NTCIR-4 QAC Experiments at Matsushita

NOMOTO, Masako FUKUSHIGE, Yoshio SATO, Mitsuhiro SUZUKI, Hiroyuki
{nomoto.masako, fukushige.yoshio, sato.mitsuhiro, suzuki.suzuki}@jp.panasonic.com

Network Systems Development Center, 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-4 QAC2, the second attempt to evaluate the technology of Japanese question answering (QA). Our basic approach is a combination of information 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 show the effects of refinements of answer extraction.

We also analyze the QAC2 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 the measurement of difficulty of the test collection and definition of tasks.

Keywords*: question answering (QA), named entity extraction, pattern matching, information 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.

We participated in subtask 1 of QAC1 and 2 succeedingly. 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 information retrieval and named entity (NE) extraction using pattern matching. No pre-processing is performed except for indexing.

Section 2 gives the overview of our system. Section 3 analyzes our results in QAC2 subtask 1. Section 4 reports the results of additional experiments to improve the performance of our system.

In section 5, we analyze the QAC2 test collection to identify features relevant for measuring the difficulty of the questions in the collection. Section 6 makes proposals for the next QAC tasks based on the analysis of section 5.

2. System Descriptions

2.1 The Architecture

Figure 1 shows the basic architecture of our system.

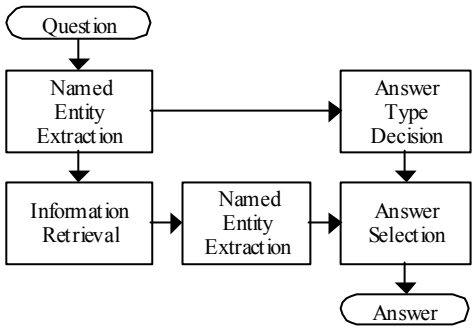

Figure 1.Architecture of the MEI QA system

The processing steps of our system are the followings:

(1) The NE extraction module annotates an input question with named entity categories.
(2) The information retrieval module extracts keywords from the annotated question and retrieves top *n* ranking documents.
(3) The NE extraction module annotates the retrieved documents 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 documents that match the answer type and selects an answer.

Information retrieval and index pre-processing are performed using the *MEISTER* software libraries, which have been used in our IR systems in NTCIR-1, 2 and 3 [3] [4] [5]. The NE extraction module is

based on the NE tool in IREX NE task [6] using hand created matching rules.

2.2 Methods

2.2.1 Information retrieval Module

The information retrieval module features the following methods:

-The unit of retrieval is a document.

-*Coordination Level Scoring* (*CLS*)[3] to rank retrieved documents, among which top 20 documents are used.

We switched from passage retrieval (QAC1) to document retrieval, based on our experimental results in QAC1 that showed the performance of the information retrieval got the best when the unit of retrieval, i.e. a passage was defined as a document [5].

2.2.2 NE Extraction Module

The NE extraction module annotates questions and retrieved documents with NE category tags using pattern matching rules (271) and dictionaries.

We defined 40 tags, of which 17 basic tags are shown below.

ARTIFACT, DATE, LOCATION, MONEY, ORGANIZATION, PERCENT, PERSON, TIME, EVENT, FREQ, LANG, NUM, ORDER, PERIOD, PRIZE, PRODUCT_CLASS, QUANT

Basic tags may also have subclassess. The first 8 of the above tags follow the IREX NE task [6]. ARTIFACT is used as a default category in our system, and includes miscellaneous NEs and non-NEs that are not classified in other categories.

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 62 pattern matching rules. When no rule matches, the module uses ARTIFACT as a default category. Examples of matching rules are shown below.

[Answer category]		[Rule]
ORGANIZATION	←	*(kaisha|daigaku|...).*doko*
		(company|university).*where
DATE	←	*nan* (*nen*|*gatsu*|*niti*)
		what(year|month|day)

For example, the question,

"baiagura wo kaihatsu shita no wa doko," Viagra delevoped where

meaning as a whole, "Which company developed Viagra?," matches the first of the above rules and the answer category is correctly referred to as ORGANIZATION though the category is not explicitly expressed in the question.

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 answer category tags in the retrieved documents. The score of each candidate NE $s(NE)$ is calculated by the following formula:

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

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

$kwj(w)$ = 2 if w is immediately followed by word j that belongs to a subset of JOSHI, a Japanese part of speech

=1 otherwise

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

$rank(doc)$ = the rank of the retrieved document that includes the NE.

Values of constants are:

$R_{max} = 20$, and $D_{max} = 100$.

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.

Main changes from QAC1 are:

-introduction of $kwj(w)$

-the settings of R_{max} and D_{max} (in QAC1, $R_{max} = 30$, and $D_{max} = 50$).

-extracting muptiple answer candidates from a document.

