 IAD task is a simulation of interactive use of QA systems in dialogues.

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. The modifications, which are discussed later, were needed for handling QA specific features that contrast with standard document retrieval. Systems are mainly evaluated by means of this modified F measure over all questions. The judgment as to whether or not a given answer is correct 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, it is regarded as incorrect even if the answer itself is correct. The correctness of an answer is determined according to the interpretation of a given question performed 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.

The formal run of the IAD task was accompanied by two runs using reference test sets in order to evaluate context processing abilities isolated from several kinds of abilities concerning QA, and to examine the degree of context dependency of questions in the test set. 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. Although most of the questions in the second test set are semantically underspecified, such as asking the date of a birthday without specifying whose birthday, all the questions are syntactically well formed in the case of Japanese. In a sense, the first reference test set measures the ceiling of the context processing in a given original test set, while the second measures the floor. In addition, the run using the first reference test set has the role of increasing the number of candidates of correct answers for sufficient pooling.

2.2 Elaboration in QAC3

In QAC3, based on the experience gained from QAC2, two aspects of the IAD task have been improved. First, the scope of answers and questions was redefined to eliminate the vagueness that comes from the definition of "names" and to progress to handling more complicated questions. Second, the evaluation measure was elaborated and became more intuitive, introducing multi-grade evaluation and the concept of a correct answer set.

2.2.1 Redefinition of the scope of answers and questions

The following expressions concerning values were included in the scope of answers in QAC3, although it was unclear whether they were or not. Note that there can be huge differences in syntactical categories of expressions representing the same items between the Japanese and the English language; the following examples can be expressed in the form of compound nouns in Japanese.

- Numerical expressions with some additional expressions for specifying their nature, such as "300 bottles per year", "30 cm in length and 50 cm in width", " 3 liters per person" and "3 tons in weight".
- Conventional or colloquial range expressions, such as "10 to 12 percent", "from the end of the 8th century to the beginning of the 9th century", "more than 30" and "between 30 and 50". They include spatial/areal expressions, such as "between Tokyo and Osaka", "Haneda to Chitose" and "in Chiba."
- Expressions of approximate or round numbers, such as "around 100 persons", "almost 3 billion." They include spatial approximations such as "around Chicago", "near Tokyo", "in front of Maihama Station" and "the back of the embassy."

Those expressions are necessary to make some answers more informative and fluent. For example, it is not sufficient to answer 300 bottles to the question How much mineral water is consumed? because it is unclear whether the amount is for a month or a year. It is awkward to answer in a length when asked about size, such as answering "50 cm" to "How large is the packet?" Excluding them also causes problems for answer enumeration and duplication checking: "50 cm" and "30 cm" should be enumerated instead of "50 cm in length and 30 cm in width"; "10" alone extracted from "10 to 12 percent" is not valid because it is not accompanied by a unit expression; it is borderline whether "100 persons" provides the same information as "102 persons" when "around 100 persons" does; and so on.

With the aim of progressing to handling questions asking for reasons and situations such as "Why did it happen?" and "What happened?" descriptions of events were included in the scope as long as they have a form of noun compound, although it is unusual to call them names. Examples are "bursting of the levee", "crash and burning up", and "alkali aggregate reaction." Questions asking about features and definitions were also included, which should be answered using a noun phrase that explains those features and definitions rather than using only names or nouns. In order to accommodate those questions in the criteria, the head noun is selected as the correct answer. In addition, some nouns in the noun phrase, which can be semantically replaced with the head noun, were also considered as correct answers. For example, having

the support text "... in the swamp near Slatani Airport in the southern part of Thailand, where the Thai airplane crashed ...", the correct answers to the question "Where in Thailand did the airplane crash take place?" are "the swamp", "near Slatani Airport" and "the southern part of Thailand."

2.2.2 Evaluation Measure

Several difficult decisions must be made on how to evaluate the systems' output in the case of a list-type task, in which systems are requested to return one list that consists of all and only correct answers [6]. The difficulties and decisions are explained in the following section.

Duplications There is more than one expression denoting the same item: person names with/without a position name, varieties of notations of foreign objects, the same amount of money in different monetary units, and the same time in different time zones. Systems may include more than one expression of those items in their answer list, which can be seen as producing duplicate answers. A method for evaluating this situation should be determined.

