NTCIR-3 QAC Experiments at Matsushita

NOMOTO, Masako nomoto@trl.mei.co.jp
SATO, Mitsuhiro msato@trl.mei.co.jp
SUZUKI, Hiroyuki suzuki@trl.mei.co.jp
Multimedia Systems Research Laboratory, Matsushita Electric Industrial Co., Ltd.
4-5-15, Higashi-Shinagawa, Shinagawa-ku, Tokyo 140-8632 JAPAN

Abstract

This paper investigates our experimental results for NTCIR-3 QAC1, the first attempt to evaluate the technology of Japanese question answering (QA). Our basic approach is a combination of passage retrieval and named entity (NE) extraction based on pattern matching. The results show that the accuracy of NE extraction crucially affects the overall performance of our system. Additional experiments prove the effects of refinements of passage retrieval and NE extraction.

We also analyze the QAC1 test collection to identify features relevant for measuring the difficulty of the questions in the collection. Based on the analysis, we make some proposals for the future QAC tasks, as regards to answer categories, technical aspects, and definition of the tasks.

Keywords: *question answering (QA), named entity extraction, pattern matching, passage retrieval*

1. Introduction

Question answering (QA) represents a promising alternative approach to information retrieval. Using information extraction techniques, it can directly pinpoint answers and reduce the costs of searching the information from documents.

The TREC question answering tracks [1], started in 1999 (TREC-8), have focused on English QA.

The NTCIR-3 QAC1 [2] is the first attempt to evaluate the technology of Japanese QA. It differs from TREC in that it requires the exact answer for each question and allows answer expressions that do not exist in the given documents and are generated using other information sources such as encyclopedia.

We participated in NTCIR-3 QAC1 tasks. Our QA system (MEI QA system) aims at processing large-scale dynamic data such as web pages. We take a shallow approach based on a combination of passage retrieval and named entity (NE) extraction using pattern matching. No pre-processing is performed except for indexing. The basic approach we used in each task (task1, 2) is essentially the same, and we deal here with our results in task1.

Section 2 gives the overview of our system. Section 3 analyzes our results in QAC1 task1. Section 4 reports the results of additional experiments to improve the performance of our system.

In section 5, we analyze the QAC1 test collection to identify features relevant for measuring the difficulty of the questions in the collection. Section 6 makes some proposals for the next QAC tasks based on the analysis of section 5.

2. System Descriptions

2.1 The Architecture

The basic architecture, we believe, is typical of most of the participating QA systems, as shown in Figure 1.

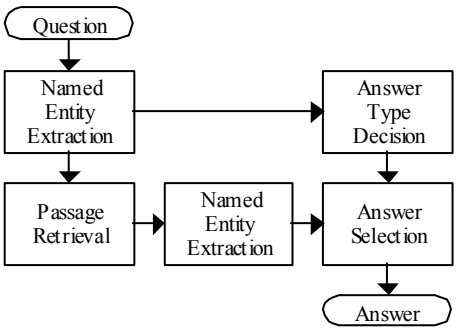

Figure 1:Architecture of the MEI QA system

The processing steps of our system are the followings:

(1) The NE extraction module annotates an input question with named entity categories.
(2) The passage retrieval module extracts keywords from the annotated question and retrieves top *n* ranking passages.
(3) The NE extraction module annotates the retrieved passages with NE categories.
(4) The answer type decision module decides on the type of the questions and adequate answer category.
(5) The answer selection module scores each NE in the passages that match the answer type and selects an answer.

Passage retrieval and index pre-processing are performed using the *MEISTER* software libraries, which has been used in our IR systems in NTCIR-1 and 2 [3] [4]. The NE extraction was developed from the NE tool in IREX NE task [5] using hand created matching rules.

2.2 Methods

2.2.1 Passage Retrieval Module

The passage retrieval module features the following methods:

-A paragraph is defined as a passage.

-*Coordination Level Scoring* (*CLS*)[3] to rank retrieved passages, among which top 30 passages are used.

2.2.2 NE Extraction Module

The NE extraction module annotates questions and retrieved passages with NE category tags using pattern matching rules (178) and dictionaries.

We defined 29 tags, of which 11 basic tags are shown below:

DATE, TIME, PERCENT, MONEY, PERSON, LOCATION, ORGANIZATION, ARTIFACT, PERIOD, FREQ, and QUANT.

The first 8 tags follow the IREX NE task [5]. ARTIFACT is used as a default category in our system, and includes miscellaneous NEs that are not classified in other categories. DATE, TIME, MONEY, and LOCATION may also have subclasses. In addition, multiple category tags, such as PERSON_OR_ORGANIZATION, are used for the NEs that may belong to plural categories and are not determined the adequate one from the context.

