Answering Contextual Questions Based on the Cohesion with the Knowledge
— Yokohama National University at NTCIR-5 QAC3 —

Tatsunori MORI and Shinpei KAWAGUCHI
Graduate School of Environment and Information Sciences, Yokohama National University
79-7 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
{mori,kawaguchi}@forest.eis.ynu.ac.jp

Abstract

*In this paper, we introduce the notion "*the cohesion with the knowledge*", and, based on it, propose a question answering system to answer contextual questions using a non-contextual QA system. The contextual questions usually have some cohesive relation to their context like reference expressions. Therefore, systems have to detect the cohesion and resolve the reference relations to answer the questions. Previous works usually address this problem in terms of the cohesion with the context. On the other hand, we address the problem by using the notion "the cohesion with the knowledge."*

First of all, the proposed method detects reference expressions in a given question. Second, it generates all possible completed question candidates by gathering antecedent candidates corresponding to reference expressions using I) the selectional restriction based on a case-frame dictionary and a thesaurus, or II) a modified version of the centering theory and I). As the degree of cohesion with the knowledge for each question candidate, we adopt the score of answer for each question candidate produced by a non-contextual QA system. The experimental results show that Strategy I is effective to improve the accuracy of answering the question series of the gathering type, on the other hand, Strategy II is effective for the question series of the browsing type.

Keywords: *Reference resolution, QA score, selectional restriction, centering theory, non-contextual QA system.*

1 Introduction

The technology of question-answering (QA) is widely regarded as an advancement on the combination of information retrieval (IR) and information extraction (IE). QA systems do not provide us with the relevant documents; instead, they provide answers to questions. In recent years, *contextual QA systems* have gained attention as a new technology to access information. In this paper, we define the *contextual QA* as "answering questions by taking into account the context, i.e., previously asked questions and their answers." Contextual QA systems are expected to be one of core modules for users to be able to access information interactively.

In this paper we propose a method to construct a contextual QA systems using an existing non-contextual QA system. Although a question for contextual QA systems generally has reference expressions like pronouns, ellipses, coreferences and so on, we expect that a non-contextual QA system is able to find answers for such a question if the reference expressions in the question are properly completed with their antecedents before the question is submitted to the QA system. The completion of a question may be performed in the following steps: 1) detect ellipses and reference expressions based on some linguistic information like a case frame dictionary, and then 2) find an antecedent for each reference expression. Both of these steps is problematic because one verb usually has multiple entries of case frame and there may be multiple candidates of antecedent for an ellipsis or a reference expression. In the research area of discourse understanding, there are many studies of the reference resolution in terms of *the cohesion with the context*. The centering theory is one of the most widely used methods[14]. This type of reference resolution tries find an optimal interpretation so as to maximize the cohesion between a newly introduced sentence and the context. This type of reference resolution would work surely, but it does not resolve the ambiguity in the detection of ellipses.

In this paper, we propose an another extreme side of reference resolution and a method to answering contextual question using the way of reference resolution. It is based on *the cohesion with the knowledge* instead of the cohesion with the context. Here, it should be noted that the QA system can refer to not only the context of dialogue but also the knowledge base when it is interpreting a question. It is also notable that "answering a question" can be regarded as finding an object, i.e. an answer, whose context in the knowledge base is coherent with the question. Therefore, the cohesion with the knowledge also may be one of the best influential criterion in finding the best interpretation of the question and consequently obtaining the best answer, if the question is ambiguous in a dialogue. It would be considered as a "to-the-best-of-my-knowledge" type of reference resolution."

Our implementation of *the cohesion with the knowledge* is summarized as follows. First, it generates all possible question candidates by completing each reference expression in a question with all possible antecedent candidates. In the completion, we take one of the following two strategies: I) a minimal semantic requirement for the antecedent, i.e., the selectional re-

striction based on the semantic consistency between the category of an antecedent and that of the case frame argument, or II) a modified version of the centering theory and I). Second, calculate the *degree of cohesion with the knowledge* for each completed question candidates. Here, our hypothesis is that the degree is analogous to the goodness of the answer for a question candidate, i.e. the score that calculated by the non-contextual QA for the question. Therefore, the question candidate and its answer with the highest score is considered as the best interpretation of the (original) question and the best answer for the question.