Qualities Two qualities of answers should be distinguished. First, there are differences of quality in expressions that denote the same item. An abbreviation and an official name differ in this quality. On dates and places, this problem relates to the difference of granularity (specificity or particularity) such as between "2000" and "3rd of Jan 2000", and between "Japan" and "Urayasushi, Chiba." This difference in quality of expressions should be taken into account in the evaluation in an intuitive manner. Second, there are differences of quality in the answer itself (information or items denoted) rather than in their expressions. For example, a number is incorrect because of a mistake in the news source or the newspaper even though it was reported to be true in a newspaper article. A scheduled date was changed, which was described as determined in an article. Although the answers are judged correct because they have a supporting article, they are considered to have a lower quality in a sense and should be distinguished from genuine correct answers. These differences should be considered in the evaluation.

Enumeration method There is more than one way to enumerate all correct answers. For example, "Three prefectures of the Tokai region" conveys the same information as a list containing Mie, Aichi, and Gifu prefectures. When enumerating answers by extracting from the text "fish and shellfish such as carp, shrimp, and crab", which includes some examples, the enumeration of all, such as fish, shellfish, carp, shrimp, and crab, appears odd, while it is not clear which of the enumerations, fish and shellfish, or carp, shrimp and crab, is better. This problem occurs in combination with the difference of granularity. Both ways are possible for enumerating the sites of a series of events in city names and in country names. For an example of more complicated cases, let us suppose that some event took place on both December 10 and December 20. The answer of just "December" does not convey the information that this event occurred twice. This is also true when the answer is a list containing "December" and "December 10." The problem of granularity is originally related to the quality of expressions. However, when an answer is so coarse-grained that it is undistinguishable from other answers, like in this example, the problem goes beyond a simple matter of expressions and becomes a problem of enumeration. Variation of enumeration also comes from the inclusion of range expressions in the scope of answers. We must give the same evaluation to both "from the end of the 8th century to the early 9th century" and the list containing "the end of the 8th century" and "the early 9th century."

We took these problems into consideration and devised the following measure in order to make the evaluation as intuitive as possible. The points are introduction of the concept of a correct answer set and multigrade evaluation according to the two qualities of answers.

One correct answer set (hereafter CAS) corresponds to one way of enumerating the correct answers. More than one CAS are allowed for one question if needed. Intuitively, in the cases mentioned above, {"Three prefectures of Tokai region"} is one CAS, while {"Mie", "Aichi", "Gifu"} is another; {"December"} is one, while {"December 10", "December 20"} is another. A factor, h, which ranges from 0 to 1, is given for each CAS, which decides the value of information when all the answers in that CAS are enumerated. In many cases that factor equals one, but in the case above, for example, the set of {"December"} is given 0.5 for its h, as the enumeration of the answers of this set gives half the information compared to the other set.

More precisely, the CAS is a collection of expression sets (hereafter ES), each of which is a collection of several correct expressions denoting the same items. Actually, since the judgment on correctness is made for each pair of answer expression and its support article, members of an ES are such pairs, and the same expressions from different articles belong to the same ES. Each ES is given a factor, g, which ranges from 0

to 1, and reflects the quality of its denotation, the item or the information itself. Each expression in an ES is given a factor, f, which also ranges from 0 to 1, and reflects the quality of the expression. One expression can appear in only one ES within the same CAS.

For output set O, the answers that a system output is given, its precision P and recall R according to one CAS, $CASi$, are derived by the expressions shown on the next page. The F measure is calculated in the standard manner, the highest F measure over all CASs is the evaluation of the output set O. For a question with no answer, the empty set gets a full score, 1, and the others get 0, even though it is not shown in the expression. We call this evaluation measure $MF1$ (Modified F measure). The systems are evaluated by $MMF1$, the mean of $MF1$ over all questions in the test set.

The following explains the intended principles of this evaluation measure.

- The quality of expressions is represented by factor f. Using the sum of these factors in the numerator of the precision and recall definitions instead of the number of correct answers, the differences in the quality of expressions of given answers are reflected in the evaluation.
- The quality of the answers themselves is represented by factor g. Using the sum of these factors in the numerator and denominator of the recall definition instead of the number of correct answers, the balance between the answers according to that quality is reflected in the evaluation.
- Detecting duplications and eliminating them is considered part of the abilities measured. The precision is lowered when more than one expression from one ES is included in a given output set, by regarding only one of those with the highest quality as being correct and the others as being incorrect.
- One enumeration method should be chosen by a system. A given output set is evaluated according to each CAS and the highest evaluation is chosen. The mixture of correct answers from different CASs has no positive effect on the evaluation, but it has no negative effect either since correct answers in other CASs are distinguished from incorrect answers by being excluded from the calculation of the precision denominator.

