NTT DATA Question-Answering Experiment at the NTCIR-4 QAC2

Toru Takaki
Research and Development Headquarters
NTT DATA Corporation
takakit@nttdata.co.jp

Abstract

In this paper, we give an overview of our question-answering system for the NTCIR-4 QAC2. Our system is based on an information-retrieval technique and an information-extraction technique based on pattern matching. The system has three main stages: question analysis, passage retrieval, and answer extraction. We have submitted our results for all three sub-tasks in the NTCIR-4 QAC2 official runs.
Keywords: *question answering, information retrieval, information extraction, passage retrieval, question analysis, pattern matching, answer mining.*

1 Introduction

Our participation in the NTCIR-4 Question and Answering Challenge (QAC2) was NTT DATA's second effort, and our question-answering system for this was almost the same as for the first, the QAC1 [2,5], where we combined a traditional information-retrieval and an information-extraction technique. In this paper, we describe the processing in our QA system, and the evaluation results we obtained in the NTCIR-4 QAC2 official runs.

2 System overview

This section describes the processing in our QA system, which was achieved by combining a fundamental information-retrieval and an information-extraction system. The QA procedure consisted of three main components. We will first explain the processing for each of these components and will then explain the task-oriented processing for the sub-tasks.

The processing procedure is outlined in Fig. 1. We only used Mainichi and Yomiuri newspapers (1998-1999) in the NTCIR-4 QAC2 document set as the information source in this system; other sources, such as encyclopedias or external Web data, were not used.

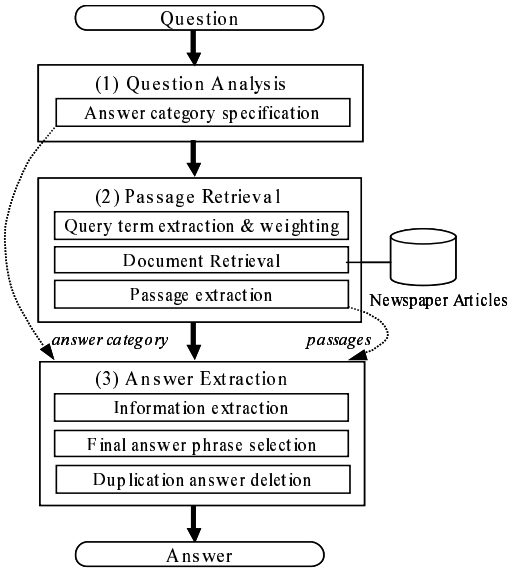

Figure 1: Processing Procedure

(1) Question-analysis component

This component determined the answer categories that matched the inputted question.

(a) Answer-categories definition

The answer categories were defined using a three-level hierarchy. We defined five categories in the top level, where the answer categories were abstract: (1) Noun, (2) Non-noun, (3) Quantity, (4) Time, and (5) Unknown. The categories were given more detailed answer-type definitions in the lower levels. For example, second-level categories under the Noun category were Person, Organization, Structure, and Location. There are some answer-type categories listed in Fig. 2.

(b) Answer-category specification

The answer-category specification was a processing step where what type of answer was required for a

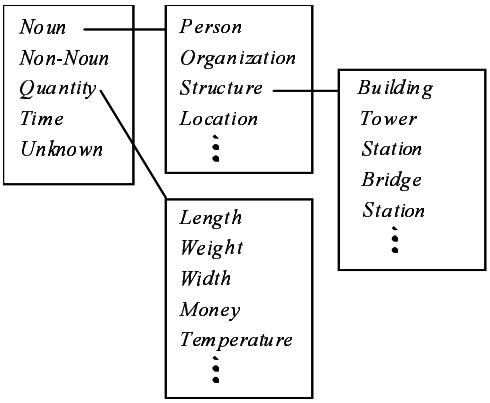

Figure 2: Answer categories

given question was determined. Characteristic expressions in the question sentence were extracted using a pattern-matching engine and matched to corresponding answer types. The pattern-matching engine used manually created rule patterns that were defined by a combination of a morphological character sequence and a part of speech [1].