2.2.3 Answer Type Decision Module

The answer type decision module determines the type of answer category, using 30 pattern matching rules. When no rule matches, the module uses ARTIFACT as a default. Examples of matching rules are shown below.

[Answer category]		[Rule]
ORGANIZATION	←	*doko*. *(*hatsubai*|*kiyou*|*gappei*|…)
		Where.*(sell| appoint| merge)
DATE	←	*nan* (*nen*|*gatsu*|*niti*)
		What(year|month|day)

For example, the question,

"*Jatco wa doko to gappei shima shita ka,*"
Jatco where with merge did

meaning, "With which company did Jatco merge?", matches the first of the above rules, and the answer category is referred to as ORGANIZATION.

2.2.4 Answer Selection Module

The answer selection module selects answers from the answer candidates. The answer candidates are the NEs that are annotated with the answer category tag in the retrieved passages. The score of each candidate NE *s*(*NE*) is calculated by the following formula:

$$s(NE) = \sum_{w \in Q} amb(NE)\,kwne(w)\{D_{\max} - dist(NE, w)\} + \{R_{\max} - rank(psg)\} \cdots\cdots (1)$$

where,

amb(*NE*) = 1/2: if *NE* is tagged with a multiple category tag

= 1: otherwise

kwne(*w*) = 2: if $w \in NE_q$ (NE_q : a set of NEs extracted from the question)

= 1: otherwise

dist(*NE*,*w*) = min(distance between *NE* and *w*, D_{max},) (bytes)

rank(*psg*) = the rank of the retrieved passage that includes the *NE*.

Values of constants are:

$R_{max} = 30$, and $D_{max} = 50$.

The answer candidates are ranked based on the scores calculated by the above method. The top 5 NEs are selected as the final answers of the question.

3. Formal Run Results and Analysis

Table 1 shows the result of task1[1].

Table 1. Task1 results

MRR	Q1 (RQ1)	Q5 (RQ5)
0. 387	61 (0. 313)	98 (0. 503)

MRR: Mean Reciprocal Rank, defined as the sum of RR devided by the number of questions

RR: Reciprocal Rank, defined as the inverse number of the highest rank among those of correct answers

Q1(RQ1) The number of questions that the system answered correctly in :the first rank (the rate of Q1)

Q5(RQ5) The number of questions that the system answered correctly in :up to the fifth rank (the rate of Q5)

MRR (Mean Reciprocal Rank) is a formal measure for evaluating performance in the task. The MRR in Table 1 suggests that for the averaged question in task1 we can include the correct answer in top 3 ranking. On the other hand, RQ5 says we could not include correct answer in top 5 ranking in about half of the questions of the task.

Table 2 gives the number of questions for which each module made errors. The errors on passage retrieval module are classified in 2 levels.

Table 2. The errors made by each module

Module	# of questions
Passage Retrieval	21
(Document Level)	6
(Passage Level)	15
NE Extraction	48
Answer Type Decision	9
Answer Selection	19
Total	97

NE extraction is most problematic, and about half of the errors occurred at this module. The passage

[1] Our system for task1 had a few bugs. The results in Table 1 are slighly better than the official one due to the bug fixes.

retrieval module retrieved relevant documents for most of the questions, but missed the relevant passages in the documents for 15 questions. The types of answers were correctly determined in most cases, but the answer selection failed to select correct answers for 19 questions.

Table 3 gives the failure ratio of NE extraction for each answer category of the questions, as classified by the answer type decision module.

Table 3. Failure ratio of NE extraction for each answer category

Answer category	# of Questions	NE failure(%)
ARTIFACT	66	20 (30. 30)
DATE	14	1 (7. 14)
FREQ	1	0 (0. 00)
LOCATION	31	7 (22. 58)
MONEY	3	0 (0. 00)
ORGANIZATION	17	6 (35. 29)
PERCENT	3	0 (0. 00)
PERIOD	4	1 (25. 00)
PERSON	39	7 (17. 95)
QUANT	17	6 (35. 29)
total	195	48 (24. 62)

The failure ratio of ORGANIZATION, QUANT, and ARTIFACT was higher than others.

Errors for ORGANIZATION might due to the lack of corresponding entries in the dictionary. It seems difficult to construct the dictionary for the category with sufficient coverage for various purposes.

Failures on QUANT (quantity) could be the lack of pattern definition. QUANT consists of a numerical expression and a measurement noun, such as "4.5 *kiro guramu* (4.5 kilograms)". QUANT has a wide variety of measurement nouns, and the definition of the pattern for the category costs higher than other numerical categories.

About 34% of the questions are classified as ARTIFACT by the answer type decision module. As the category is used as a default, it may include NEs that should have been classified otherwise. The fact that a considerable number of questions are classified as ARTIFACT implies the lack of categories. Finer grained classification scheme is needed for a precise error analysis.

