Question Answering Experiments at NTCIR-5: Acquisition of Answer Evaluation Patterns and Context Processing using Passage Retrieval

Yuichi Murata *Tomoyosi Akiba*
Department of Information and Computer Sciences, Toyohashi University of Technology
1-1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, 441-8580, JAPAN
akiba@cl.ics.tut.ac.jp

Atsushi Fujii
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Katunobu Itou
Graduate School of Information Science, Nagoya University
1 Furo-cho, Nagoya, 464-8603, JAPAN

Abstract

In our participation in QAC3, our Question Answering system developed for QAC2 is extended in two ways, separately. The first system is constructed to improve the performance of answer evaluation. The automatic lexico-syntactic pattern acquisition from large corpora and the method to incorporate the patterns into QA system are developed and evaluated. The second system is constructed to implement the ability of context processing for information access dialogue (IAD), which is a main target of QAC3 evaluation. The system exploits passage retrieval for selecting an appropriate context from the history of the series questions, in order to compose a complete question.

1 Introduction

Our QA system [3] was initially constructed at the time of QAC2 evaluation. In our participation in QAC3, the system was extended in two ways from different points of view. Because the extensions were implemented separately and has not yet merged in a single system, we participated in QAC3 by using two different systems.

The first system, which is used as the run for QAC3 labeled 'sys1', extends the original QA system to improve the answer evaluation performance. For the purpose, the automatic lexico-syntactic pattern acquisition from large corpora and the method to incorporate the patterns into QA system were developed and tested.

The second system, which is used as the runs for QAC3 labeled 'sys2' and 'sys3', was constructed to implement the ability of context processing for information access dialogue (IAD), which is a main target of QAC3 evaluation. The system exploits passage retrieval for selecting an appropriate context from the history of the series questions, in order to compose a complete question.

Section 2 describes the issues about first extension, that is automatic acquisition of lexico-syntactic patterns from large corpora. Section 3 described the issues about second extension, that is context processing using dynamic passage retrieval.

2 Automatic Lexico-Syntactic Pattern Acquisition for Answer Evaluation

In TREC and NTCIR, a question often contains the word or phrase that directly express the semantic category for the correct answer. For example, the question *"kokumin eiyo shou wo jushou shita eiga kantoku wa dare desu ka?"* (Who is the film director received the national honorary prize?) implies that the answer should be an instance of *"eiga kantoku"* (film director). Since the answer candidate *"kurosawa akira"* is an instance of the QF, it is likely to be a correct answer. We shall call these words (or phrases) representing the

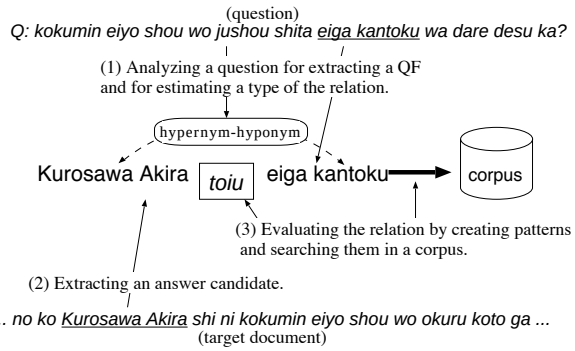

Figure 1. The process of evaluating the hypernym-hyponym relation between QF and an answer candidate.

semantic categories for the correct answers "Question Focus (QF)".

We previously proposed the method that evaluates those relations that stand between a QF and a correct answer in a data-driven method using unorganized corpora as knowledge resources [3]. For the answers that are name expressions, which are focused on in this paper, the hypernym-hyponym relation between the QF and each answer candidate (AC) is examined [1]. Figure 1 depicts the process of the method. (1) Question analysis is performed for extracting the QF *"eiga kantoku"*(film director) from the question. The analysis also estimates the relation that stands between the QF and a correct answer is hypernym-hyponym, as the question requires a name expression as its answer. (2) From the target documents obtained by passage retrieval, an answer candidate *"Kurosawa Akira"* is extracted. (3) The pattern *"Kurosawa Akira toiu eiga kantoku"* (film director such as Kurosawa Akira) is created using the extracted QF and the candidate according to the estimated relation, and it is searched if it is appeared anywhere in a large corpus. If the pattern is found in the corpus, the candidate will be promoted, while it will be undervalued if not found.