3. Formal Run Results and Analysis

Table 1 shows the result of subtask 1 in QAC1 and QAC2. Note that in QAC2, the jugment on the correctness of answers is more strictly made using document IDs in which the answers appeared[7].

Table 1. Subtask 1 results in QAC1 and 2

Run	MRR	RQ1	RQ5
QAC2	0. 418	0 .344	0 .538
QAC1	0. 387	0. 313	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

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

RQ5: The rate of 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 subtask[7]. The MRR in Table 1 suggests that for the averaged question in QAC2 subtask 1 we can include the correct answer in top 3 ranking but not in top2. The score is a little better than that of QAC1; it may be caused by the difference of difficulty of test collections, or improvement of our system, or both of them. On the other hand, RQ5 says we cannot include correct answer in top 5 ranking in about 46.2 % of the questions of the subtask.

Table 2 gives the number of questions for which each module made errors in QAC1 and 2. The QAC2 test collection consists of 197 questions[7], but 2 of them lack correct answers. Our system got correct answers for 135 questions and failed to answer 60, out of 195 questions that have at least one answers in the test collection.

Table 2. The errors made by each module

Module	# of questions failed(ratio)	
	QAC2	QAC1
Information Retrieval[1]	12 (0.135)	21 (0.216)
(Incorrect documents)	-	6 (0.062)
(Incorrect passages)	-	15 (0.155)
NE Extraction	38 (0.427)	48 (0.495)
Answer Type Decision	15 (0.167)	9 (0.093)
Answer Selection	24 (0.267)	19 (0.196)
(bug of answer set)	1	-
Total	90	97

NE extraction is still most problematic for us, and about 42.7 % of the errors in QAC2 occurred at this module. The information retrieval module, switched from passage retrieval (QAC1) to document retrieval (QAC2), got relevant documets for about 96.5 % of questions in QAC2. But expanding the unit of retrieval may affect the performance of answer selection module, which seriously got worse.

The increase of failure ratio on answer type decision should also be noted. We modified 11 rules of QAC1 rule set and added 32 rules for QAC2. An additional run proved that the revision of rule set as a whole worked well as shown in Table 3:

Table 3. Effects of revision of answer type decision rules

rule set	Correct	Recall	Precision	MRR
QAC2	135	0.351	0.135	0.418
QAC1	124	0.322	0.125	0.383

The causes of failure on this module should be investigated more in detail.

Table 4 gives the failure ratio of our system and its NE extraction module, Answer Type Decision module, Answer Selection module, for each answer category of the questions classified by our system.

Note that the category ARTIFACT, used as a default category in our system, may include NEs that should have been classified otherwise, and non-NEs.

Major categories of QAC2 in the above classification are[2]:

ARTIFACT, PERSON, LOCATION, ORGANIZATION, QUANT(or QUANTITY)

The fact that a considerable number of questions are classified as ARTIFACT implies the lack of answer categories. Finer grained classification scheme is needed for a precise error analysis. We should also notice that 71.4% of errors on ARTIFACT occurred at NE extraction module.

It is easy to identify answer candidates of questions asking for PERSON or LOCATION, but selecting correct answers from these candidates is difficult.

On the contrary, for questions asking for QUANT, answer candidates are not easily distinguished, but if rightly tagged, correct answers of this type can be selected easily.

The distribution of errors on ORGANIZATION among these modules is not so biased.

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.

As the error analysis showed that errors on answer selection increased in QAC2, we ran experiments to compare alternative settings.

We tested 3 additional weighting methods:

・*kwj'(w)*

kwj'(w), a modification of *kwj(w)* to limit the length of keywords to be weighted, is used instead of *kwj(w)* in Formula (1) in 2.2.4 and defined as follows:

kwj(w) = 2 if *w* is more than 2 bytes and immediately followed by a subset of words comprising JOSHI, a Japanese part of speech.

= 1 otherwise.

・*kww(w)*

kww(w) is introduced to weight the keywords that occur around wh-words in a question. When *kww(w)* is used, the score of answer candidate NE *s(NE)* is calculated by the following formula:

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

where,

[1] Errors by passage retrieval module(QAC1) are classified here in 2 levels for the purpose of evaluation.

[2] Questions on numeric expressions are classified into groups, and not dealt with as a category in our scheme.

kww(*w*) = 2: if the distance between *w* and wh-word *wh* in a question is within DW_{max} bytes
= 1: otherwise

Values of a constan is:

DW_{max} = 10.