The following example shows variations in handling correct answers. Let us assume that there are two correct answers, "Urayasu-shi, Chiba" and "in front of Maihama Station", to the question "Where is Tokyo Disneyland located?" As the first case, if these two answers are considered as variations of expressions that denote the same place, they should belong to the same ES. In this case, systems are expected to give only one of these answers. The precision becomes lower when both are given as the answers. As the second case, if these two answers are considered as different information, and the systems must enumerate both for answering the question completely, they should belong to different ESs in the same CAS. In this case, the recall becomes lower when only one of the answers is given. And as the third case, if you think that these are two different ways of answering and each answer provides enough information, the two should belong to different CASs. In this case, it is sufficient to give only one of the answers and the precision does not become lower even when enumerating both. In the second case, in which the system should give both answers, if you think "Urayasu-shi, Chiba" is preferable to "in front of Maihama Station", you can represent this preference by giving a small value to factor g of the latter's ES. By doing that, the answer {"Urayasu-shi, Chiba"} gets 0.67 of the recall, while {"In front of Maihama Station"} gets 0.33, for example. Moreover, if you think "Chiba" is another correct answer that denotes the same place as "Urayasu-shi, Chiba" but is less specific, it should belong to the same ES as "Urayasu-shi, Chiba" but with a smaller value for factor f.

In addition to this $MMF1$, we also evaluate systems using a supplementary measure, MRC (Mean Reciprocal Cost), the mean of RC defined by the following expression over all questions [7]. MRC is a precision-based measure, which is a natural extension of MRR used in the ranked list task. In MRC, duplication of correct answers does not cause any problems. The thresholds should be decided on the quality levels over which we regard the answer to be correct, since answers should be correct or incorrect, as multi-grade evaluation is not incorporated into MRC.

$$C = O \cap \bigcup_{ES \in \bigcup_i CAS_i} ES$$

$$RC = \begin{cases} \frac{|C|+1}{|O|+1} & if\ C \neq \phi \\ 0 & otherwise \end{cases}$$

3 Conducting Workshop QAC3

3.1 Schedule and Process

QAC3 follows the course of QAC based on the success of the previous two workshops, with its task limited to the IAD task. This concentration is, firstly, to avoid dispersing research resources and to focus on this important and novel subject. We also anticipated

$$P_{CAS_i} = \frac{\sum_{ES \in CAS_i} \begin{cases} \max_{e \in O \cap ES} f(e) & if\ O \cap ES \neq \phi \\ 0 & otherwise \end{cases}}{|O| - |(O - \bigcup_{ES \in CAS_i} ES) \cap \bigcup_{ES' \in \bigcup_{j \neq i} CAS_j} ES'|}$$

$$R_{CAS_i} = \frac{h(CAS_i) * \sum_{ES \in CAS_i} g(ES) * \begin{cases} \max_{e \in O \cap ES} f(e) & if\ O \cap ES \neq \phi \\ 0 & otherwise \end{cases}}{\sum_{ES \in CAS_i} g(ES)}$$

$$F_{CAS_i} = \frac{2 * P_{CAS_i} * Q_{CAS_i}}{P_{CAS_i} + Q_{CAS_i}}$$

$$MF1 = \max_i F_{CAS_i}$$

that the reference run, discussed above, could substitute for other subtasks. In addition, limited human resources on the organizer side would have made it difficult to conduct multiple subtasks.

QAC3 was declared at the NTCIR-4 workshop meeting held in June 2004, and the basic plan, which included the decision to conduct only the IAD task in QAC3, was decided and confirmed at a round table meeting in September. CFP and the details of the task definition were announced at the end of November, and the deadline for participation application was established as the end of 2004. A format checker and the previous test sets were delivered to the participants. No dry runs were conducted. A formal run was conducted over a period of one week starting from April 25, 2005. The test set was delivered using a WWW system, and the results were submitted as e-mail attachments. The processing time for submission was limited to 48 hours after downloading a test set for one system. An additional 24 hours was given for each submission if one team tried to make more than one submission. After the formal run, in May, a reference run was conducted in the same manner but without a strict deadline. Samples of the correct answers were delivered at the end of June and the beginning of August, and the final evaluation was delivered at the end of August.

3.2 Construction of the test set

Test set construction consists of collecting and making up questions and choosing and organizing them into series of questions.