2 Related work

2.1 Contextual question answering

The approaches of the systems participated to NTCIR-4 QAC2 Subtask3 are mainly based on the cohesion with the context. In general, they are classified into two types described as follows. The first type of approaches is based on the effort in the document/passage retrieval. It expands the query submitted to the IR system with the words/phrases that appeared in the previously asked questions. Takaki[13] proposed the method of query expansion in which the query is formulated with not only the words/phrases obtained from the current question but also those from the last question.

The other type of approaches is based on the completion of questions by resolving reference expressions. One completed question is submitted to the non-contextual QA system. Fukumoto et al.[1] proposed a method to completing questions by using contextual information. They classify questions with reference expressions or ellipses into three types: the ellipsis of adjective expressions, ellipsis of the topic presentation parts, and the pronouns. The antecedents of each type of reference expressions are completed with the expressions of the same type in the previous question.

Our method is similar to the latter approach because the method is based on the completion of reference expressions in a question and the non-contextual QA system. However, as described before, it is based on the cohesion with the knowledge instead of the cohesion with the context.

2.2 Reference resolution

The process of reference resolution consists of the following two steps: 1) the detection of reference expressions, and 2) the identification of the antecedent for a detected reference expression.

2.2.1 Detection of reference expressions

In Japanese, there are several types of reference expressions like demonstratives, pronouns, and zero pronouns. Zero pronouns are ellipses of obligatory case elements and they behave like ordinary pronouns in Japanese. Especially, the detection of zero pronouns is very important and is studied from various viewpoints. One of the most widely used methods is the detection using a case-frame dictionary.

A case-frame dictionary has entries for declinable words such as verbs. Each entry consists of a set of sentence patterns, i.e. case-frames for the word. For example, each entry of the IPAL-BV (basic verbs)[3] has the information of case-frames. The "nihon-go goi taikei" (a Japanese lexicon)[8] also includes the information. A case-frame dictionary is used to find the unoccupied case in the input sentence by comparing the dependency relations in the sentence with the case-frame for the main declinable word of the sentence. Seki et al.[12] utilizes IPAL-BV for the detection of zero pronouns.

2.2.2 Identification of antecedents

Kawahara et al.[7] propose a method to detect and resolve zero pronouns using a case-frame dictionary that is automatically constructed from a corpus. The case-frame dictionary provides fine-grained selectional restriction that filter out inadequate antecedent candidates. They also introduce the structural preference of antecedents to narrow down the candidates.

The centering theory is one of the most widely used methods for anaphora resolution[14]. Nariyama[11] proposes a modified version of the centering theory for resolving Japanese zero pronouns. It utilizes a "*salient referent list (SRL)*", which pools all overt arguments which have appeared up to the sentence in question. An SRL is created for each new sentence by modifying the one for the preceding sentence. If a new argument appears with an identical grammatical relation to another argument already existing in the SRL, the new argument takes its place because of recency. In an SRL, arguments are listed in the following order, termed "salient referent order list":

> Topic (with the topic marker WA) > Nominative (with the case marker GA) > Dative (with the case marker NI) > Accusative (with the case marker O) > Others.

A zero pronoun is resolved by selecting the most salient argument in the SRL. If a sentence has multiple zero pronouns, the zero pronouns are resolved in the same order of the salient referent order list.

Iida et al.[2] proposed a method that combines i) the feature of local contextual factors, and ii) a learning model based on the comparison between two antecedent candidates, in order to resolve Japanese zero pronouns. In i), they use the SRL method.

3 Proposed method

Figure 1 shows the overview of the proposed method. The method obtains an answer list for each question in a given series of questions by the following procedure. It is should be noted that the non-contextual QA system can perform the list-type question answering. The list-type question answering is the task in which a system requested to enumerate all correct answers, i.e. an answer list, to a given question. Each answer in an answer list has its own score, but we adopt the maximum score of them as the score of answer list.

1. Detect reference expressions including zero pronouns in a new question using a case frame dictionary, then generate question candidates with zero pronouns.

2. Find antecedent candidates using one of two strategies: Strategy I or Strategy II. Strategy I gathers all nouns from i) the completed preceding question, ii) its answer list, and iii) *topic phrases* (described later) in the first question as antecedent candidates for each reference expression. On the other hand, Strategy II is based on a modified version of Nariyama's SRL-based centering theory.