In the step (3) above, the lexico-syntactic patterns are used, e.g., *"AC toiu QF"* (QF such as AC), *"AC igaino QF"* (QF other than AC), *"QF · AC"*, in which *QF* and *AC* are surface expressions of a QF and an answer candidate, respectively. In our previous work, the twelve patterns were constructed manually.

For our method of answer evaluation, the selection of the lexico-syntactic patterns is crucial for the performance of question answering. In this work, automatic pattern acquisition is investigated, in order to improve the performance of our method.

[1]For the answers that are numerical expression, the value of the number and the unit expression appeared in the answer candidate is examined with respect to the QF. See [3].

2.1 Related Work

A lot of works initiated by Hearst [7] try to acquire the word pairs in some semantic relation from text corpora. Fleischman et al. [5] presented the method to extract the concept-entity pairs of words from text corpora. Girju et al. [6] presented the method to extract the pair in part-whole relations.

With respect to the hypernym-hyponym relations, Ando et al. [4] extracted the pattern for evaluating the hypernym-hyponym relation automatically in order to acquire the word pairs in that relation. Though our method is similar to their work, we focused on acquiring the patterns aiming for question answering: we tried to extract the patterns effective for question answering and applied the acquired pattern for evaluating the relation between the QF and an answer candidate on the fly.

2.2 Process of Lexico-Syntactic Pattern Acquisition

Our method for extracting the lexico-syntactic patterns consists of the following three steps.

1. Extract a set of hypernym-hyponym word pairs as seed patterns from an existing language resource [1].
2. Extract lexico-syntactic pattern candidates from a large corpus using the seed patterns.
3. Evaluate the performance for each patterns, and select the subset of the patterns as the final set of lexico-syntactic patterns.

2.2.1 Extracting Seed Patterns

We selected 72,878 hypernym-hyponym word pairs from the EDR dictionaries [1]. The pairs are used as seed patterns to extract the lexico-syntactic patterns in the following steps.

2.2.2 Extracting Candidate Patterns

For each word pairs, the sentences that have both two words are extracted from the newspaper articles. Then, the common patterns in the set of sentences are extracted as the candidate patterns.

The method of finding the common patterns is as follows. The extracted sentences are morphologically analyzed. For each sentence, the hypernym noun and the hyponym noun are labeled with *hypernym* and *hyponym*, respectively. The other words in the sentence are labeled with their POS tags.

The subsequence in the sentence between *hypernym* and *hyponym* are investigated to find the common patterns. We introduced the following constraints to restrict the pattern extracted.

Table 1. The candidates of lexico-syntactic patterns.

Patterns	Examples	Precision
[Y] igai-no [X]	[Banana] igai-no [Kudamono]	0.569
[Y] nado [X]	[Dosei] nado [Wakusei]	0.621
[Y] toiu [X]	[Sakura] toiu [Hana]	0.450
[Y] wa [X]	[Inu] wa [Doubutu]	0.250
[X] [Y]	[Megane] [Rougankyou]	0.217

- Surface expression of the nouns are discarded and always expressed by the POS tag *noun* in a pattern.
- The number of words between *hypernym* and *hyponym* must be less than 3.

The 57 patterns were extracted by the process mentioned above. Table 1 shows the example of extracted patterns.

2.2.3 Selecting Lexico-syntactic Patterns from the Candidates

The automatic process mentioned above extracts various candidate patterns: some pattern is reliable, while another is not. For example, the pattern "*hyponym* igai no *hypernym*" is reliable for evaluating the hypernym-hyponym relation but does not expected to appear frequently in corpora. On the other hand, the pattern "*hypernym* no *hyponym*" appears frequently in corpora but is not so reliable, because the pattern is often used for expressing the other relations, e.g. the ownership, than the hypernym-hyponym relation.

In order to select good patterns for answer evaluation of question answering, we performed the performance evaluation for each candidate patterns by using the test collection of hypernym-hyponym word pairs, called "development set".

Development set

The training set of hypernym-hyponym candidates called "development set" was constructed by using the QAC1 DryRun and QAC1 ReferenceRun collections. Firstly, the 527 questions that ask about a name expression and have QF were selected from the collections. For each question, its QF was extracted manually.