When a question was matched with a pattern, the answer category was determined by referring to a table that defined the correspondence between the pattern and the category, and a category score was awarded for the pattern.

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

(2) Passage-retrieval component

This component extracted candidate passages containing answer phrases from the newspaper articles in the data set. It extracted query terms, retrieved documents, and processed extracted passages.

(c) Important-question-phrase extraction

The interrogative phrase was removed from the original question in this process.

(d) Query-term extraction

Query terms for document retrieval were extracted from the question.

(e) Query-term weighting

When the parts of speech of the query terms were a proper noun and an out-of-vocabulary word, a higher score was awarded to the query terms.

Extracted phrase	Passage ID	Score
orange	981212999-071	5.0
apple	990206777-003	4.0
apple	990905555-024	3.5

Extracted phrase	Total Score
apple	7.5
orange	5.0

Figure 3: Phrase selection

(f) Document Retrieval

The system searched newspaper articles in the database using the extracted query terms and their scores to find documents that included 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 in the subsequent processing.

(g) Passage extraction

Candidate passages that may have included the answer were specified and extracted from the top-ranked documents obtained in the previous step. Each passage was awarded a score that depended on the importance of the query terms and the degree of concentration with which the terms appeared [4].

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

(3) Answer-extraction component

The answer-extraction component extracted a phrase that matched the answer category from the obtained candidate passage and outputted the final answer phrase.

(h) Information extraction

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

When the answer category was "Unknown", a proper noun was generally extracted as the answer phrase. The extracted answer-candidate phrases were awarded scores that were calculated using the answer category's score and passage score. Thus, even identical phrases could have different phrase scores depending on the extracted candidate passage.

Table 1: Evaluation results for Subtask 1 of NTCIR-4 QAC2 official questions

Run name	MRR	#Q at answer rank						#Q	#Q
		1st	2nd	3rd	4th	5th	Not found	≥ Ave	> Ave
R11	0.405	65	14	14	5	5	94	87	84
R12	0.397	63	14	14	8	4	94	84	81

Table 2: Evaluation results for Subtask 2 of NTCIR-4 QAC2 official questions

Run name	Output	Correct	Recall	precision	MF
R21	928	159	0.245	0.171	0.229
R22	1480	225	0.347	0.152	0.219

Run name	#Q at MF range						#Q	#Q	#Q
	≥0.8	≥0.6	≥0.4	≥0.2	>0	=0	Best	≥ Ave	> Ave
R21	12	16	27	26	23	95	45	95	81
R22	7	11	24	41	41	75	32	99	88

(i) Final-answer phrase selection

This as the process that determined which phrases would be output from the extracted answer-candidate phrases as the final answer. The scores for an identical answer phrase appearing in different passages were totaled and the output order of the answer phrases was based on the total scores. There are examples of phrase selection in Fig. 3, where three answer phrases have been extracted. The phrase "apple" has been extracted from two separate passages.

When not using our method, which outputs the order of the answer phrase score given for each passage, the first answer phrase would have been "orange" and the second would have been "apple". With our method, where the scores for an identical answer phrase were totaled, the first answer phrase was "apple".

(j) Duplication-answer deletion

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

(4) Task-oriented component

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

(k) Determination of the number of answer

Subtasks 2 and 3 were evaluated with a F-measure. The QA system needed to determine the number of answers to output. Our QA system determined the number of answers with the ratio of phrase scores. When the ratio of the n-th phrase's score and the $n+1$-th score was larger than the threshold, the system only outputted n answers.

We also implemented another method to determine the cutoff point for outputting answers by the answer type given during answer-categories specification processing. When a question was given an "Unknown" category, the system could not determine what kind of phrase should have been outputted. Although a proper noun was generally extracted as the answer phrase, the possibility that the answer would be mistaken was higher. We applied the few-answer (FA) method to the "Unknown" categories' question where the system restricted the number of answers.