3. Generate all possible completed question candidates by completing reference expressions in the question candidates with pronouns generated in Step 1. Then select the M-best completed question candidates according to the semantic consistency in reference resolution.

4. Submit the question candidates to the non-contextual QA system, and obtain answer lists and their score. The best answer list is the final output for the question.

With the above procedure, on a generate-and-test basis, the method tries to find the best interpretation of reference expressions in terms of the cohesion with the knowledge. Here, our hypothesis is that the degree of cohesion with the knowledge is analogous to the goodness of the answer list for a question candidate, i.e. the score that calculated by the non-contextual QA for the question. The method can be regarded as the reference resolution by maintaining the coherence of the question with the description in the document collection.

3.1 Example

Before we explain the detail of each step, we describe the flow of the procedure (Strategy I) with the following series of questions in this subsection.

(1) a. Tai-de Hikouki-Jiko-ga
Thailand-LOC airplane-accident-NOM
Oki-ta-no-wa
happen-PAST-NOUN-TOP
98-nen-no Itsu de-su-ka
98-year when BE-POL-INTERROG
When did an air accident occur in 1998?

b. Tai-no doko-de
Thailand-REL where-LOC
Oki-ta-no de-su-ka
occur-PAST-NOUN BE-POL-INTERROG
Where did ϕ occur in Thailand?

Since Question (1a) is the first question of the series, the system gathers *topic phrases* from the question, and looks up them in a thesaurus to obtain the semantic categories of them. The topic phrases and their semantic categories are preserved for the following questions. We define topic phrases as the candidate phrases for the (potential) topic of each question. As described in Section 3.4, in this paper, we adopt all nouns and the noun phrase with the Japanese topic marker "wa" in the first question. In this example, the topic phrases are the following four phrases:

(2) a. "Tai-de Hikouki-Jiko-ga Oki-ta-no"(category UNKNOWN),

b. "98-nen" (YEAR),

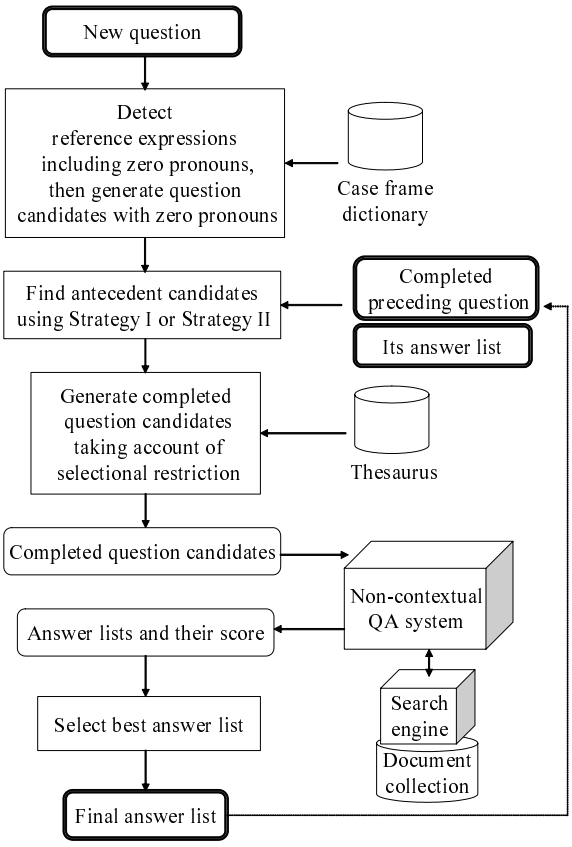

Figure 1. Overview of the proposed method

c. "Tai" (HUMAN, NATION, LOCATION), and

d. "Hikouki-Jiko" (ACCIDENT),

where the categories are assigned based on the "nihongo goi taikei"(a Japanese lexicon).

Then the question is submitted to the non-contextual QA system without any modification because it is the first question, and we obtain an answer list. The system also assigns a semantic category to each answer in the list as follows:

(3) a. "12-nichi (Day 12)" (category DAY),

b. "11-nichi-yoru (Night of Day 11)" (NIGHT AND DAY).

This is the end of the process for the first question.

