NTT DATA Question-Answering Experiment at the NTCIR-3 QAC

Toru Takaki Yoshio Eriguchi
Research and Development Headquarters
NTT DATA Corporation
Kayabacho Tower Bldg., 1-21-2, Shinkawa,
Chuo-ku, Tokyo 104-0033 Japan
{takakit,eriguchiy}@nttdata.co.jp

Abstract

In this paper, we give an overview of our question-answering system for the NTCIR-3 QAC. Our system is based on an information-retrieval technique and an information-extraction technique by pattern matching. The system has three main stages: question analysis, passage retrieval, and answer extraction. In the passage-retrieval stage, two types of retrieval method are applied sequentially to narrow down the document quantity for the following answer-extraction stage. We have submitted our results for all three sub-tasks of the NTCIR-3 QAC official runs.

Keywords: *question answering, information retrieval, information extraction, passage retrieval, question analysis, pattern matching.*

1 Introduction

Our participation in the NTCIR-3 Question and Answering Challenge (QAC) represents NTT DATA's first participation in NTCIR.

The question-answering approach - an application of "real" information retrieval rather than "document" retrieval - is a promising means of providing a user with required information precisely and efficiently. The goal of question-answering is that a system returns a concise package of information that answers the user's question through knowledge acquisition from a data source on the Internet or an intranet. The Text REtrieval Conference (TREC) has set out an English question-answering track each year from TREC-8 in 1999 to TREC-2002 in 2002 [6,7]. We participated in the TREC question-answering track in 1999 and 2000, and have built an English question-answering system [4,5].

The NTCIR-3 QAC is the first evaluation workshop concerning large-scale question-answering using Japanese [2]. We rebuilt our question-answering system for the Japanese language, and have submitted the results for all three sub-tasks (task 1, 2, and 3) of the NTCIR-3 QAC.

To perform these tasks, we constructed a question-answering system by combining a traditional information-retrieval technique and an information-extraction technique. In this paper, we describe the processing of our QA system, and discuss and analyze the evaluation results we obtained for the NTCIR-3 QAC official runs.

2 System overview

This section describes the processing of our QA system. The system was built by combining a fundamental information-retrieval system and an information-extraction system. The QA procedure consists of three main components. First, we will explain the processing for each of these components and then will explain the task-oriented processing for the sub-tasks.

The processing procedure is shown in Fig. 1. For this system, we used only the Mainichi Newspapers (1998-1999) in the NTCIR-3 QAC document set as an information source; other sources, such as an encyclopedia or external Web data, were not used.

(1) Question-analysis component

This component determines the answer categories that match the inputted question.

(a) Answer-categories definition

The answer categories are defined using a three-level hierarchical structure. In the top-level of the structure, where the answer categories have abstract answer types, we defined five categories: (1) Noun, (2) Non-noun, (3) Quantity, (4) Time, and (5) Unknown. In the lower levels of the structure, the categories are given more detailed answer-type definitions. For example, second-level categories under the Noun cate-

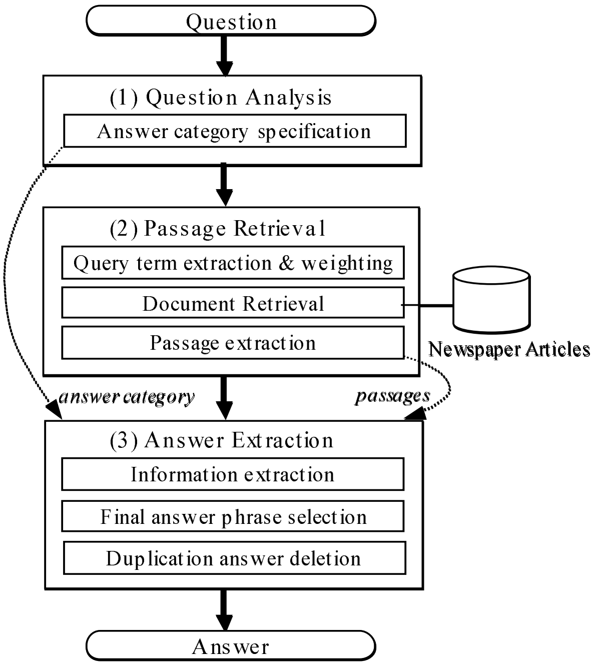

Figure 1: Processing Procedure

gory are Person, Organization, Structure, Location, etc. An example of answer-type categories is shown in Fig. 2.