Secondly, each question extracted was submitted to the QA system to obtain the twenty answer candidates. Though the system has the ability to evaluate the answer candidates by the hypernym-hyponym relation with QF described in this section, we did not use this ability when obtaining the candidates. We shall call this baseline system that does not used the ability *baseline*.

Table 2. Four categories of the judgment.

	Judgment results	
	Correct	Incorrect
Positive examples	CA	CD
Negative examples	IA	ID

Finally, we collected the word pairs and assign the label "positive" or "negative" to each pair, to construct the final set. For each question, the pairs of its QF and each correct answer were added to the positive examples of the "development set". On the other hand, the pairs of the QF and each of the other answer candidates obtained by the QA system were added to the negative examples. Note that the correct hypernym-hyponym pair is not always added to the positive examples by this process, because there might be a candidate that is the hyponym of the QF but is not the correct answer of the corresponding question. The adopted this process in order to reduce the cost of the development.

As the result, we collected 9272 pairs for the development set that consists of 986 positive examples and 8286 negative examples. [2]

Evaluating the performance of each candidate

The pattern is used to judge whether or not hypernym-hyponym relation stands between a given pair of words (or phrases). The part labeled *hypernym* and *hyponym* are replaced by the given words (or phrases) respectively, and the resulting pattern is investigated if it appears in a large text corpus. If the pattern is found in the corpus, the pair is judged to be in hypernym-hyponym relation. If it is not found, the pair is judged not to be in the relation.

The effective patterns for checking hypernym-hyponym relation should be selected from the candidates. For each pattern, how correctly it can judge the relation against the pairs in the development set was investigated. Each example in the development set is divided into four categories against a lexico-syntactic pattern as shown in Table 2.

As evaluation measures, recall and precision are calculated as follows.

$$Precision = \frac{CA}{CA + IA} \quad (1)$$

$$Recall = \frac{CA}{CA + CD} \quad (2)$$

[2] Because some questions have less than 20 candidate answers, total number of test pairs are less than 20 times of number of questions.

Selecting the effective patterns

We used the precision as the measure to select the effective patterns. The lexico-syntactic patterns that have the precision above a given threshold were selected. The value of the threshold was investigated changing from 0.1 to 0.5 at 0.1 intervals. Table1 shows some examples of lexico-syntactic patterns with their precisions.

2.3 Experimental Evaluation

In order to examine the effectiveness of the acquired lexico-syntactic patterns, two experimental evaluations were performed. The one evaluates the performance of the acquired lexico-syntactic patterns directly. The other evaluates the performance indirectly by applying them to question answering task.

2.3.1 Evaluating the performance of hypernym-hyponym judgment

Test data

Using QAC1 subtask 1 FormalRun test collection, we constructed the test data for evaluating the performance of acquired lexico-syntactic patterns. The test data consist of word or phrase pairs, each of which is tested if they are in hypernym-hyponym relation.

The test data were constructed as follows. Firstly, we selected the questions that request name expressions for their answers. The 153 out of 200 questions (including four questions with no answer) in QAC1 ask for name expression as their answers. From each of those questions, we manually extracted a correct QF that was a hypernym of the correct answer(s). We excluded the QF that was a synonym of the correct answer(s) in this experiments. We could extract the QFs from 121 out of 153 questions (79.1%).

Secondly, for each QF, the hyponym candidates were collected. One group of the candidates was the correct answers denoted in QAC1 test collection. The other group was the answer candidates obtained by submitting the corresponding question to the *baseline* QA system. Twenty candidates were extracted for each QF.

Finally, the extracted word or phrase pair was judged by human whether or not they were in hypernym-hyponym relation. After removing duplications, there obtained 1642 pairs, in which 980 pairs were correct (in hypernym-hyponym relation) and 662 were incorrect (not in hypernym-hyponym relation).

Applying the patterns independently

We investigated two methods to apply the acquired lexico-syntactic patterns to the task of hypernym-hyponym relation judgment. The first method applies the patterns independently, referred as *automatic acquisition*: it decides a test pair to be positive when in it finds the text corpus any one of the filled patterns that the places labeled *hypernym* and *hyponym* are replaced by the corresponding words in the pair.

Table 3. The Performance of hypernym-hyponym judgment.