Next, the system receives Question (1b). Since the question is not the first question and has a context, the system tries to detect reference expressions as described in Section 2.2.1. After the system checks that the question does not have either demonstratives or pronouns, the system tries to detect zero pronouns as follows. First, the system looks up the verb of the question in a case-frame dictionary and obtains case-frames. Second, unoccupied cases in the question are detected by comparing the dependency relations in the question with the case-frames. The system also obtains the information of semantic categories for unoccupied cases from the case-frames. In the case of Question (1b), the system obtains a case-frame for the verb "okiru" (occur) [1] and detects that

(4) a noun phrase with the case marker "GA" whose semantic category should be INCIDENT

is omitted and there is a zero pronoun.

The antecedent candidates for the zero pronoun are gathered from the context. As described in Section 3.4, in Strategy I, the set of antecedent candidates consists of i) all nouns and the noun phrases with the topic marker "wa" in the preceding question (the phrases (2a), (2b), (2c), and (2d)), ii) all phrases in the answer list of the question (the phrases (3a) and (3b)), iii) and topic phrases[2].

According to the similarity calculation with the semantic category of the zero pronoun (4) (as described in Section 3.5), the system narrows down the antecedent candidates (the phrases (2a), (2b), (2c), (2d)), (3a), and (3b)) to the two phrases "Tai-de Hikouki-Jiko-ga Oki-ta-no" (2b) and "Hikouki-Jiko" (2c). Thus, we have the following completed question candidates, although these candidates describe almost same question and Question (5a) is redundant:

(5) a. Tai-de Hikouki-Jiko-ga
Thailand-LOC airplane-accident-NOM
Oki-ta-no-ga
happen-PAST-NOUN-NOM
Tai-no doko-de
Thailand-REL where-LOC
Oki-ta-no de-su-ka
occur-PAST-NOUN BE-POL-INTERROG
Where did the event that an air accident occured occur in Thailand?

b. Hikouki-Jiko-ga
airplane-accident-NOM
Tai-no doko-de
Thailand-REL where-LOC
Oki-ta-no de-su-ka
occur-PAST-NOUN BE-POL-INTERROG
Where did an air accident occur in Thailand?

The system selects the M-best completed question candidates according to the semantic consistency in reference resolution as described in Section 3.7, and submits them to the non-contextual QA system. Since we hypothesize that the question candidate whose answer list has the highest score is the best interpretation in terms of the cohesion with the knowledge, the system outputs the answer list as the final answer.

3.2 Non-contextual Japanese QA system

The non-contextual Japanese QA system we used is a Japanese real-time QA system based on Mori[9]. As shown in Figure 2, it consists of five modules, i.e., the question analyzer, the search engine, the passage extractor, the sentential matcher and the answer generator. The question analyzer receives a question from

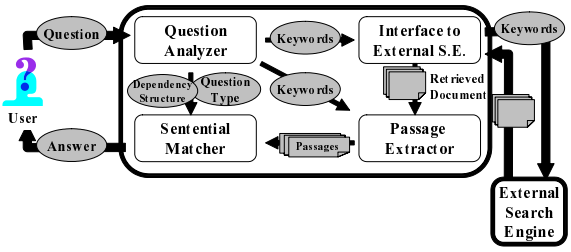

Figure 2. Overview of the Japanese QA system

a user and extracts a list of keywords, the question type, and the dependency structure. Here, we define the term *keywords* as content words in a given question.

The search engine retrieves documents related to keywords, which are obtained by the question analyzer. Although the QA system may use any kind of search engine, we currently use our original search engine. It is based on an ordinary tf*idf method for term weighting and the vector space model for calculating similarity between a list of keywords and a document.

Since the information related to a question is usually contained in a very small part of the document, the passage extractor segments each document, which is retrieved by an external search engine, into small passages and selects suitable passages that are related to keywords. In our experiment, we defined one passage as a sequence of three sentences, similar to Murata et al.[10].

The sentential matcher receives a set of sentences in retrieved passages. The module treats each morpheme as an answer candidate and assigns it a matching score. The matching score represents the fitness of each answer candidate for the answer. We adopt a composite matching score shown in Equation (1), which is a linear combination of the following sub-scores for an answer candidate AC in

[1] The expression "okiru" is the root form of "oki-ta" in Question (1b)

[2] In this case, they are identical to the phrases in i).