・*kww'(w)*

kww'(w), a modification of *kww(w)* to limit the length of the keywords to more than 2 bytes, is used instead of *kww(w)* in Formula (1') and defined as follows:

kww'(w) = 2: if *w* is more than 2 bytes and the distance between *w* and wh-word *wh* in a question is within DW_{max} bytes

= :1 otherwise

Table 5 shows the results of the experiments. QAC2NOJWH2, using *kww(w)* instead of *kwj(w)*, got the best MRR, though the value is not so high as expected. What we observed here are the followings:

-*kww(w)* can be used as an alternative to *kwj(w)*, but not to be added as in Formula (1')(cf. QAC2ORG, QAC2WH2, and QAC2NOJWH2). Also, *kww(w)* seems to be a little better than *kww'(w)*(cf. QAC2NOJWH2 and QAC2NOJWH2UP2).

-*kwj(j)* is harmful(cf. QAC2ORG and QAC2NOJ, QAC2R10 and QAC2NOJR10, and also, QAC2DST50 and QAC1R20M). Using *kwj'(w)* instead of *kwj(w)* slighly raises the value of MRR but not the values of all the other measures(cf. QAC2JUP2 and QAC2ORG).

-DST_TH 100 is better than DST_TH 50 under R_LIMIT 10 and 20(cf. QAC2NOJ and QAC1R20M, and QAC2NOJR10 and QAC1R10M).

-Setting R_LIMIT to 10 or 20 does not make difference under DST_TH 100(cf. QAC2ORG and QAC2R10). Under DST_TH 50, R_LIMIT should be set to 20 (cf. QAC1R10M, QAC1R20M, and QAC2R30M).

-Extracting multiple answer candidates from a document (multi) works well to improve MRR and other measures(cf. QAC1R30ORG and QAC1R30M, and also, QAC2ORG and QAC2RANS1).

4.1 Discussions

The error analysis above revealed that NE extraction and answer selection module are problematic in our system.

Failures of our system, NE extraction module, answer type decision module, and answer selection module are classified for each answer category. The result proved that the distributions of failures among

Table 4. Failure ratio of our system and its modules for each answer category

Answer category	number of ques	failure(ratio)											
		MEI QA			NE extraction			Answer Type Decision			Answer Selection		
	(a)	(b)	((b)/(a))	(c)	((c)/(a))	((c)/(b))	(d)	((d)/(a))	((d)/(b))	(e)	((e)/(a))	((e)/(b))
ARTIFACT	50	28	0.560	20	0.400	0.714	0	0.000	0.000	5	0.100	0.179
DATE	7	1	0.143	0	0.000	0.000	0	0.000	0.000	1	0.143	1.000
EVENT	3	2	0.667	1	0.333	0.500	0	0.000	0.000	1	0.333	0.500
FREQ	0	-	-	-	-	-	0	-	-	0	-	-
LANG	0	-	-	-	-	-	0	-	-	0	-	-
LOCATION	39	12	0.308	3	0.077	0.250	1	0.026	0.083	5	0.128	0.417
MONEY	2	0	0.000	0	0.000	0.000	0	0.000	0.000	0	0.000	0.000
NUM	0	-	-	-	-	-	0	-	-	0	-	-
ORDER	4	2	0.500	2	0.500	1.000	0	0.000	0.000	0	0.000	0.000
ORGANIZATION	19	13	0.684	4	0.211	0.308	3	0.158	0.231	4	0.211	0.308
PERCENT	1	1	1.000	0	0.000	0.000	1	1.000	1.000	0	0.000	0.000
PERIOD	5	3	0.600	0	0.000	0.000	3	0.600	1.000	0	0.000	0.000
PERSON	45	15	0.333	4	0.089	0.267	1	0.022	0.067	6	0.133	0.400
PRIZE	1	0	0.000	0	0.000	0.000	0	0.000	0.000	0	0.000	0.000
PRODUCT_CLASS	4	4	1.000	1	0.250	0.250	1	0.250	0.250	1	0.250	0.250
QUANT	13	7	0.538	3	0.231	0.429	3	0.231	0.429	1	0.077	0.143
TIME	2	2	1.000	0	0.000	0.000	2	1.000	1.000	0	0.000	0.000
total	195	90	0.462	38	0.195	0.422	15	0.077	0.167	24	0.123	0.267