Method	Threshold	#patterns	Precision	Recall
manually constructed	-	10	0.906	0.249
automatic acquisition	0.5	7	0.944	0.035
	0.4	10	0.920	0.096
	0.3	15	0.911	0.270
	0.2	24	0.830	0.359
	0.1	26	0.772	0.461
	0.0	57	0.773	0.466
automatic acquisition with SVM	0.3	15	0.917	0.219
	0.2	24	0.919	0.213
	0.1	26	0.908	0.227
	0.0	57	0.791	0.328

The evaluation measures are precision and recall, which are defined by the equation (1) and (2), respectively. We compared the performance of acquired patterns with that of the pattern manually constructed [3], referred as *manually constructed.*

The result are shown in Table 3. It indicates that the pattern acquired with the threshold 0.3 is best performed. It performed better than *manual* with respect to both precision and recall.

Applying the patterns integratedly

The second method integrates the results of judgments obtained by each pattern, referred as *automatic acquisition with SVM*. The hypernym-hyponym judgment is considered as a binary classification problem. Each judgment result by a pattern is treated as a feature for the classification problem. We used Support Vector Machine (SVM) to solve the problem. The development set is used as the training data for training the SVM.

The result is shown in 3. The result indicates that it improves the precision, while it degrades the recall.

2.3.2 Evaluating the performance of question answering

The acquired lexico-syntactic patterns were applied to our QA system and the performance was investigated. We report the results with respect to QAC1 and QAC3 test collections.

QAC1

The 153 out of 200 questions in QAC1 ask for name expression as their answers. Among them, the 121

Table 4. The performance of question answering with respect to QAC1.

Method	Threshold	#Correct	#+	#-	MRR
baseline	-	89	-	-	0.431
manual	-	88	25	11	0.498
automatic	0.4	85	28	12	0.426
	0.3	87	27	13	0.491
automatic (SVM)	0.3	87	26	13	0.476
automatic	0.4	89	25	10	0.500
+ manual	0.3	88	28	12	0.491

out of the 153 questions have QFs. The question answering performance for the 121 questions was investigated. The Mean Reciprocal Rank (MRR) is used as the evaluation measure.

We compared five systems. The *baseline* system does not use the hypernym-hyponym answer evaluation between QF and each candidate. The other four systems use the answer evaluation. They differ in the lexico-syntactic patterns used for the evaluation and in the way to use them.

The system labeled *manual* used the patterns constructed manually, which is the same system used for QAC2 [3]. The systems labeled *automatic* and *automatic (SVM)* used the patterns automatically acquired by using the method described in this paper. The system labeled *automatic* applies the patterns independently, as the first method described in section 2.3.1. The system labeled *automatic (SVM)* applies the patterns integratedly by using SVM, as the second method described in section 2.3.1. The system labeled *manual + automatic* used the both patterns and applies them independently.

Table 4 shows the results. The column "#Correct" indicates the number of questions correctly answered by the method. The column "#+" and "#-" indicate the number of questions increased and decreased the MRR as compared with *baseline*.

The system using automatic acquired patterns did improve *baseline*, but did not improve *manual*. This is partly because the patterns used in *manual* are associated with their confidence score assigned by human, while the patterns in *automatic* are not. The score is used to promote the answer candidate successfully evaluated its hypernym-hyponym relation with the QF by the pattern.

The system using both patterns performed best among them, in which the confidence score of the automatic acquired patterns are assigned to the unique value equal to the minimum value of the score among *manual* patterns.

Table 5. The performance of question answering with respect to QAC3 (MMF1).

System	Total	First	Rest
QAC3 sys1	0.137	0.224	0.124
+ Bugfixed	0.184	0.272	0.170
+ noISA	0.188	0.272	0.175
QAC3 sys2	0.193	0.286	0.180
+ noISA	0.188	0.272	0.174

Table 6. The performance of question answering with respect to QAC3 Reference Run 1 (MMF1).

System	Total	First	Rest
QAC3 sys1	0.177	0.224	0.169
+ Bugfixed	0.237	0.272	0.232
+ noISA	0.243	0.272	0.239
QAC3 sys2	0.256	0.286	0.251
+ noISA	0.245	0.272	0.241

QAC3

The QA performance was evaluated with respect to QAC3 test collection. Our first system used as the run for QAC3 (referred as *sys1*) has been extended to use the automatic acquired lexico-syntactic patterns for answer evaluation and to apply the patterns integratedly by using SVM at run time as described in this section, which corresponds to *automatic (SVM)* in Table 4. However, the system was found to be buggy in the process of string extraction from the target documents. Therefore, the bug-fixed version of the system (referred as *BugFixed*) was also evaluated. We also evaluated the system that did not use the answer evaluation for hypernym-hyponym relation to see the effect of the method (referred as *noIsa*).