the i-th retrieved sentence L_i with respect to a question sentence L_q: 1) the matching score in terms of 2-grams, $Sb(AC, L_i, L_q)$, 2) the matching score in terms of keywords, $Sk(AC, L_i, L_q)$, 3) the matching score in terms of dependency relations between an answer candidate and keywords, $Sd(AC, L_i, L_q)$, and 4) the matching score in terms of the question type, $St(AC, L_i, L_q)$. In the calculation of $St(AC, L_i, L_q)$, we employ an NE recognizer that spots NEs in eight types defined in the IREX-NE task[4].

$$\begin{aligned} & S(AC, L_i, L_q) \\ = \; & Sb(AC, L_i, L_q) + Sk(AC, L_i, L_q) \\ & + Sd(AC, L_i, L_q) + St(AC, L_i, L_q) \end{aligned} \quad (1)$$

An answer candidate obtained by the sentential matcher is a morpheme; a morpheme may be either a word or a part of a longer compound word. In the latter case, the system finds the compound word including the answer candidate, and outputs it.

We also add an extension of the list-type QA processing proposed by Ishioroshi et al.[5] to the system.

3.3 Detecting of reference expressions

Our method treats the three types of reference expressions, namely, i) demonstratives, ii) pronouns, and iii) zero pronouns. The detection of reference expressions of the types i) and ii) is not hard because they explicitly appear in questions. On the other hand, in order to detect zero pronouns we need some extra linguistic knowledge because only based on a given question we cannot guess what case elements are omitted. We employ an existing method based on a case-frame dictionary as described in Section 2.2.1. With regard to the case-frame dictionary, we use the "nihon-go goi taikei" (a Japanese lexicon)[8]. We also obtain the information of semantic categories for each case elements from the case-frame. The information is used to check the selectional restriction about demonstratives and pronouns as well as zero pronouns. If no reference expressions are detected in a question, the system suppose that a topic phrase is omitted in the question, and introduces a zero topic phrase, which is marked with the topic marker "wa", in order to force the question to have a relation with the context.

3.4 Gathering candidates of antecedents

3.4.1 Strategy I: gathering all possible nouns as candidates

In strategy I, we adopt the following phrases as the antecedent candidates for each reference expression in the current question:

- the noun phrases with the Japanese topic marker "wa" and all nouns in the preceding completed question,
- all phrases in the answer list of the preceding (completed) question, and
- topic phrases, that is, all nouns and the noun phrase with the topic marker "wa" in the *first* question.

Here, it should be noted that the system search not the original preceding question but the *completed* preceding question for antecedent candidates. The expressions in the questions before the preceding question may be retained in the completed question if the questions keep referring to them.

The system also looks up these phrases in a thesaurus to obtain the semantic categories of them. The information is used to narrow down the antecedent candidates by checking the selectional restriction. We adopt the "nihon-go goi taikei" (a Japanese lexicon)[8] as a thesaurus.

With respect to topic phrases, we focus on the first question because the question tends to play an important role in making the context of the question series. It is notable that we regard the noun phrases with the topic marker "wa" as important. The marker "wa" explicitly represents that the noun phrase with it is the main topic of the current context. Therefore, we adopt not only nouns in it but also itself.

Of course, we have to treat the topic shift appropriately. However, as an approximation we take the following simple strategy for it. Topic phrases are basically retained untile the end of question series. But a topic phrase is discarded and will not be used in the following questions if the topic phrase is not nominated for any antecedent candidates for the current question.

3.4.2 Strategy II: a method based on Nariyama's SRL-based centering theory

In this strategy, for each reference expression in a question candidate with zero pronouns, the system select one *noun phrase* (i.e. *bunsetsu* segment) according to a modified version of Nariyama's SRL-based centering theory described in Section 2.2.2. The difference from Nariyama's method is as follows:

- Although Nariyama's method takes into account all preceding sentences by maintaining the SRL, our method is a simplified version of it and it only looks up the SRL obtained from the *completed* preceding question.
- Demonstratives and pronouns in a new question are resolved before zero pronouns.
- The interrogative in the completed preceding question is replaced with one of answers in the answer list. Thus, we may have multiple completed question candidates even if the verb of the question has only one case frame, when the case corresponding to the interrogative is the most salient in the SRL.

3.5 Narrowing down antecedent candidates using the selectional restriction

For each reference expression in the current question, the system narrows down antecedent candidates obtained by the method described in Section 3.4 based on a selectional restriction.