(b) Answer-category specification

The answer-category specification is a processing step in which a determination is made as to what type of answer is required for a given question. Characteristic expressions in the question sentence are extracted using the pattern-matching engine and matched to corresponding answer types. The pattern-matching engine uses manually created rule patterns that are defined by a combination of a morphological character sequence and a part of speech [1].

When a question sentence is matched with a pattern, the answer category is determined by referring to a table that defines the correspondence between the pattern and a category, and a category score is given for the pattern.

If the answer category of a detailed lower-level category is given, the categories of related higher-level categories are also given as next-candidate types with the lower category's score. If a question does not match any of the patterns, an "Unknown" category is given.

(2) Passage-retrieval component

This component extracts the candidate passages containing answer phrases from the newspaper articles in the data set. It performs the query-term-extraction, document-retrieval, and passage-extraction processing.

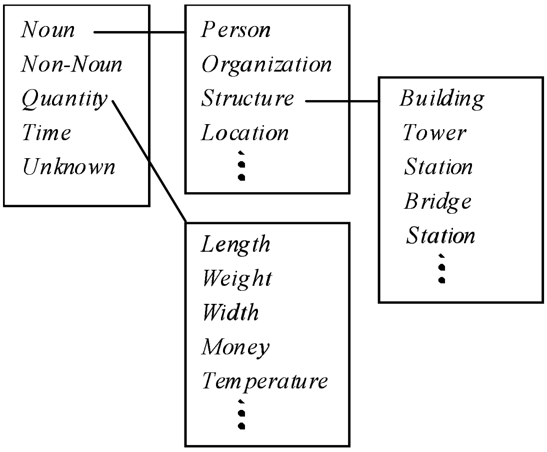

Figure 2: Answer categories

(c) Important-question-phrase extraction

In this processing, the interrogative phrase is removed from the original question sentence.

(d) Query-term extraction

Query terms for document retrieval are extracted from the above question sentence. We implemented two types of query-term-extraction sub-component: **QTE1** and **QTE2**.

QTE1 All the combination of sequential morphemes is extracted as query terms. In some cases, query terms may become quite long. Although the number of query terms becomes sum of 1 to n when the number of morphemes is n, we used only twenty selected terms with a low document frequency as query terms.

QTE2 Only terms of a morpheme's combination with specific parts of speech, such as a noun and an out-of-vocabulary word, are extracted as query terms.

QTE2 is a query-term-extraction method that is applied as part of the general-information retrieval. Through the **QTE1** method, a newspaper article that includes a longer query term can be searched with a higher score.

(e) Query-term weighting

When the parts of speech of the query terms are a proper noun and an out-of-vocabulary word, a higher score is given to the query terms.

(f) Document Retrieval

Using the extracted query terms and their scores, the system searches newspaper articles in the database to find documents that include the question's answer phrase. We did a relevance ranking of the articles using the BM25 probabilistic retrieval formula [3], and used the ten top-ranked documents for the subsequent processing.

(g) Passage extraction

Candidate passages that may include the answer are specified and extracted from the top-ranked documents obtained in the previous step. Each passage is given a score that depends on the importance of the query terms and the degree of concentration with which the terms appear [4].

Passages with a score above a set threshold become candidate passages.

(3) Answer-extraction component

The answer-extraction component extracts a phrase that matches the answer category from the obtained candidate passage, and outputs the final answer phrase.

(h) Information extraction

A phrase that belongs to the answer categories given by the answer-categories specification is extracted from the candidate passage. We use the same pattern-matching engine as is used in the answer-categories specification to extract an answer phrase. The answer-extraction-pattern rules for each answer category were created manually.

When the answer category is "Unknown", a proper noun is generally extracted as the answer phrase.

The extracted answer-candidate phrases are given scores that are calculated using the answer category's score and the passage score. Thus, even identical phrases can have different phrase scores depending on the extracted candidate passage.