4. Experimental Results

Based on the error analysis in the previous section, we made attempts to improve the performance of the system. Below, we discuss what results are for our attempts.

4.1 NE Extraction Rules

As the result of error analysis implied the possibility of lack of pattern definitions or matching rules, we revised them as follows:

- addition of 39 pattern definitions that consist of 121 expressions on QUANT (66), NUMBER (26), DATE (26), and FREQ (3)
- modification and addition of 33 matching rules on PERSON (20), LOCATION (10), and so on.

The results using the revised pattern definitions and rules are shown in Table 4.

Table 4. Effects of revision of rules

Rule	MRR	Q1 (RQ1)	Q5 (RQ5)
Formal Run	0. 387	61 (0. 313)	98 (0. 503)
Revised	0. 399	63 (0. 323)	102 (0. 523)

MRR and Q5 (RQ5) improved, though the increase of Q1 was modest.

Further analysis demonstrated that rule set modification is effective for numerical expressions. For example, in questions that required QUANT type answers, NE failure rate decreased from 35. 3% to 17. 6%.

4.2 Definition of a Passage

As shown in Table 2, the passage retrieval module could not retrieve adequate passages in the relevant documents for 15 questions. The result may be due to the definition of passage.

We run experiments with the following alternative definitions of passages.

- **A:** a paragraph of a document
- **B:** the headline and a paragraph of a document
- **C:** a segmentation of a document divided by special symbols (the marker of headings, etc. ex. squares) or the limit of maximum of length (1,024bytes)
- **D:** a document

Table 5 shows the results.

Table 5. Effects of using various definitions of the passage

Psg	AVL	RelP (%)	ESA (%)	MRR	Q1 (RQ1)	Q5 (RQ5)
A	145	174 (89.2)	19 (10.9)	0. 387	61 (0. 313)	98 (0. 503)
B	217	181 (92.8)	19 (10.5)	0. 395	62 (0. 311)	99 (0. 508)
C	353	186 (95.4)	24 (12.9)	0. 404	67 (0. 344)	99 (0. 508)
D	1098	191 (97.9)	28 (14.7)	0. 408	66 (0. 388)	98 (0. 503)

AVL: average passage length (bytes)

RelP: # of questions for which a relevant passages was retrieved in top 30

ESA: # of questions for which a correct answer was NOT selected from relevant passages of top 30 (cf. RelP)

The longer the average length of a passage was, the more easily the relevant passages were retrieved and MRR were improved. However, Q5 did not improved in all trials. This could be caused by the difficulties with answer selection in longer passages, as shown in the column of ESA.

4.3 Number of Passages

Table 6 shows the number of questions in which up to top *n* rankings relevant passages were retrieved.

Table 6. Performance of passages retrieval

ranking of relevant passages	1	~ 5	~ 10	~30
# of questions	92	141	160	174
ratio (%)	47	72	82	89

The passage retrieval module ranked relevant passages within top 10 most of the time.

We also compared the number of retrieved passages used for selecting answers, as shown in Table 7.

Table 7. The effect of modifying # of passages used for answer selection

# of passages	MRR	Q1 (RQ1)	Q5 (RQ5)
top 5	0. 349	53 (0. 272)	90 (0. 462)
top 10	0. 393	62 (0. 318)	99 (0. 508)
top 30	0. 387	61 (0. 313)	98 (0. 503)

Though we used top 30 passages for the formal run, all the results with top 10 passages are slightly better than what we got for top 30. As for the RR for each question, going from 30 to 10 passages, it was found, increases RR for 15 questions, while decreasing the RR for 8 of the questions.

4.4 Discussions

The error analysis above revealed that NE extraction module is most problematic in our system. As shown in the results of experiments, refining the rules for NE extraction effectively improved the performance of our system.

Another possible refinement of NE extraction would be to increase the number of entries listed in dictionaries. Indeed, prior to the formal run, we added about 66,400 entries to the person dictionary. Contrary to our expectation, the increase in performance was found not so impressive, perhaps because the additions of family names, many of which are also used as NEs of other categories could have hurt precision. What this suggests for us is that some care must be exercised on features of NEs to be added. If they include NEs that could be used in other categories, a scheme for disambiguation should also be worked out.

The passage retrieval module at the document level, worked well for the task. Further analysis showed that among top 30 passages retrieved by the module, there was one passage from relevant documents for 97% of the questions, while among top 10 passages, a relevant passage was found for 91 % of the questions.

The module still has problems at passage level. Experiments show that the definition of a passage seriously affects MRR, but it is almost impossible to give a definition that could work for various purposes. Another problem is how we might determine the optimal number of passages the system retrieves.

At any rate the discussion so far suggests that we look into new approaches to passage retrieval, so that relevant passages are placed higher in the rank. For example, when the number of relevant passages is small, such passages could be easily identified by some keywords in the question.