The selectional restriction is based on the similarity between the semantic categories of antecedent candidates and those of reference expressions. The similarity is calculated with the following equation, that is the

same as Kawahara et al.[7]:

$$sim(x,y) = \begin{cases} \frac{2\times L_{xy}}{l_x+l_y} & \text{if } x \notin y \\ 1 & \text{if } x \in y \end{cases} \quad (2)$$

where l_x and l_y are the depthes of the categories x and y in the thesaurus respectively, and L_{xy} is the depth of the lowest common ancestor of x and y.

We determine a threshold value Th_{sim} of the similarity, and filter out each antecedent candidate whose similarity is less than the threshold value.

3.6 Generating completed question candidates

By completing each reference expression in the current question with all possible antecedent candidates, the system generates all possible candidates of the completed question.

3.7 Narrowing down completed question candidates

The process described so far may generate a lot of question candidates, and the non-contextual QA systems may take a very long time to process them. Therefore, we introduce a measure for a completed sentence in terms of *the degree of consistency in reference resolution*, and narrow down the question candidates by using the measure. We defined the degree as Equation (3). It is the summation of the consistency between each reference expression and its antecedent candidate in a question candidate:

$$C(S) = \sum_{<r_i,a_i>\in resolv(S)} c_1(r_i,a_i) \quad (3)$$

$$c_1(r,a) = \begin{cases} 1 & \text{if } a \in r \wedge a\text{is not an NE} \\ 1.5 & \text{if } a \in r \wedge a\text{is an NE} \\ sim(r,a) & \text{if } a \notin r \end{cases}$$

where $resolv(S)$ is the set of pairs of a reference expression and its antecedent candidate in the sentence S. We define $c_1(r,a)$ as 1.0 if the category of the antecedent candidate a is a descendant of the category of the reference expression r, because the situation is totally consistent. Some extra point is added to the value if the antecedent candidate a is a named entity because of our observation that an named entity tends to be an antecedent. If the situation is not totally consistent, we define $c_1(r,a)$ as the similarity $sim(r,a)$.

According to the degree of consistency in reference resolution, we select the M-best candidates of the completed question.

3.8 Finding the best answer by the non-contextual QA system

The selected question candidates are submitted to the non-contextual QA system. Since we hypothesize that the question candidate whose answer list has the highest score is the best interpretation in terms of the cohesion with the knowledge, the final answer is the answer list with the highest score.

4 Experimental results in NTCIR-5 QAC3

We evaluate the proposed systems in terms of the accuracy of reference resolution and the accuracy of question answering by using the test set of NTCIR-5 QAC3. The test set consists of 50 series and 360 questions. In these series, 35 series are of the *gathering type* and 15 series are of the *browsing type*. A question series of the gathering type contains questions that are related to one topic. On the other hand, in a series of the browsing type, the user does not have any fixed topic of interest and the topic of interest varies as the dialogue progresses. Here, it should be noted that the systems cannot use the type of series in answering questions. The document collection as the knowledge source consists of all (Japanese) articles in Mainichi Shimbun Newspaper and Yomiuri Shimbun Newspaper published in 2000 and 2001. The non-contextual QA system used in the systems is identical to the system that participated in NTCIR-4 QAC2 subtask1[9] except for an extension of the list-type QA processing[5]. The threshold value Th_{sim} for the selectional restriction is 0.5, and the number M of completed question candidates to be selected is 20.

With regard to the measures for the accuracy of reference resolution, we adopt the recall $R_{r.res.}$, the precision $P_{r.res.}$, and the F measure $F_{r.res.}$ defined as follows:

$$R_{r.res.} = N_{\text{reference}}/N_{\text{correct}}$$
$$P_{r.res.} = N_{\text{reference}}/N_{\text{detected}}$$
$$F_{r.res.} = \frac{2\cdot R_{r.res.}\cdot P_{r.res.}}{R_{r.res.}+P_{r.res.}}$$

where $N_{\text{reference}}$, N_{correct}, and N_{detected} are the number of reference expressions to be resolved, the number of reference expressions that are correctly resolved by the system, the number of reference expressions that are detected by the system.

With regard to the the measures for the accuracy of question answering, we use the recall $R_{ans.}$, the precision $P_{ans.}$, and the mean of the modified F measure $MMF1$ defined by Kato et al.[6].

Tables 1, 2, and 3 show the experimental results of reference resolution. The experimental results of question answering are shown in Tables 4, 5, and 6.