(i) Final-answer phrase selection

This is the processing that determines which phrases will be output from among the extracted answer-candidate phrases as a final answer. We implemented two types of selection sub-components: **ANS1** and **ANS2**.

ANS1 The output order of the answer phrases is based on the score given to each phrase in each passage. The same answer phrase is output only once.

ANS2 The scores of an identical answer phrase that appears in different passages are summed and the output order of the answer phrases is based on the summed scores.

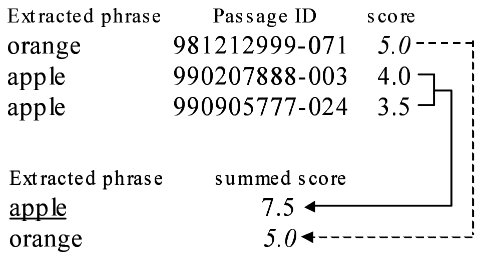

Figure 3: Phrase selection

An example of phrase selection is shown in Fig. 3. Three answer phrases are extracted in this example. The phrase "apple" is extracted from two separate passages.

When using the **ANS1** method, which outputs in the order of the answer phrase score given for each passage, the first answer phrase would be "orange" and the second would be "apple". On the other hand, with the **ANS2** method, where the scores of an identical answer phrase are summed, the first answer phrase would be "apple".

(j) Duplication-answer deletion

The system does not output the same answer phrase within the question sentence.

(4) Task-oriented component

In addition to the processing components explained above, we implemented task-oriented processing components. We explain each component here with respect to its sub-task characteristics.

(k) Passage retrieval by phrase structure (PRPS)

We prepared another passage-retrieval component to implement a passage-retrieval method that retrieves similar passages with regard to the relationships between terms determined through syntactic analysis. The syntax of the question sentence and each newspaper article is analyzed using the Kurohashi-Nagao Parser (KNP). This method was applied to task 1 and task 2.

(l) Query-term extraction for series questions (QTESQ)

Task 3 is question answering for a series of questions that are assumed to be continuously input.

Table 2: Evaluation results for task 1 for NTCIR-3 QAC official questions

Run name	MRR	#Q at answer rank						#Q	#Q	#Q
		1st	2nd	3rd	4th	5th	not found	Best	≥ Ave	> Ave
NDQAS11	0.362	53	23	10	8	4	97	70	91	80
NDQAS12	0.313	47	19	5	6	7	111	63	75	64
NDQAS13	0.404	63	22	7	7	3	93	79	102	91
NDQAS14	0.436	70	21	9	3	4	88	84	107	96

Table 3: Evaluation results for tasks 2 and 3 for NTCIR-3 QAC official questions

Run name	FM	#Q at FM range						#Q	#Q	#Q
		≥0.8	≥0.6	≥0.4	≥0.2	>0	=0	Best	≥ Ave	> Ave
NDQAS21	0.149	7	16	24	61	62	138	40	87	10
NDQAS31	0.144	0	0	0	17	18	22	14	30	0

Table 1: Applied components

Task No	Run name	Applied components
1	NDQAS11	**QTE1+ANS1**
1	NDQAS12	**QTE1+ANS1+PRPS**
1	NDQAS13	**QTE1+ANS2**
1	NDQAS14	**QTE2+ANS2**
2	NDQAS21	**QTE1+ANS1+PRPS**
3	NDQAS31	**QTE1+ANS1+QTESQ**

For such a task, it is necessary to use the information obtained for previous questions to get answers. In our system, we use query terms that combine the terms extracted from the present question and the terms extracted from the previous question.

3 Experimental Results

In this section, we discuss the results of experimental evaluations that used the data set of the NTCIR-3 QAC formal run.

3.1 Experimental conditions

In each task, the question-answering processing of the formal run's question was performed by choosing appropriate processing components. The list of applied components is shown in Table 1. When the number of NDQAS12 output answers was less than five, the output of NDQAS11 was added behind the NDQAS12 answers.