5. Analysis of the QAC1 test collection

In this section, we will look at answer categories for the QAC1 test collection and identify some features of the questions that make them difficult or easy to answer.

To see how difficult or easy each question of the test collection is for systems participating in the task1, we consider *RR(AVG)*, or the average of the RR(reciprocal rank)s of all the systems, given as the following:

$$RR(AVG) = (AvgSys5 * N(Sys\#5)) / N(SysAll) \cdots\cdots (2)$$

where,

AvgSys5 : The average of the RRs of the systems that obtained more than zero in RR.

N(*Sys#5*) : the number of the systems that obtained more than zero in RR,

N(*SysAll*): the number of all the systems participated.

In the following, *MRR(AVG)*, the averaged RR(AVG)s for a set of questions, refers to the averaged performance of all the systems.

5.1 Categories of Answers

In response to the analysis of errors in section 3, which suggests the need for a finer grained classification scheme of answer categories, we formulated a new classification scheme for answers so as to cover the 195 questions used in the task1. We defined 8 basic categories and 27 sub categories.

Table 8 shows the number of questions and the performance of systems for each answer category.

For 75% of the questions, an answer is one of the following categories:

ARTIFACT, PERSON, LOCATION, NUMBER.

Hard questions the MEI system and the average system failed on are those that require answer categories of the following types:

(basic categories)

ASTRO, and LIVING_THINGS.

(sub categories)

LOCATION: NATURE, NUMBER: PERCENT, LIVING_THINGS: PLANT, and ORGANIZATION: OTHER.

The following categories were difficult for AVG:

(sub categories)

LOCATION: PREFECTURE, and TIME: PERIOD.

Easy categories for AVG and MEI were:

(basic categories)

PERSON, and LOCATION.

Table 8: # of questions and performance of systems for each answer category

Answer Categories	# of Questions	MRR (AVG)	MRR (MEI)
PERSON	42	0. 36	0. 41
PERSON: JAPANESE	31	0. 35	0. 40
PERSON: FOREIGN	11	0. 37	0. 43
ARTIFACT	44	0. 28	0. 44
ARTIFACT: PRODUCT_CLASS	6	0. 29	0. 50
ARTIFACT: PRODUCT_NAME	6	0. 33	0. 58
ARTIFACT: WORK	10	0. 32	0. 53
ARTIFACT: OTHER	22	0. 25	0. 34
LIVING_THINGS	8	0. 20	0. 04
LIVING_THINGS: ANIMAL	1	0. 27	0. 00
LIVING_THINGS: PLANT	5	0. 14	0. 00
LIVING_THINGS: OTHER	2	0. 34	0. 17
ASTRO	2	0. 16	0. 00
LOCATION	32	0. 33	0. 44
LOCATION: COUNTRY	13	0. 41	0. 66
LOCATION: STATE	1	0. 43	0. 00
LOCATION: PREFECTURE	3	0. 08	0. 33
LOCATION: CITY	3	0. 32	0. 33
LOCATION: CAPITAL	3	0. 51	0. 33
LOCATION: TOWN	2	0. 28	0. 00
LOCATION: SPOT	5	0. 27	0. 50
LOCATION: NATURE	2	0. 08	0. 00
ORGANIZATION	20	0. 27	0. 28
ORGANIZATION: COMPANY	13	0. 27	0. 28
ORGANIZATION: POLITICS	3	0. 40	0. 33
ORGANIZATION: SPORTS	2	0. 21	0. 50
ORGANIZATION: OTHER	2	0. 14	0. 00
NUMBER	29	0. 28	0. 39
NUMBER: NUMBER	3	0. 20	0. 33
NUMBER: PERCENT	2	0. 13	0. 17
NUMBER: QUANT	21	0. 31	0. 40
NUMBER: MONEY	3	0. 30	0. 57
TIME	18	0. 31	0. 43
TIME: DATE	14	0. 35	0. 44
TIME: PERIOD	4	0. 18	0. 38

5.2 Causes for failure

Below we go through a component by component analysis of what caused failures or poor performance on some of the questions.

We start by dividing the questions into 3 groups based on the number of systems that output at least one correct answer in task1, as shown below:

Q_EASY: 8-15 systems
Q_MOD: 3-7 systems
Q_DIF: 0-2 systems

All questions in Q_DIF are listed in the APPENDIX with the answer categories.

5.2.1 Document Retrieval

Let us look at the following features of questions, which we believe may influence performance on document retrieval:

RelD (the average number of relevant documents), and

RKey (the average ratio of keywords appeared in a relevant document to those extracted from the question).

Table 9 shows the results on RelD and RKey.