3.2.1 Collection

Preparation: 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 between 800 and 1600 characters long and an abstract around 100 characters long were constructed using a full text search system on the newspaper articles. Four experts shared the preparation work.

Collection by questionnaire: 80 topics were selected from among the original 101 on the basis that enough information was gathered and compiled into the summary. Twelve subjects participated in the question collection experiment and 20 topics were assigned to each subject. That is, each topic was handled by three subjects. The subjects were asked to make up questions on the given topics using the following procedure consisting of two phases. Instruction, distribution and collection of materials were performed through postal mail. First, presenting only the topic and its abstract, we asked the subjects to make up a series of questions asking for appropriate information assuming a situation in which a report had to be written on the given topic. The questions were restricted to being in the form of a sentence with an interrogative pronoun, but were allowed to contain reference expressions. The recommended number of questions for each topic was 10. Second, presenting the topic and its summary, we asked the subjects to judge the questions made in the first phase and decide whether or not the information asked was appropriate to their report. We also asked the subjects to make up additional questions asking for information that they found necessary after reading the summary.

Collection by WoZ method: 20 topics were selected from the above 80. Six subjects participated in the question collection 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. The four experts who wrote the summaries played the role of simulated QA systems like a WoZ (Wizard of Oz), and each expert participated in dialogues on the topic for which she/he wrote the summary, and tried to answer questions from users using the summary, a full text search system, and his/her memory. Presented with the topic and abstract, the subjects, the users of the "QA system," were directed to think about questions beforehand assuming the situation in which they had to write a report on that topic, and then they participated in an information access dialogue with the "QA system". The subjects were taught that the system could answer only simple factoid questions, and the Wizards of Oz were instructed not to answer complicated questions such as asking for reasons and opinions.

A total of 2,416 questions were created in the first phase of collection by questionnaire, 1,874 of which were judged appropriate in the second phase. A total of 620 questions were asked in the collection by the WoZ method, and 504 of those were answered properly by the Wizards of Oz. Other questions were clarified, judged as being too complicated, or judged as not having any answer. We could not find any difference in the content and expressions between the questions collected by questionnaire and those by the WoZ method. Further investigation may be needed on this point.

We chose 502 questions from those collected by the WoZ method and answered appropriately and from those created in the first phase of collection by questionnaire and judged appropriate in the second phase, and checked whether or not the answers to those questions existed in the document set. The questions were expected to be spontaneous and natural in both content and expression, because they were created or asked without much knowledge on the topic. They were also expected to be appropriate because they were judged so by the subjects or by the Wizards of Oz.

We also made up another 200 questions separate from the experiments and checked the existence of the answers.

3.2.2 Constructing question series

Series of the gathering type, which are series of questions on a specific topic, were constructed by choosing questions from those collected on a given topic and confirming the existence of the answers, and by modifying and reordering to ensure appropriateness to the context in which they were placed. While the gathering type in QAC2 had a precise definition on the use of reference expressions, in QAC3, we loosened that restriction and decided to use the term gathering type for the series constructed in the above manner. In contrast, the browsing-type series are series of questions that are apparently related to different topics. Browsing-type series were constructed by choosing questions from those collected and made up, using them as seeds of a sequence, and adding new questions to create a flow to/from those questions.

3.3 Characteristics of the test set

Figure 1 shows examples of the series in the test set for QAC3. The first two belong to the gathering type. The topic of Series 30002 is the "Harry Potter" series, and that of Series 30004 is low-malt beer. In Series 30002, the fifth question has a pronoun that refers to the first volume rather than the series itself. In Series 30004, the interest seems to have moved from a specific brand of low-malt beer to low-malt beer in general. As illustrated, there is a wider variety of gathering-type series in QAC3 than in QAC2, and they are difficult to define by characteristics through the use of referential expressions. The third series in Figure 1 is an example of the browsing type. The topic moves from a theme park to an actor and then to a movie. The final question has no relation to the theme park, which was the original topic. This is in contrast to the gathering-type series, in which all questions are somehow related to the same topic. Figure 2 shows example questions of the first reference set. They correspond to the questions in Series 3002 shown in Figure 1.