We also compared them with the second system used as the run for QAC3 (referred as *sys2*), which corresponds to the *baseline* system described in Section 3. The system uses the patterns extended from QAC2 system (corresponding to *manual* in Table 4) by manually adding the several patterns found by the method described here. Though the *sys2* system is developed separately from *sys1* and additional improvements were performed, the system can be said to correspond roughly to the *automatic + manual* system in Table 4.

The performance was measured by MMF1 [8]. Table 5 shows the results with respect to QAC3 formal-run. The result indicates that the use of answer evaluation is less effective for QAC3 than for QAC1. We

think one of the reasons seems to be because QAC3 consists of the series of questions and the extraction of QF from the question in the series requires context processing.

Table 6 shows the results with respect to QAC3 reference run 1, in which the context of each question is complemented correctly. With respect to *QAC3 sys2*, the hypernym-hyponym evaluation was more effective for the reference run than for the formalrun. However, the reverse is true with respect to *QAC3 sys1*. We will investigate the details about the difference in the performance of our method between QAC1 and QAC3, to improve the method in the future work.

3 Context Processing using Dynamic Passage Retrieval

Information Access Dialogue (IAD) task have been evaluated in the recent NTCIR QAC series, specifically QAC2 subtask3 and QAC3. IAD task assumes the situation in which users interactively collect information using a QA system. The QA Systems aiming at the task need the abilities of context processing. Suppose the following series of questions

Q1 *"Whose monument was displayed at Yankees Stadium in 1999?"*

Q2 *"When did he come to Japan on honeymoon?"*

Q3 *"Who was the bride at that time?"*

The second question *Q2* can be answered by selecting the fragments *"Joe DiMaggio"* that is the answer to the first question and composing the complete question *"When did Joe DiMaggio come to Japan on honeymoon?"* Similarly, the third question *Q3* can be answered by selecting appropriate fragments from the previous questions and their answers (*"Joe DiMaggio"* and *"come to Japan on honeymoon"*) and composing the complete question. If the fragments is selected incorrectly, e.g. *"Yankees Stadium"* and *"1954"*(the answer of the second question), the resulting complete question is useless, rather harmful, to find the correct answer. Therefore, this can be seen as a problem of ambiguity resolution, and can be resolved by selecting the fragments from the history in order to compose the most appropriate question.

We propose the method of resolving the disambiguation problems in context processing by exploiting passage retrieval. The basic idea of the proposed method is as follows. Suppose an input question has at least one correct answer in the target document collection, there must be at least one similar passage in it. Therefore, the similarity with some passage in the target documents can be used to select the appropriate context from the history of the questions.

We previously applied this idea to N-best rescoring of spoken questions returned by a large vocabulary continuous speech recognition (LVCSR) decoder [2]. The experimental results showed that the method considerably improve both the speech recognition accuracy and the QA performance in speech-driven QA task.

3.1 Dynamic Passage Retrieval

A passage, i.e. a text fragment in target documents, is used to calculate the similarity against the question. Some systems use a sentence as a passage, while other systems use a paragraph. The longer the size of a passage is selected, the more candidates of the answer can be picked up. It raises the recall of the answer, while it reduces the precision because the more incorrect candidates are also picked up. Developing a good passage retrieval method is one of the common research topics for question answering [9].

We have proposed a dynamic passage retrieval method [3]. The method selects an appropriate size of the passage on the fly by using F-measure based similarity with the question. We recast the method below.

Let $C(s)$ be a set of passage candidates with respect to a sentence s in the target documents. [3] Though $C(s)$ can include any size of text fragments surrounding s theoretically, only the following sentences are considered in our implementation if each of them should be included in the passage.

s_{-1}**:** the sentence immediately before s.

s_{+1}**:** the sentence immediately after s.

h_A**:** the headline of the article A that s belongs.

d_A**:** the date string of the article A that s belongs.