Table 9: Difficulty of Document Retrieval

	RelD	**RKey**	MRR (AVG)	MRR (MEI)
Q_EASY	8. 492	0. 773	0. 531	0. 700
Q_MOD	5. 242	0. 762	0. 249	0. 318
Q_DIF	2. 333	0. 694	0. 051	0. 026
Total	5. 744	0. 752	0. 303	0. 387

Note that the RelD declines dramatically as we go from Q_EASY to Q_DIF. This shows that the number of relevant documents has an impact on performance of QA systems, and thus serves as a potential indicator of how hard a question is.

We find a similar pattern in the behavior of RKey, though not as obvious as RelD.

5.2.2 Passage Retrieval / Selection

Next we turn to features on passage retrieval and look at how they affect system's overall performance in QA tasks.

Here we invoke two notions PSG_EASY and PSG_NO_EASY to discriminate between easy and hard tasks in passage retrieval. If many keywords in a question appear in relevant passages in a document, and not in non-relevant passages, relevant passages are supposed to be easily distinguished from non-relevant ones (PSG_EASY). On the contrary, if many keywords appear in non-relevant passages and not in relevant ones, it would be difficult to identify relevant ones(PSG_NO_EASY).

We call a question that allows easy passage retrieval a "Q_PSG_EASY" question, and the other quesitons, a "Q_PSG_NO_EASY" question. We divided all the questions into the two groups, by the following steps. Here, a paragraph is used as a passage.

For a given document that contains the correct answer of the question:

- segment the document into passages $psg_1, \ldots psg_n$
- classify all passages into 2 groups:

  CRP: set of passages that include the correct answer

  ICP: set of passages that do not include the correct answer

- calculate $kwdnum(psg_i)$ of each passage

  where $kwdnum(psg_i)$ is the number of keywords, which were extracted from the question and appeared in the passage psg_i

- calculate $PX = max_{j \in ICP}\ (kwdnum(psg_j))$
- count *ncg* : number of passage $psg_x \in NCG$

  where $psg_x \in NCG$ if $psg_x \in CRP \wedge kwdnum(psg_x) > PX$

- count *ncl* : number of passage $psg_y \in NCL$

  where $psg_y \in NCL$ if $psg_y \in CRP \wedge kwdnum(psg_y) \leq PX$

- the question is classified as **Q_PSG_EASY** if *ncg* > *ncl*

If no such document is found, the question is classified **Q_PSG_NO_EASY**.

Table 10 shows the distribution of Q_PSG_EASY and Q_PSG_NO_EASY questions across Q_EASY, Q_MOD, and Q_DIF.

Table 10: Difficulty of passage selection

	Q_PSG_ EASY	Q_PSG_ NO_EASY	Total
Q_EASY	38	27	65
Q_MOD	39	52	91
Q_DIF	6	33	39
Total	83	112	195
MRR (AVG)	0. 376	0. 249	0. 303
MRR (MEI)	0. 463	0. 331	0. 387

Notice that the MRRs both for AVG and MEI correlate nicely with Q_PSG_EASY and Q_PSG_NO_EASY. The systems consistently produce a better MRR on Q_PSG_EASY questions than on Q_PSG_NO_EASY questions. The result suggests that the distinction between Q_PSG_EASY and Q_PSG_NO_EASY questions may be usefully exploited to predict performance on questions in QA tasks.

5.2.3 NE extraction

As a way of examining how performance in NE extraction affects that in QA tasks, we focus on what we call the "context of answer NEs." A context here is to be understood as a small textual stretch in which an NE appears.

We classified the context of answers into the following 6 groups:

(a): an answer involves an open class NE such as company name, and its context contains no cue words or non linguistic symbols to identify that NE,

(b): an answer involves a closed class NE such as a name of a prefecture. It comes with no cue words or no linguistic symbols in the context,

(c): an answer is marked by a pair of symbols used for punctuation on NEs in Japanese, *kagi kakko*,

(d): an answer is marked by any symbols other than (c)(ex. "(" and ")"),

(e): the context contains at least one keyword that functions as a unit of something (ex. "*en* (yen)," "3 *jikan* (3 hours)," "*Saitama ken* (Saitama prefecture") ,

(f): at least one cue word other than (e) appears in the context (ex. "*daitouryou* (President)").

Note that what counts as a cue word in (a), (b), (c), (d) is determined more or less arbitrarily. Table 11 shows the result of above classifications.

Table 11: Difficulty of NE Extraction

	(a)	(b)	(c)	(d)	(e)	(f)	Total
Q_EASY	4	6	13	4	16	22	65
Q_MOD	14	3	20	3	26	25	91
Q_DIF	8	3	4	7	6	11	39
Total	26	12	37	14	48	58	195
MRR (AVG)	0. 222	0. 413	0. 312	0. 262	0. 307	0. 318	0. 303
MRR (MEI)	0. 269	0. 569	0. 556	0. 214	0. 418	0. 311	0. 387

As (a) and (b) have no available indicator for NEs, we need some dictionary to identify NEs of either type. A poor MRR for (a) may suggest that our dictionary is not large enough to deal with NEs of the type (a) and (b).