The test set constructed for QAC3 contains 50 series and 360 questions, with 35 series of the gathering type and 15 series of the browsing type. The number of questions in one series ranges from 5 to 10, and the average is 7.2. The topics in the gathering-type series consist of 8 persons, 2 organizations, 11 events, 9 artifacts, and 5 animals, plants and so on. Eighteen questions out of 360 can be answered by event descriptions or by noun phrases. For example, Question 30001-05 "What was the submarine doing at the accident?" could be answered by "launching torpedoes" and Question 30017-01 "What type of company is the Arabian Oil Company?" could be answered by "an oil field developing company."

4 Analysis on the Workshop

4.1 Test set and evaluation principle

We found two mistakes in the question IDs for the QAC3 test set, which were corrected at the time of the formal run. Question 30015-07 contains a typo, which was not corrected, and was used as is.

There were several problems specific to the IAD task, in which some questions have a reference expression referring to the answer of the previous ques-

Series 30002
What genre does the “Harry Potter” series belong to?
Who is the author?
Who are the main characters in that series?
When was the first volume published?
What title does it have?
How many volumes were 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?
Which train station is the nearest?
Who is the actor who attended the ribbon-cutting ceremony on the opening day?
What is the movie he was featured in that was released in the New Year season of 2001?
What is the movie starring Kevin Costner 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 of Questions

tion. Question 30009-05 “Which country decided not to adopt the Euro monetary system by referendum?” had unintended correct answers, “England” and “Sweden,” in addition to the intended “Netherlands”, which affected the interpretation of the next questions such as “When was it conducted?” However, since we could not find the answer in the unintended interpretation, the correct answers used in the evaluation of those questions remain the original ones. Question 30022-04 was ambiguous, and had two possible interpretations: “the opponent who took the championship from him” and “the opponent who lost the championship to him.” Using multiple CASs, the answers corresponding to both interpretations were treated as correct, causing a discrepancy between the formal run and reference run, in which the question was resolved ac-

What genre does the “Harry Potter” series belong to?
Who is the author of the “Harry Potter” series?
Who are the main characters in the “Harry Potter” series?
When was the first volume of the “Harry Potter” series published?
What is the title of the first volume of the “Harry Potter” series?
How many volumes of the “Harry Potter” series were published by 2001?
How many languages has the “Harry Potter” series been translated into?
How many copies of the “Harry Potter” series have been sold in Japan?

Figure 2. Examples of Questions of the Reference Set

cording to the former interpretation. The same type of discrepancy exists in Question 30043-06, in which the question in the reference test set could be interpreted as referring to a product of the same name but different from the one being referred to in the context of the original test set.

The average number of correct answers was 1.98; the number of questions with only one correct answer was 204.

More than one CAS were needed in 37 questions out of 360 for intuitive evaluation. Many of them were needed for handling problems of granularity. For example, in Question 30016-02, the pregnancy of the princess was announced twice, on April 16 and May 15. These two days comprise one CAS. Another CAS must be used in order to treat the more coarse-grained answer “this year” as being correct. In this case, using a factor h, answering the latter got a lower recall than enumerating the two dates. Others relate to ways of extracting answer expressions. In Question 30018-07, we have two pieces of information on the origin of a fire: “third floor” and “elevator hall”. Multiple CASs are needed to treat both ways of answering by enumerating these two and by one combined phrase “elevator hall on the third floor” equivalently. In this case, the factor h is 1.0 for both CASs. In addition, we used multiple CASs for handling cases in which it is difficult to determine that possible answers differing in expressions denote the same item, such as “1994” and “six years ago” in Question 30010-07. Although it may deviate from the original principle, using two CASs for those answers does not lower the precision even when the system gives both answers.

Two grades are set for both the quality of expressions and the quality of the answer itself, and factors f

and g are 0.5 for the lower and 1.0 for the higher. The principles on both quality and answer enumeration follow. As for the quality of expressions, the following answers were determined to have lower quality.

- Answering in years to questions asking for the date of an event that took place in 2000 or 2001
- Answering with "Japan" to questions asking for a place that appears domestic.
- Answering only with family names or nicknames to questions asking for a person's name with the exception of some famous foreigners such as "Clinton."

As for the quality of the answer itself, the following answers were determined to have lower quality.

- The dates abroad for events that took place on different days in Japan and abroad, such as a publishing or release date, when the question does not specify whether abroad or domestic is required.
- Scheduled dates eventually changed, described as determined in an article.
- Minor alternatives for values that should be unique, most likely due to a mistake made by the newspaper or information source.
- Exceptional answers, such as an illegal cheap price as the answer to the question "How much should you pay for it?"