Table 1. Evaluation of reference resolution (all series)

Strategy	Recall ($R_{r.res.}$)	Precision ($P_{r.res.}$)	F ($F_{r.res.}$)
I (Forst3)	0.285	0.236	0.248
II (Forst1)	0.369	0.339	0.346

5 Discussion

5.1 Performance of reference resolution

As shown in Table 1, Strategy II, which uses the centering theory together, is more accurate than Strategy I in terms of both of the recall and precision. It

Table 2. Evaluation of reference resolution (series of the gathering type)

Strategy	Recall ($R_{r.res.}$)	Precision ($P_{r.res.}$)	F ($F_{r.res.}$)
I (Forst3)	0.342	0.285	0.297
II (Forst1)	0.404	0.365	0.374

Table 3. Evaluation of reference resolution (series of the browsing type)

Strategy	Recall ($R_{r.res.}$)	Precision ($P_{r.res.}$)	F ($F_{r.res.}$)
I (Forst3)	0.152	0.121	0.130
II (Forst1)	0.288	0.277	0.281

Table 4. Evaluation of question answering (all series)

Strategy	Recall ($R_{ans.}$)	Precision ($P_{ans.}$)	$MMF1$
I (Forst3)	0.161	0.198	0.156
II (Forst1)	0.158	0.197	0.156

Table 5. Evaluation of question answering (series of the gathering type)

Strategy	Recall ($R_{ans.}$)	Precision ($P_{ans.}$)	$MMF1$
I (Forst3)	0.172	0.207	0.168
II (Forst1)	0.154	0.186	0.154

Table 6. Evaluation of question answering (series of the browsing type)

Strategy	Recall ($R_{ans.}$)	Precision ($P_{ans.}$)	$MMF1$
I (Forst3)	0.136	0.176	0.128
II (Forst1)	0.168	0.222	0.161

is an unsurprising result, because the centering theory is a method with an established reputation and works well in many cases. However, it should be noted that there is another reason for the difference. The reason is that all nouns in the preceding completed question can be antecedents of reference expressions in Strategy I, on the other hand, in Strategy II, only noun phrases with case markers or topic markers can be antecedents. Since the number of candidates in Strategy I is usually much larger than the number of reference expressions including zero pronouns, almost all zero pronouns in the current question are filled with candidates in generating possible completed question candidates, even if some of zero pronouns have to be blank according to the context.

In the current implementation, zero pronouns are only treated as ellipsis, and we do not take into account the ellipsis of the modifier "NP_1-NO" in the noun phrase "NP_1-NO NP_2," which can not be handled by the centering theory. However, there are many questions that have this type of ellipsis.

With regard to the types of series, Tables 2 and 3 show the difference between the gathering type series and the browsing type series. Both strategies process the gathering type series better than the browsing type series.

5.2 Overall performance of question answering

To the contrary, the result of evaluation of question answering shown in Table 4 is very interesting because Strategy I has the almost same accuracy as Strategy II in spite of its insufficient performance in reference resolution. By comparing Tables 5 and 6, we can see that Strategy I has much better performance for series of the gathering type than the browsing type series. In the next section, we will discuss the reason why Strategy I has better performance than we expected.

On the other hand, Strategy II is well-balanced. It works for the gathering type as well as the browsing type with almost same accuracy.

5.3 Failure analysis

A detailed analysis of success and failure is summarized in Table 7. In this table, "Success" means that the answer list generated by the system contains at least one correct answer, otherwise. The other cases are "Failure."

With respect to the success, there are many cases that the reference resolution in a question is failed but the system successfully finds the answers for the question. Strategy I has stronger tendency to succeed in such cases than Strategy II. It means that the introduction of many expressions of the preceding questions into the current question has a good effect on the performance of answering questions even if the accuracy of reference resolution is insufficient. One of the reasons is that these newly introduced expressions may work well in the early stages of question answering like document/passage retrieval, in which deep linguistic processing is not performed. Another reason is that the non-contextual QA system is robust to non-grammatical questions because of the composite matching score as described in Section 3.2.