While (c) and (d) both have NEs marked by some symbols, NEs of the type (c) turned out to be easier to identify than those of the type (d). The symbols in (c) typically indicate NEs in Japanese newspapers. It is worth noting that symbols other than those above such as "(" and ")" could act as indicators of an NE. Four questions in (d) under Q_DIF, asked for an alias of a particular NE. Obviously, to answer them involves more than identifying NEs.

Notice also in Table 11 that while on average system performs better on (f) than on (e), the opposite is the case with the MEI system.

5.2.4 Answer Category Identification

We divided each question in the test collection into the following four groups to examine the difficulty involved in selecting an appropriate answer category:

A: the answer category can be identified by looking at interrogative words in the question,
ex. "*dare* (who)," "*nan nen* (how many years)"

B: the answer category can be identified by looking at interrogative words and other words in the question. The question contains words indicating how specific an answer should be,
ex. "*zasshi no namae ... nani* (the name of the magazine ... what)," "*kimeta ... doko* (decided... where)"

C: Same as the above, except that the question lack information on the specifics of answer expressions,
ex. "*2006 nen no touki gorin no kaisai yoteiti wa doko desu ka* (Where will the 2006 Olympic Winter Games be held?)."

D: the answer category cannot be determined.
ex. "*Kafun shou no genin wa nan desu ka* (What are the causes of pollen allergy?).").

Table 12 shows the result of above classifications.

The questions in type A and B are distributed evenly across Q_EASY, Q_MOD and Q_DIF. The MRR for AVG decreases as one goes from A to D, but remains stable on A through C for the MEI QA system.

The answer categories of questions in D, listed below, were more difficult than others.

QAC1-1040-01: *Supo-tsu yougo de shiruba- buru-mu wa nani wo imi shimasu ka.* (In the world of sports, what does the term Silver Bloom mean?)

QAC1-1102-01: *Kafun shou no genin wa nan desu ka.* (What are the causes of pollen allergy?)

QAC1-1162-01: *Makao wa Porutogaru go de dono youni arawashi masu ka* (How is Macao spelled in Portuguese?)

Table 12: Difficulty of answer category decision

	A	B	C	D	Total
Q_EASY	38	25	2	0	65
Q_MOD	42	36	12	1	91
Q_DIF	18	15	4	2	39
Total	98	76	18	3	195
MRR(AVG)	0. 328	0. 296	0. 231	0. 104	0. 303
MRR (MEI)	0. 396	0. 389	0. 389	0. 067	0. 387

6. Proposals for Future QA tasks

In this section, we will identify some of the issues the future QAC tasks need to address, based on the discussion in the previous sections.

6.1 Categories of Answers

In the test collection, 75% of the answers of the questions are classified as ARTIFACT, PERSON, LOCATION, or NUMBER. The previous discussion suggests that PERSON and LOCATION represent easy cases.

More of the difficult questions such as one that requires one to identify ORGANIZATION, should be represented in the future test collection.

More of the questions that requires more difficult answer categories should be represented

6.2 Technical Aspects

One of the problems with the present QAC setup is that it is not clear what technical challenge each question poses. It fails to address questions like "Are some questions more important to answer than others?" or "Are they equally important?" In the QAC1, technical challenges involved in the test are not fully explained. The analysis of technical aspects of a test collection by participants takes time and is rather difficult itself.

Technical challenges or issues we found through the analysis of the test collection include:

- how to retrieve relevant documents when they come in small number.
- how to retrieve passages when keywords form a question are unable to distinguish between relevant and non-relevant passages.
- how to identify an NE when the context of answer NEs lack any cue, and those marked by some unknown symbols.
- the problem of answering a question which contains no information on categories it might belong to or on the level of specificity required of its answer.

In addition, two questions of the test collection required pulling out answers from tables, which no system was able to answer correctly. The following issue might be added to the above list:

- how to identify the part of a document in which the answer exists.

We should explore novel technical challenges, which reflect realistic QA situations. For example, the poverty of keywords in the question might be one of the problems that should be faced by QA systems.

In contrast, most of the questions in the test collection are long and rich with indicators. Some questions include the expressions which seem to be unnecessary to specify the answer, as in QAC1-1104-01, "With which company did the automatic transmission manufacturer Jatco, whose shares owned by Matsuda were all purchased by Nissan Motor, merge?", instead of "With which company did Jatco merge".

6.3 Definition of Tasks

We like to see future QAC tasks include a primary task that holds to a rigorous evaluation, and some tasks of more experimental nature.