Besides ordinary cases, the enumeration of all information was requested in the following cases:

- Numbers that changes over different articles as time passes, such as the number of casualties.
- Numbers that are inconsistent among articles, even when they should be unique such as the size of a ship. In this case, an apparent minority gets a lower quality as mentioned above, and mixtures of specific values and round values are handled using multiple CASs.
- Addresses and names of facilities to questions asking for a place, such as "Where was it held?"

4.2 Submissions and their evaluations

Seven teams participated in QAC3 and 16 systems' results were submitted, although some violated the specifications for document sets and/or the deadline. All teams participated in the reference run, but some teams had a problem with the document set used and some participated in only one of the two reference runs.

Figure 3 shows the evaluation of all participant systems, each of which is designated by its run ID, the alphabetical part indicating the team who submitted the system. The $MMF1$ is depicted for three categories of questions: the entire test set, the first in each series, and the second and later. Figure 4 shows the evaluation by MRC, which differs very little from the one by $MMF1$, and the differences among systems are too small to change the system ranking. One of the possible reason is that more than half of the questions have only one correct answer, It narrows the range of recall; recall is always 1 when precision in not zero for those questions. Figure 5 shows the difference in performance according to the type of series: the $MMF1$ for the gathering type and the browsing type. For the majority of questions, those in the browsing-type series are more difficult than in the gathering type as anticipated. One system, however, received a better evaluation in the browsing type. Moreover, we could see some divergence in the balance of evaluations for gathering type and browsing type among systems submitted from the same team. This raised the expectation of various context processing techniques employed in those systems.

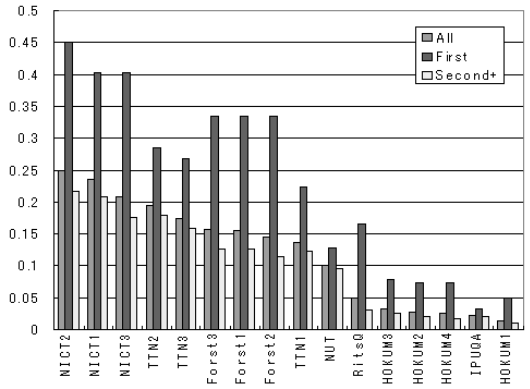

Figure 3. MMF1 Evaluation of the Formal Run

Figure 6 shows the evaluation by $MMF1$ of the first reference run. System `Forst1`, `Forst2` and `Forst3` in Figure 3 behave exactly the same in the reference run and are summarized in system `Forst1` in Figure 3. For the others, the same system has the same ID in both figures. Even in the reference run there is an obvious tendency for the evaluation of the second and later questions to fall compared to the first questions. We found the same tendency in QAC2. It appears not intuitive but understandable probably because when making a series of questions on a given topic, prominent questions, which are easier than the others, tend to be put at the head of the series. Nevertheless, the evaluation becomes lower in the range of 50–80% by employing context processing. As the

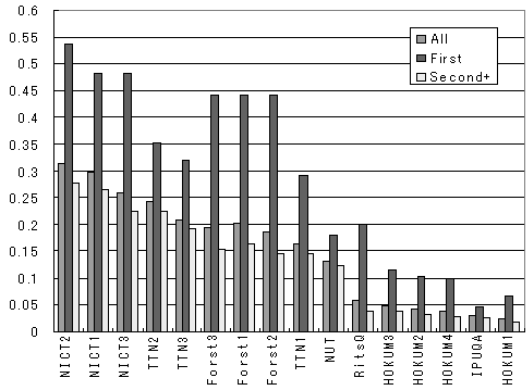

Figure 4. MRC Evaluation of the Formal Run

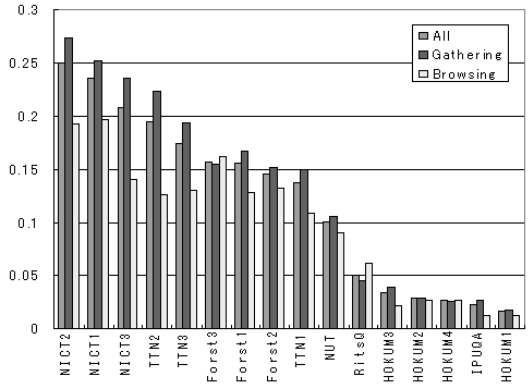

Figure 5. Differences on Series Types in MMF1

amount of deterioration differs among systems, we expected from this point also that there are some variations in context processing employed.