Therefore, we adopted the following candidate $C(s)$ in practice.

$$C(s) = \{\{s\} \cup E | E \in 2^{\{s_{-1}\ s_{+1}\ h_A\ d_A\}}\}$$

The proposed method selects a best passage $\hat{c}$ from $C(s)$ by using following F-measure based similarity $F(q,c)$ with a question q,

$$\hat{c} = \operatorname*{argmax}_{c \in C(s)} F(q,c) \tag{3}$$

$$F(q,c) = \frac{(1+\beta^2)P(q,c)R(q,c)}{\beta^2 P(q,c) + R(q,c)} \tag{4}$$

$$P(q,c) = \frac{\sum_{t \in T(q) \cap T(c)} \mathrm{idf}(t)}{\sum_{t \in T(c)} \mathrm{idf}(t)},$$

$$R(q,c) = \frac{\sum_{t \in T(q) \cap T(c)} \mathrm{idf}(t)}{\sum_{t \in T(q)} \mathrm{idf}(t)}$$

[3] More specifically, the candidates should be considered with respect to an answer candidate a. However we approximate a to be identical with s that includes a.

where $T(c)$ is a set of terms included in c and $idf(t)$ is the inverse document frequency (IDF) of a term t. We chose $\beta = 2$ to emphasize the recall for the selection for question answering, while $\beta = 1$ for the context processing explained below.

The passage retrieval score $S_{\text{passage}}(q)$ is defined as the max value of $F(q, s)$ with respect to the target document collection D.

$$S_{\text{passage}}(q) = \max_{s \in D} \max_{c \in C(s)} F(q, c) \qquad (5)$$

Again, we cannot examine all of sentences in D because of the computational cost, only the sentences included in the documents that a document retrieval engine returns by submitting q are examined to calculate the equation (5) for an approximation.

3.2 Formulation of Context Processing for IAD task

The third question *Q3* of the last example of IAD task can be combined with any set of the text fragments extracted from the history of the series of questions and their answers, and formed a candidate of the appropriate question, e.g. *"Joe DiMaggio, come to Japan on honeymoon, Who was the bride at that time?"* is one of the candidates, while *"Yankees Stadium, 1954, Who was the bride at that time?"* is another. Suppose the passage *"Joe DiMaggio and Marilyn Monroe went to Japan for their honeymoon."* is found, the first candidate is more likely to be the appropriate question because of the higher similarity between the candidate and the passage.

This context processing problem is formulated as follows. Let $H(q_i)$ be a history of a question q_i, which is a set of text fragments appeared in either a previous questions $q_1 \cdots q_{i-1}$ or their answers $a_1 \cdots a_{i-1}$. Any unit can be used for the text fragment that correspond an element of $H(q)$: it can be a word $w \in q_1 \cup \cdots \cup q_{i-1} \cup a_1 \cup \cdots \cup a_{i-1}$, or a sentence $s \in \{q_1 \cdots q_{i-1}\ a_1 \cdots a_{i-1}\}$. In the following, we use a sentence s as the unit.

Giving a question q and its history $H(q)$, A candidate of the complete question of q is composed by adding a set of text fragments in the history $h \in 2^{H(q)}$ to q, i.e. $h \cup q$. Now, the problem of context processing is defined as selecting the best context $\hat{h} \in 2^{H(q)}$ that compose the best complete question $\hat{h} \cup q$. The proposed method try to solve this problem by maximizing the passage retrieval score $S_{\text{passage}}(h \cup q)$ as follows.

$$\hat{h} = \operatorname*{argmax}_{h \in 2^{H(q)}} S_{\text{passage}}(h \cup q) \qquad (6)$$

The computational cost of calculating the equation (6) exactly gets higher with the size of $H(q)$, because all of the combinations of the elements in $H(q)$ must be compared. Therefore, we introduced the approximation to (6): we restricted the context to $\tilde{H}_{QA}(q_i) = \{q_1\ q_{i-1}\ a_1\ a_{i-1}\}$. We also exclude the case with no context. Those result in the following equation (referred as *HQA* in our experiment).

$$\hat{h} \approx \operatorname*{argmax}_{h \in 2^{\tilde{H}_{QA}(q)} - \{\phi\}} S_{\text{passage}}(h \cup q) \qquad (7)$$