A primary task should be simple and similar to the task1 of QAC1 (QAC1-T1), which enjoyed participation by most of the systems.

A unique feature of the QAC1-T1 is that it requires systems to supply exact answers. The feature is not shared with TREC and should stay as part of the primary task.

Another feature of QAC1-T1 is that it allows systems to use external resources and deliver answer expressions that may or may not exist in test documents. We believe, however, that answers should be limited to those one can extract from test documents in the primary task, for it makes fair and easy the evaluation of a system's performance. Perhaps it would be wise to restrict the use of external resources to particular interests like query expansion.

The following list some of the possible extensions/modifications to the current QAC scheme:

- allowing variations in answer expressions with the same reference.
- upholding a most specific answer expression found in documents,
- searching exhaustively for answers in documents, and
- finding a specified number of answers in documents.

The future evaluation scheme should be modified to address the above issues. An additional modification would be to generate different scores according to the specificity required of an answer, which would be required in case one needs some specific answer to a question.

7. Conclusions

We analyzed the results of NTCIR QAC1 task1.

As for passage retrieval, we should investigate a new method for defining passages appropriately, which seriously affects MRR. Also, a new measure should be introduced to evaluate the relevance of passages, incorporating their properties on, for example, the appearance of keywords.

As for NE extraction, the results of experiments showed that a modification of rules for NE extraction contribute to an improved MRR.

In the latter half of this report, we analyzed the test collection and made proposals for future QAC tasks.

The Questions in the test collection are classified in terms of answer categories. The analysis identified some features relevant for measuring the difficulty of the questions.

We made some proposals for the future QAC tasks in respect of categories of answers, technical aspects, and definition of tasks, based on the analysis of the test collection. We hope the future QAC test collection serves as a vehicle for the evaluation of the system's ability to solve the practical problems in QA.

References

[1] E. M. Voorhees, "Overview of the TREC 2001 Question Answering Track", in Proceedings of the Tenth Text REtrieval Conference (TREC 2001), 2002.

[2] J. Fukumoto, T. Kato, and F. Masui, "Question Answering Challenge (QAC-1) Question answering evaluation at NTCIR Workshop3", Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, to be published in Tokyo, 2003.

[3] M. Sato, H. Ito, and N. Noguchi, "NTCIR Experiments at Matsushita: Ad-hoc and CLIR task", Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.71-81, Tokyo, Aug. 1999.

[4] M. Sato, and N. Noguchi, "NTCIR-2 Experiments at Matsushita: Monolingual and Cross-Lingual IR tasks", Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, pp.5-173-178, Tokyo, Mar. 2001.

[5] H. Ito, and Y. Fukushige, "IREX NE Experiments at Matsushita" (in Japanese), Proceedings of the IREX Workshop, pp.163-169, Tokyo, Sept. 1999.