A few systems actively handled the range expressions and numerical expressions with some additional expressions, which is one of the highlights of QAC3. Few systems could correctly handle questions to be answered by event descriptions or by noun phrases, which is still difficult using current technologies.

4.3 Technologies employed

The basis of context processing in QA is to generate a question to be processed by concatenating previous questions or their keywords to the current question with reference expressions. The NICT system, which had the best performance in QAC3, employs this technique, and treats as the current question the concatenation of all the questions from the first of the series to the current [13]. Coupled with the high performance of the background QA system, this method achieved very good results. Infecting the current question with useless and possibly harmful keywords should pose a

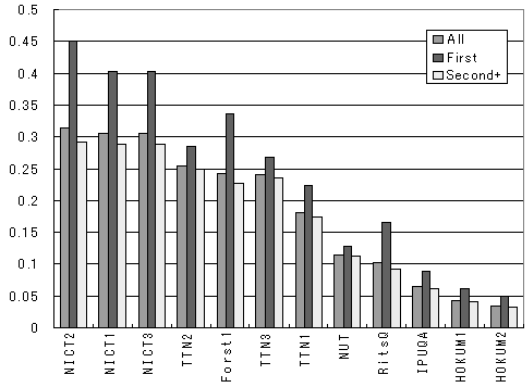

Figure 6. MMF1 Evaluation of the Reference 1 Run

problem, but the NICT system is supposed to be highly robust against such noise. This must be one of the reasons why this system is able to achieve excellent results both for isolated questions and for contextual questions with simple context processing.

In contrast, Rits is closely tackling context processing adopting traditional techniques for natural language understanding [10]. The system categorizes reference expressions into three types: pronouns, zero anaphora of a case element of verbs, and zero anaphora of a modifier or modificand of nouns. The latter two types are processed using case frames and co-occurrence data from the EDR Japanese Co-occurrence Dictionary, respectively. This reference resolution is based on application-independent linguistic knowledge and linguistic analysis, and attempts to address a wide range of phenomena. Unfortunately, it has not achieved satisfactory results for the present, because of insufficient linguistic knowledge and morphological and syntactic analysis with numerous errors. However, its future success is expected as a context processing method with wide coverage.

An important and impressive proposal presented in QAC3 is that the appropriateness of context processing can be measured by the appropriateness of the answers found to the resultant question. This was presented by both TTN and Forst, separately [14][12]. The TTN system selects which set of keywords should be linked to the current question from those in the first question, those in the previous question, the union of those two, and so on. This selection is made by measuring from which of those combinations the background QA system can obtain the most plausible passage containing the correct answer. It is a novel concept that the QA system itself can judge the appropriateness of the context processing. Although the NICT system can be regarded as implicitly doing the same thing, this explicit proposal is significant.

Forst combines this idea, called cohesion of knowl-

edge, with candidate narrowing using case frames and the focusing theory. In other words, an approach from natural language understanding similar to Rits's is synthesized with the idea of a QA system itself judging the appropriateness of the context processing. The Forst system handles only zero anaphora of a case element of verbs. It determines which of the candidates derived using case frames in Nihongo Goi Taikei and the focusing theory is most appropriate as a referent using the scores of answers that the QA system outputs for the question with that referent. There are some interesting observations. Adopting the focus theory, which slightly lowers the scores of the gathering-type series, improves the scores of the browsing-type series, which is usually much lower than the gathering type, to higher than those of the gathering type. This can be explained by the supposition that linguistic cues are frequently used when focus shifts occur and the focus theory can successfully capture them. In addition, it was observed that failure of context processing does not always cause failure of the whole QA process, which provides some ideas about context processing used for QA systems.

New trials were also examined for fields other than context processing. To our knowledge, the WWW filtering tried by HOKUM is the first one in QAC [8]. Several techniques were examined for selecting answer candidates, such as automatic acquisition of patterns [14], using the machine learning method for evaluating several measures [11], and using multiple support documents [13].

5 Review of the workshop

The following points should be considered for future workshops on QA technologies.

QAC3 was insufficient for the purpose of constructing a reusable test set. First, the pooling was not adequate, which was clear from the low coverage of the submitted answers over the correct answers previously found manually. Moreover, the variations in extractions were not covered in the current test set. That is, although variations such as "Eisaku Sato" and "Prime Minister Eisaku Sato," should be included in the correct answers, when "the late Prime Minister Eisaku Sato" was found to be correct, we did not expand the correct answers in the current test set. This deficiency is a significant problem, as one of the purposes of the evaluation workshops is to construct a test set.