(l) Query-term extraction for series questions

Subtask 3 involved answering questions in a series that were assumed to have been continuously input. Here, it was necessary to use the information obtained from previous questions to obtain answers. We used query terms in our system that combined the terms extracted from the present question and those extracted from the previous.

3 QAC2 Results

We used 197 questions to evaluate Subtask 1, 199 for Subtask 2, and 251 for Subtask 3. Although the number of released questions for Subtasks 1 and 2 was 200, questions that had no correct answer were exempted from the evaluation. The evaluation measures were the mean reciprocal rank (MRR) for Subtask 1, and the Mean F-values (MF) for Subtasks 2 and 3. We submitted two results for each Subtask. The run names are **R11** and **R12** for Subtask 1, **R21** and **R22** for Subtask2, and **R31** and **R32** for Subtask 3

Table 3: Evaluation results for Subtask 3 of NTCIR-4 QAC2 official questions

Run name	Output	Correct	Recall	precision	MF
R31	672	45	0.083	0.067	0.095
R32	912	59	0.109	0.065	0.099

Table 4: Best and worst results and questions (R11 for Subtask 1)

	Q No.	MRR			Question
		R11	Ave.	Diff.	
Best-1	10175	1.000	0.130	0.870	*In response to a critical accident at JCO, an Accident Countermeasure Headquarters consisting of the Minister for Science and Technology as chief was established at 3 o'clock in the afternoon. What time was the Government Task Force for the Accident, headed by Prime Minister Obuchi, established? What is absolute zero in centigrade?*
Best-2	10174	1.000	0.219	0.781	*An earthquake occurred in Taiwan on September 21, 1999. What was the magnitude of the earthquake?*
Best-3	10148	1.000	0.263	0.737	*What is the general term for a car that runs 100 kilometers on three liters of fuel?*
Best-4	10075	1.000	0.286	0.714	*How many people died in the Great Hanshin Earthquake?*
Best-5	10173	1.000	0.328	0.672	*There was a strong earthquake in Taiwan on September 21, 1999. What time did it occur?*
Worst-1	10059	0.000	0.720	-0.720	*Who did Kobayashi Asei file a suit against for an infringement of copyright?*
Worst-2	10180	0.000	0.648	-0.648	*Who were the successive presidents of Indonesia?*
Worst-3	10066	0.000	0.640	-0.640	*Who was the first person to succeed in swimming across the Atlantic Ocean?*
Worst-4	10023	0.000	0.613	-0.613	*Who is the landowner of the Kitora Ancient Tomb?*
Worst-5	10090	0.000	0.611	-0.611	*When did the luxury passenger liner Titanic sink?*

in this paper. The difference between **R11** and **R12** was the value of the category score. The few-answer (FA) method was applied to **R21** and **R31** for the "Unknown" categories' questions, while this was not applied to **R22** and **R32**. The evaluation results for Subtasks 1, 2 and 3 are listed in Tables 1, 2 and 3. Questions with large differences in the evaluation values between our results and the average are listed in Table 4 for Subtask 1.

In the answer-category specification, 111 of 197 questions were correctly assigned to a category other than "Unknown" in Subtask 1. Seventy questions were assigned to the "Unknown" category was (36%). The MRR was 0.493 for questions assigned to the correct category for Subtask 1, and was 0.281 for the "Unknown" category questions.

4 Summary

We explained processing in our question-answering system, and briefly discussed and analyzed the evaluation results. We applied a question-answering approach to the NTCIR-4 QAC2 based on the combination of an information-retrieval and an information-extraction technique.

References

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

[2] J. Fukumoto, T. Kato and F. Masui: Question and Answering Challenge (QAC-1): Question answering evaluation at NTCIR workshop 3, *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2002.

[3] S. E. Robertson and S. Walker: Okapi/ Keenbow at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 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 and Y. Eriguchi: NTT DATA Question Answering Experiment at the NTCIR-3 QAC, *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2002.