3.2 Experimental results

The numbers of questions for the evaluation of tasks 1, 2, and 3 were 195, 200, and 40, respectively. (Although the number of released questions for task 1 was 200, five questions that had no correct answer were excepted from the evaluation.) The evaluation measures were the mean reciprocal rank (MRR) in task 1, and the F-measure (FM) in tasks 2 and 3. The evaluation results for task 1 are shown in Table 2, and those for tasks 2 and 3 are shown in Table 3.

We submitted two results for task 1 (NDQAS11 and NDQAS12) and one result for each of tasks 2 and 3 (NDQAS21 and NDQAS31, respectively). Each task's average evaluation value for the formal run was 0.303 (task 1, MRR), 0.141 (task 2, FM), and 0.107 (task 3, FM). The questions where there was a large difference in the evaluation value between our results and the average are shown in Table 4 (best) and Table 5 (worst).

3.3 Processing result for each component

Here, we analyze the processing of each component based on the result for NDQAS11 in task 1.

(1) Question-analysis component

In the answer-category specification, 126 of 200 questions were correctly assigned to a category other than "Unknown" in task 1. The number of questions assigned to the "Unknown" category was 55 (28%).

Many of the "Unknown" category questions had a question expression such as "〜は何ですか" (*What is/are ...?*) and "〜は何といいますか" (*What is/are ... called?*). The correct category for these questions could not be determined through the interrogative pattern rule only, and the answer category-judging pattern for these questions was not yet defined in our system.

Another question expression "〜はどこですか" also led to many errors. The meaning of such a question

Table 4: Best results and questions

	Q No.	MRR			Question
		NDQAS11	Ave.	Diff.	
1	1067	1.000	0.100	0.900	*What is absolute zero in centigrade?*
2	1105	1.000	0.167	0.833	*What are the sequels to J.K. Rowlings Harry Potter and the Sorcerers Stone?*
3	1172	1.000	0.217	0.783	*Who is the TV personality who got married saying that she felt a trembling beam, beep, beep, beep?*
4	1030	1.000	0.232	0.768	*What is the name of the government agency that oversees financial businesses and that became independent from the Ministry of Finance?*
5	1180	1.000	0.236	0.764	*What is the name of the publishing house that publishes a magazine titled Shukan Nichiroku 20 Seiki (Weekly Daily Records of the 20th Century)?*
		NDQAS12	Ave.	Diff.	
1	1104	1.000	0.089	0.911	*With which company did the automatic transmission manufacturer Jatco, whose shares owned by Matsuda were all purchased by Nissan Motor, merge?*
2	1055	1.000	0.113	0.887	*For which television drama did Koki Mitani write his first script?*
3	1170	1.000	0.217	0.783	*Where is the first store which Costco, a major US supermarket chain, opened in Japan?*
4	1172	1.000	0.217	0.783	*Who is the TV personality who got married saying that she felt a trembling beam, beep, beep, beep?*
5	1154	1.000	0.222	0.778	*What planet is Europa a satellite of?*

cannot be specified from only the expression, and a common problem was that questions asking for an "Organization" name were categorized as "Location" questions in our system.

(2) Passage-retrieval component

In the passage-retrieval component, after the number of articles was narrowed down in the document-retrieval stage, more detailed passage-selection processing was done. We analyzed whether documents or passages that included a correct answer were correctly searched in the document-retrieval and passage-extraction stages.

Figure 6 shows the number of questions for which we could retrieve a document/passage that included the correct answer at a cutoff point when N top-ranked document/passages were output. In the experiment, the candidate passage was chosen from the top ten documents from the result of the document-retrieval stage and for 176 questions (90.3%) we could retrieve a document that included the correct answer. Moreover, in the passage extraction, the rates at cutoff points of 10 and 100 passages were about 80% and 90%, respectively. Thus, the passage-retrieval component worked well for most questions.

(3) Answer-extraction component

The answer-extraction stage extracted a correct answer phrase for 168 questions, but the answer phrases for 71 questions were given a low phrase score, so these questions were not output within the final answer.

Furthermore, seven questions had errors caused by a shortage of extraction patterns or an error regarding the extraction character sequence length. For example, when the correct answer was "9.79 seconds", the system output only "9 seconds".