The problems related to task definition and test set construction found in QAC2 remain unsolved. The history of a system's output is not taken into account in the evaluation. There is no empirical background on constructing browsing-type series.

We have to discuss the direction of QA technologies. Although QAC3 attracted almost the same number of people as the previous challenge, that is, Subtask 3 in QAC2, the total number of participants decreased compared to the QAC2 workshop. We emphasized that the IAD task could evaluate QA technologies in general using its reference run, but it did not attract more participants. We must make the task more attractive to a greater number of researchers on QA technologies. We have to discuss what is most important for accomplishing that. Do QA systems need context processing? Is the limitation to answering only factoid questions realistic? We should also further discuss the appropriate measure for intuitive evaluation.

6 Conclusion

QAC3 was a great success, although we face some problems for future work. Through QAC3, the IAD task has become more sophisticated, a new evaluation measure was proposed that could be used for the list-type task in general, and a new WoZ method for test set construction was devised and examined. The most important thing is that several new methods of context processing for question answering were tried and evaluated in this workshop.

Acknowledgements

We express our deep appreciation to Dr. Yutaka Sasaki for his valued comments on our evaluation measure. We also wish to thank all participants of QAC3 for their important contribution to the workshop and discussions.

References

[1] AQUAINT Home Page: Advanced Question & Answering for Intelligence. `http://www.ic-arda.org /InfoExploit/aquaint/`. 2003.

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

[3] Jun'ichi Fukumoto, Tsuneaki Kato and Fumito Masui. 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, 2003.

[4] Andrew Hickl, Johm Lehmann, John Williams, and Sanda Harabagiu. Experiments with Interactive Question Answering in Complex Scenarios. *Proceedings of HLT-NAACL2004 Workshop*

on Pragmatics of Question Answering, pp. 60-69, 2004.

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

[6] Tsuneaki Kato, Fumito Masui, Jun'ici Fukumoto and Noriko Kando. Characterization of List-Type Question Answering and its Evaluation Measures (in Japanese). *ISPJ SIG-NL 2004-NL-163*, pp. 115–112, 2004.

[7] Tsuneaki Kato, Jun'ici Fukumoto, Fumito Masui and Noriko Kando. Are Open-domain Question Answering Technologies Useful for Information Access Dialogues? – An empirical study and a proposal of a novel challenge – *ACL TALIP (Trans. of Asian Language Information Processing)*, 2005.

[8] Yasutomo Kimura, Kenji Ishida, Hirotaka Imaoka, et al. Three Systems and One Verifier – HOKUM's Participation in QAC3 of NTCIR-5 –. *in this proceedings*, 2005.

[9] Elizabeth D. Liddy. Preparing to Explore a New Paradigm in Information Access: A Scenario Approach to Question-Answering. `http://nrrc.mitre.org/NRRC/workshop03/Scenario_BaseQAWriteup.htm`. 2003.

[10] Megumi Matsuda and Jun'ichi Fukumoto. Answering Questions of IAD Task using Reference Resolution of Follow-up Questions. *in this proceedings*, 2005.

[11] Yasuharu Matsuda, and Takashi Yukawa. Synthesis of Multiple Answer Evaluation Measures using a Machine Learning Technique for a QA System. *in this proceedings*, 2005.

[12] Tatsunori Mori and Shinpei Kawaguchi. Answering Contextual Questions Based on the Cohesion with the Knowledge – Yokohama National University at NTCIR-5 QAC3 –. *in this proceedings*, 2005.

[13] Masaaki Murata, Masao Utiyama, and Hitoshi Isahara. Japanese Question-Answering Systems Using Decreased Adding with Multiple Answers at NTCIR 5. *in this proceedings*, 2005.

[14] Yuichi Murata, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. Towards Question Answering Challenge 3: Automatic Lexico-Syntactic Pattern Acquisition for Answer Evaluation and Context Processing exploiting Dynamic Passage Retrieval. *in this proceedings*, 2005.

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

[16] Sharon Small, Nobuyuki Shimizu, Tomek Strzalkowski, and Liu Ting. HITIQA: A Data Driven Approach to Interactive Question Answering: A Preliminary Report. *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 94-104, 2003.

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

[18] 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.

[19] Ellen M. Voorhees. Overview of the TREC 2001 Question Answering Track. *Proceedings of TREC 2001*, 2001.

[20] Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. *Proceedings of TREC 2004*, 2004.