Including the answers $a_1 \cdots a_{i-1}$, which are returned by the system, in $H(q)$ seems harmful, because they may be incorrect. Usually in many QA systems including ours, the string exactly appears in the question is not considered as an answer candidate. Therefore, if the system outputs an incorrect answer that is accidentally same with a future question in the same series, it will not be possible to return the correct answer to the future question. For this reason, we restricted the context to $\tilde{H}_Q(q_i) = \{q_1 q_{i-1}\}$ and introduced the following equation for context selection, (referred as *HQ*)

$$\hat{h} \approx \operatorname*{argmax}_{h \in 2^{\tilde{H}_Q(q)} - \{\phi\}} S_{\text{passage}}(h \cup q) \qquad (8)$$

$$= \operatorname*{argmax}_{h \in \{\{q_1\}\{q_{i-1}\}\{q_1 q_{i-1}\}\}} S_{\text{passage}}(h \cup q) \qquad (9)$$

As a baseline, the method using the fixed context $\tilde{h} = \{q_1\ q_{i-1}\}$ was investigated (referred as *baseline*).

Table 7. QA performance differences according to the context processing methods (MMF1).

Method	Total	First	Rest	Gather	Browse
baseline	0.193	0.286	0.179	0.222	0.125
HQA	0.169	0.286	0.151	0.188	0.125
HQ	0.194	0.286	0.180	0.216	0.143

3.3 Experimental Results

The experiment was performed by using QAC3 test collection. The performances of the three methods above were compared by the evaluation measure MMF1 [8]. The results were shown in Table 7.

The difference of the performance according to the categories of questions was investigated. The label **Total**, **First**, and **Rest** correspond to the entire questions, the first questions in each series, and the second and the latter questions in each series, respectively. The QAC3 test set consists of 35 series of the gathering type and 15 series of the browsing type. The difference according to the type was also investigated. The label **Gather** and **Browse** correspond to the gathering type and the browsing type, respectively.

The result showed that the proposed method did not improve the baseline method with respect to entire test set (**Total**): the performance of *HQA* was less than *baseline*, while the performance of *HQ* is almost equal to *baseline*. However, the performances of them were quite different according to the type of series. *HQ* outperformed *baseline* with respect to the browsing type of series (**Browse**). This result indicated that the method was effective for the browsing type, in which the context processing plays more important role than in the gathering type.

3.4 Discussion

We formulated the context processing in IAD task as a problem of context selection from previous questions and answers to compose an appropriate complete question, and proposed a novel method for the problem exploiting passage retrieval. The method uses only term statistics for context processing instead of conventional NLP such as anaphora resolution. Since the current implementation of the method is naive, we think some refined implementation can improve the performance further. The combination of our method and the conventional NLP method will be also hopeful.

4 Conclusion

In this paper, we presented two directions of extending our previous QA system. The first system was constructed to improve the performance of answer evaluation. The automatic lexico-syntactic pattern acquisition from large corpora and the method to incorporate the patterns into QA system were developed and tested. The second system was constructed to implement the ability of context processing for information access dialogue (IAD). The system exploits passage retrieval for selecting context from the history of the series questions, in order to compose the complete question. The extensions were implemented separately for our participation in QAC3. We would like to merge the results in a single system in the future work.

References

[1] *EDR Home Page*. `http://www2.nict.go.jp/kk/e416/EDR/`.

[2] T. Akiba and H. Abe. Exploiting passage retrieval for n-best rescoring of spoken questions. In *Proceedings of International Conference on Speech Communication and Technology (Eurospeech)*, pages 65–68, 2005.

[3] T. Akiba, A. Fujii, and K. Itou. Question answering using "common sense" and utility maximization principle. In *Proceedings of The Fourth NTCIR Workshop*, 2004. `http://research.nii.go.jp/ntcir/workshop/OnlineProceedings4/QAC/NTCIR4-QAC-AkibaT.pdf`.

[4] M. Ando, S. Sekine, and S. Ishizaki. Automatic extraction of hyponyms from japanese newspapers using lexico-syntactic patterns. In *Proceedings of International Conference on Language Resources and Evaluation*, pages 387–390, 2004.

[5] M. Fleischman, E. Hovy, and A. Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 1–7, 2003.

[6] R. Girju, A. Badulescu, and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 80–87, May 2003.

[7] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of International Conference on Computational Linguistics*, pages 539–545, 1992.

[8] T. Kato, J. Fukumoto, and F. Masui. An overview of NTCIR-5 QAC3. In *Proceedings of NTCIR Workshop 5*, 2005.

[9] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of ACM SIGIR*, pages 41–47, 2003.