Table 7. Detailed analysis of failure

	Success		Failure				
Strategy	Res. OK Ans. OK	Res. NG Ans. OK	Ante. NG	Q. Gen. NG	Q. Sel. NG	Ans. NG	Others
I (Forst3)	7.4% (23)	14.5% (45)	32.6% (101)	10.6% (33)	21.9% (68)	10.3% (32)	2.6% (8)
II (Forst1)	11.3% (35)	10.0% (31)	42.6% (132)	17.7% (55)	1.3% (4)	16.1% (50)	1.0% (3)

Res.: reference resolution
Ans.: question answering by the non-contextual system
Ante.: appropriate antecedent
Q. Gen.: generation of completed question candidates
Q. Sel.: selection of an appropriate question candidate

With regard to the failure, we can see in Table 7 that the main reason lies in "Ante. NG". The column expresses the ratio of cases in which the appropriate antecedents of reference expressions in the current question do not appear in either the completed preceding question or its answer list. The failure is caused by, at least, the following reasons:

- The system failed to find correct answers for some previous questions.
- The system failed to find appropriate antecedents for reference expressions in completing some previous questions.

It should be reminded that the current version of our method only searches the (completed) preceding question for antecedent candidates. An expression in a older question can be an antecedent of reference expressions in a new question only if the expression was continuously referred to by the following questions. Thus, if there is a failure in reference resolution at some point, the failure will also cause other errors of reference resolution in the following questions.

6 Conclusion

In this paper, we introduce the notion "*the cohesion with the knowledge*", and, based on it, propose a question answering system to answer contextual questions using a non-contextual QA system. First, it generates all possible question candidates by completing each reference expression in a given question with possible antecedent candidates in the previous (completed) question and its answers. Second, it estimates the degree of cohesion with the knowledge for each completed question candidates by using the answer score produced by a non-contextual QA system. For the completion of a given question, we introduced two strategies, namely, I) the selectional restriction based on a case-frame dictionary and a thesaurus, and II) a modified version of the centering theory and I).

Experimental results in NTCIR-5 QAC3 show that Strategy I has much better performance for series of the gathering type than the browsing type series and Strategy II is well-balanced. According to our failure analysis, the main reason of failure is the appropriate antecedents of reference expressions in the current question do not appear in either the completed preceding question or its answer list. Therefore, we need some other devices to keep antecedent candidates in the context like Nariyama's SRL.

References

[1] J. Fukumoto, T. Niwa, M. Itoigawa, and M. Matsuda. Rits-QA: List answer detection and Context task with ellipses handling. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 310–314, June 2004.

[2] R. Iida, K. Inui, and Y. Matsumoto. Identifying antecedents of Japanese zero-pronouns using a machine learning model with contextual cues. *Journal of Information Processing Society of Japan*, 45(3):906–918, Mar. 2004.

[3] IPA Technology center. *The lexicon of the Japanese basic verbs for Computers*. Information-technology Promotion Agency(IPA), Japan, Mar. 1987. (in Japanese).

[4] IREX Committee, editor. *Proceedings of IREX workshop*. IREX Committee, 1999. (in Japanese).

[5] M. Ishioroshi and T. Mori. A method of list-type question-answering based on the distribution of answer score generated by a ranking-type q/a system. SIG Technical Reports 2005-NL-169, Information Processing Society of Japan, Sept. 2005. (in Japanese).

[6] T. Kato, J. Fukumoto, and F. Masui. An Overview of NTCIR-5 QAC3. In *Working Notes of the Fifth NTCIR Workshop Meeting*, Dec. 2005.

[7] D. Kawahara and S. Kurohashi. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 334–341, 2004.

[8] N. C. S. Laboratories. *Nihongo Goi Taikei (A Japanese Lexicon)*, 1997.

[9] T. Mori. Japanese question-answering system using a^* search and its improvement. *ACM Transactions on Asian Language Information Processing (TALIP)*, to appear. Special Issue for NTCIR-4.

[10] M. Murata, M. Utiyama, and H. Isahara. Question answering system using similarity-guided reasoning. SIG Notes 2000-NL-135, Information Processing Society of Japan, Jan. 2000.

[11] S. Nariyama. Grammar for ellipsis resolution in Japanese. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 135–145, 2002.

[12] K. Seki, A. Fujii, and T. Ishikawa. Japansese zero pronoun resolution using a probabilistic model. *Journal of Natural Language Processing*, 9(3):63–85, July 2002. (in Japanese).

[13] T. Takaki. NTT DATA Question-Answering Experiment at the NTCIR-4 QAC2. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 402–405, June 2004.

[14] M. Walker, M. Iida, and S. Cote. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232, 1994.