APPENDIX

List of questions that less than 3 systems could answer

Question ID	Question	Target category	Correct documents	Ratio of keywords	Passage selection	NE extraction	Answer category decision
Questions that no system correctly answered							
QAC1-1007-01	日本カー・オブ・ザ・イヤーを受賞したことのあるダイハツ工業の車は何ですか。	ARTIFACT:PRODUCT_NAME	1	1.000	EASY	(a)	B
QAC1-1014-01	世界で最も高いビルは何というビルですか。	LOCATION:SPOT	4	0.750	EASY	(f)	B
QAC1-1048-01	JPEGは何をもとにした略称でしょうか。	ARTIFACT:OTHER	1	0.333	NO EASY	(d)	C
QAC1-1068-01	よみうりランドにある木でできたジェットコースターの名前は何ですか。	ARTIFACT:OTHER	2	0.500	NO EASY	(c)	B
QAC1-1081-01	小渕恵三の前に総理大臣だった人は誰ですか。	PERSON:JAPANESE	3	0.933	NO EASY	(f)	A
QAC1-1090-01	「怪談」の作者が日本に帰化する前の名前は何ですか。	PERSON:FOREIGN	2	0.250	NO EASY	(d)	B
QAC1-1137-01	肥満の判定基準となっているものは何ですか。	ARTIFACT:OTHER	2	1.000	NO EASY	(c)	B
QAC1-1149-01	柔道の井上康生の父親は何という名前ですか。	PERSON:JAPANESE	2	0.667	NO EASY	(f)	B
QAC1-1173-01	五千円札に描かれている湖の名前は何ですか。	LOCATION:NATURE	1	0.333	NO EASY	(f)	B
QAC1-1175-01	1998年の豊かさ指標で総合2位となったのはどこの県ですか。	LOCATION:PREFECTURE	2	0.833	NO EASY	(b)	B
QAC1-1176-01	1997年の国会議員の所得で13位だったのは誰ですか。	PERSON:JAPANESE	1	0.833	NO EASY	(a)	A
Questions that 1 system correctly answered							
QAC1-1003-01	NHK連続テレビ小説の平均視聴率は最高どのくらいですか。	NUMBER:PERCENT	2	0.857	NO EASY	(e)	B
QAC1-1015-01	これまでで日本の最高気温を記録したのはどこですか。	LOCATION:PREFECTURE	1	1.000	NO EASY	(b)	B
QAC1-1040-01	スポーツ用語で「シルバーブルーム」は何を意味しますか。	ARTIFACT:OTHER	1	0.833	NO EASY	(c)	D
QAC1-1062-01	宮本武蔵が生まれたのは現在の何県何町ですか。	LOCATION:TOWN	1	0.625	NO EASY	(e)	A
QAC1-1075-01	メガネの日を決めたのはどこですか。	ORGANIZATION:OTHER	1	0.500	NO EASY	(f)	B
QAC1-1094-01	日本の桜の中で最も有名な品種は何ですか。	LIVING_THINGS:PLANT	4	0.550	NO EASY	(d)	B
QAC1-1162-01	マカオはポルトガル語でどのように表しますか。	LOCATION:COUNTRY	1	1.000	NO EASY	(a)	D
QAC1-1163-01	世界最大の花の名前は何ですか。	LIVING_THINGS:PLANT	1	0.667	NO EASY	(a)	B
QAC1-1198-01	プレイステーション用ソフト「トゥームレイダー3」の主人公は誰ですか。	PERSON:FOREIGN	1	0.800	NO EASY	(f)	A
Questions that 2 systems correctly answered							
QAC1-1009-01	寄付金とお年玉くじがついた年賀はがきが発売されるようになったのはいつですか。	TIME:DATE	1	0.889	NO EASY	(e)	A
QAC1-1023-01	米ソの冷戦が終わったのはいつですか。	TIME:DATE	3	1.000	NO EASY	(e)	A
QAC1-1047-01	小沢征爾はいつからボストン交響楽団の音楽監督を務めていましたか。	TIME:DATE	2	0.944	NO EASY	(e)	A
QAC1-1066-01	相撲の小錦が所属していた部屋はどこですか。	ORGANIZATION:SPORTS	3	0.833	NO EASY	(f)	C
QAC1-1067-01	絶対零度は摂氏何度ですか。	NUMBER:QUANT	3	0.267	NO EASY	(d)	A
QAC1-1071-01	ポパイの結婚相手は誰ですか。	PERSON:FOREIGN	3	0.583	EASY	(f)	A
QAC1-1072-01	ピアノ三重奏で使われる楽器はなんですか。	ARTIFACT:PRODUCT_CLASS	10	0.550	EASY	(a)	A
QAC1-1077-01	蝶の形をしたパスタの名前はなんですか。	ARTIFACT:PRODUCT_CLASS	1	0.500	NO EASY	(a)	A
QAC1-1078-01	キーマンはどこのお茶ですか。	LOCATION:COUNTRY	3	0.500	NO EASY	(d)	A
QAC1-1082-01	ガニメデは何星の側にありますか。	ASTRO	1	0.250	NO EASY	(f)	A
QAC1-1084-01	奈良の世界遺産にはどのようなものがありますか。	LOCATION:SPOT	7	0.929	EASY	(d)	C
QAC1-1091-01	グリーンランドは何領ですか。	LOCATION:COUNTRY	1	0.667	NO EASY	(b)	B
QAC1-1096-01	日本神話で「天孫」とは誰のことを指しますか。	PERSON:JAPANESE	3	0.467	NO EASY	(d)	A
QAC1-1104-01	日産自動車がマツダ保有株を全て買い取った自動変速機メーカー「ジャトコ」は、その後どこと合併しましたか。	ORGANIZATION:COMPANY	1	0.727	NO EASY	(c)	B
QAC1-1132-01	茨木県東海村での臨界事故を受け、政府が国会に提出した原子力防災のための2つの法案は何ですか。	ARTIFACT:OTHER	5	0.646	NO EASY	(f)	A
QAC1-1179-01	毎日新聞の購読申し込みのフリーダイヤルは何番ですか。	NUMBER:NUMBER	6	0.667	NO EASY	(a)	C
QAC1-1181-01	クーヘと呼ばれる皿状のそりにあおむけに寝て滑り降りる第9回インスブルック大会から採用された氷上競技をなんといいますか。	ARTIFACT:OTHER	1	1.000	NO EASY	(a)	A
QAC1-1186-01	第48回別府大分マラソンの優勝タイムは何時間何分何秒ですか。	TIME:PERIOD	2	0.615	NO EASY	(e)	A
QAC1-1197-01	ドラマ「GTO」(フジ系)で教頭役を演じた俳優は誰ですか。	PERSON:JAPANESE	1	0.778	EASY	(f)	A