For 13 other questions, the system output answer phrases that had the same meanings within the set of final answer to each question (e.g., "Tokyo Disneyland" and "TDL").

When NDQAS11 was compared with NDQAS13 using the other answer extraction component, NDQAS13 output the exactly correct answer phrase. This suggests that the use of phrase information obtained from many passages enables more correct answering of the questions.

4 Summary

In this paper, we have explained the processing of our question-answering system, and briefly discussed and analyzed our evaluation results. In the NTCIR-3 QAC, we applied a question-answering approach based on a combination of an information-retrieval technique and an information-extraction technique.

Our experimental results suggest that applying a conventional information-retrieval technique allowed us to effectively extract documents and passages that included a correct answer. However, questions can vary in many ways, and our system did not respond adequately with regard to our present answer-category-specification module to some questions. The large-scale Japanese question-answering test was built for the first time for the NTCIR-3 QAC, so further development of the question-answering technology is expected in the future.

Table 5: Worst results and questions

	Q No.	MRR			Question
		NDQAS11	Ave.	Diff.	
1	1058	0.000	0.817	-0.817	*Who are the Japanese laureates for the Nobel Prize for Physics?*
2	1136	0.000	0.816	-0.816	*When will the Law for Recycling of Specified Kinds of Home Appliances become effective?*
3	1125	0.000	0.766	-0.766	*In which two countries was the FIFA World Cup 2002 held?*
4	1119	0.000	0.764	-0.764	*Which country owns the space station Mir?*
5	1114	0.000	0.734	-0.734	*Who bombed Yugoslavia?*
5	1131	0.000	0.734	-0.734	*Which nuclear powers signed the Protocol to the Treaty on the Southeast Asia Nuclear-Weapon-Free Zone?*
		NDQAS12	Ave.	Diff.	
1	1131	0.000	0.734	-0.734	*Which nuclear powers signed the Protocol to the Treaty on the Southeast Asia Nuclear-Weapon-Free Zone?*
2	1037	0.000	0.689	-0.689	*When did the i-mode services start?*
3	1079	0.000	0.566	-0.566	*Which world chess champion had a match against IBMs Deep Blue in 1997?*
4	1100	0.000	0.545	-0.545	*When was iMac first launched in Japan?*
5	1128	0.000	0.485	-0.485	*Who won the French Open Tennis Womens Singles Championship after an interval of three years?*

Table 6: Accuracy of Document retrieval and Passage selection

Cutoff point at x documents/passages	#Q including relevant document/phrase at cutoff point	
	Document retrieval	Passage selection
1	119 (61.0%)	83 (42.6%)
2	137 (70.3%)	110 (56.4%)
3	150 (76.9%)	126 (64.6%)
4	158 (81.0%)	139 (71.3%)
5	163 (83.6%)	142 (72.8%)
10	176 (90.3%)	157 (80.5%)
20	185 (94.9%)	168 (86.2%)
30	185 (94.9%)	172 (88.2%)
50	189 (96.9%)	174 (89.2%)
100	191 (97.9%)	178 (91.3%)

References

[1] Y. Eriguchi, T. Kitani: NTT Data Description of the Erie System Used for MUC-6, In *Proceedings of Tipster Text Program (Phase II)*, pp. 469-470, 1996.

[2] J. Fukumoto and T. Kato: An Overview of Question and Answering Challenge (QAC) of the Next NTCIR Workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, pp.375-377, 2001.

[3] S. E. Robertson and S. Walker: Okapi/ Keenbow at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pages 151-161, 2000. NIST Special Publication 500-246.

[4] T. Takaki: NTT DATA: Overview of system approach at TREC-8 ad hoc and question answering. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pages 523-530, 2000. NIST Special Publication 500-246.

[5] T. Takaki: NTT DATA TREC-9 Question Answering Track Report. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, pages 399-406, 2001. NIST Special Publication 500-249.

[6] E.M. Voorhees, The TREC-8 Question Answering Track Report. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pages 77-82, 2000. NIST Special Publication 500-246.

[7] E.M. Voorhees: Overview of the TREC-9 Question Answering Track. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, pages 71-79, 2001. NIST Special Publication 500-249.