Table 5. Effects of using various settings of the answer selection

run	settings based on QAC2 or 1				other weighting of keywords			result			
	DST_TH	R_LIMIT	multi	kwj	kwj'	kww	kww'	Correct	Recall	Precision	MRR
QAC2ORG	100	20	Y	Y	-	-	-	135	0.351	0.135	0.418
QAC2WH2	100	20	Y	Y	-	Y	-	129	0.335	0.136	0.409
QAC2WH2UP2	100	20	Y	Y	-	-	Y	135	0.351	0.135	0.418
QAC2JUP2	100	20	Y	-	Y	-	-	134	0.348	0.134	0.423
QAC2NOJ	100	20	Y	-	-	-	-	134	0.348	0.134	0.429
QAC2NOJR10	100	10	Y	-	-	-	-	134	0.348	0.134	0.430
QAC2NOJWH2	100	20	Y	-	-	Y	-	135	0.351	0.135	0.432
QAC2NOJWH2UP2	100	20	Y	-	-		Y	134	0.348	0.134	0.429
QAC2RANS1	100	20	N	Y	-	-	-	124	0.322	0.126	0.415
QAC2R10	100	10	Y	Y	-	-	-	135	0.351	0.135	0.419
QAC2DST50	50	20	Y	Y	-	-	-	130	0.338	0.130	0.416
QAC1R30ORG	50	30	N	-	-			124	0.322	0.126	0.421
QAC1R30M	50	30	Y	-				132	0.343	0.132	0.426
QAC1R20M	50	20	Y	-				132	0.343	0.132	0.426
QAC1R10M	50	10	Y	-				127	0.330	0.134	0.412

the modules show different patterns for each of the major categories.

As for NE extraction, our system classifed answers of 50 questions into miscellaneous category called ARTIFACT. We failed 56% of the questions on ARTIFACT, and 71.4% of the failures is due to NE extradtion module. We should reconsider the range of the target NE categories.

We suppose the problem on classification of categories is common to other systems. The potential categories of answers for the QAC2 task is not clearly defined and there is no official classification of answer categories shared among participants.

As for answer selection, the results of experiments show changing the setting of answer selection can improve the performance of our system. Still, We should introduce a new feature to detect answers more correctly.

5. Analysis of the QAC2 test collection

In this section, we will look at answer categories for the QAC2 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 subtask1, we consider *RR(AVG)*, or the average of the RR(reciprocal rank)s of all the systems, which we introduced for the analysis of QAC1[5], given as the following:

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

where,

*AvgSys*5 : 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 in section 3, which suggests the need for a finer grained classification scheme of answer categories, we formulated a modified classification scheme for answers so as to cover the 195 questions having at least one answer in the subtask 1. This is based on the classification scheme we used in the analysis of QAC1[5]. We defined 42 categories, which are comprised of 9 basic categories and 33 sub categories.

Table 6 shows the number of questions in QAC1 and 2 and the performance of systems for each answer category. Newly introduced categories are marked with '*.' The category 'ARTIFACT_AND_OTHER:OTHER', used for the analysis of QAC1, is obsolete and seperated into OTHER_NE and OTHER_NON_NE.

Each question is classified into one of the categories. If a question has multiple answer strings that belong to different categories, we chose one; In case of numeric expressions, we gave priority to answers that belong to subcategories other than

Table 6. # of questions and performance of systems for each answer category

Answer Categories	Number of questions		MRR(AVG)	
	QAC2	QAC1	QAC2	QAC1
ARTIFACT_AND_OTHER	46	44	0.27	0.28
:LAW*	2	-	0.20	-
:MEDICAL*	3	-	0.23	-
:PRIZE*	1	-	0.27	-
:PRODUCT_CLASS	4	6	0.09	0.29
:PRODUCT_NAME	0	6	-	0.33
:WORK	8	10	0.42	0.32
:(OTHER)	-	22	-	0.25
:OTHER_NE*	20	-	0.32	-
:OTHER_NON_NE*	8	-	0.10	-
PERSON	45	42	0.49	0.36
:FOREIGN	23	11	0.49	0.37
:JAPANESE	22	31	0.49	0.35
LIVING_THINGS	7	8	0.23	0.20
:ANIMAL	5	1	0.29	0.27
:PLANTS	2	5	0.07	0.14
:OTHER	0	2	-	0.34
ASTRO	2	2	0.35	0.16
EVENT*	3	-	0.21	-
ORGANIZATION	19	20	0.17	0.27
:COMPANY	6	13	0.31	0.27
:POLITICS	3	3	0.47	0.40
:SPORTS	1	2	0.03	0.21
:OTHER	9	2	0.19	0.14
LOCATION	39	32	0.40	0.33
:COUNTRY	8	13	0.51	0.41
:STATE	3	1	0.22	0.43
:PREFECTURE	10	3	0.46	0.08
:CITY	6	3	0.34	0.32
:CAPITAL	1	3	0.64	0.51
:TOWN	2	2	0.22	0.28
:SPOT	6	5	0.33	0.27
:NATURE	3	2	0.38	0.08
NUMBER	20	29	0.38	0.28
:NUMBER	0	3	-	0.20
:QUANT	13	21	0.33	0.31
:MONEY	2	3	0.78	0.30
:ORDER*	4	-	0.32	-
:PERCENT	1	2	0.42	0.13
TIME	14	18	0.38	0.31
:DATE	7	14	0.50	0.35
:PERIOD	5	4	0.28	0.18
:TIME*	2	-	0.23	-
total	195	195	0.35	0.30

NUMBER:NUMBER. In other cases, answers that convey less detailed meaning are selected.

For about 86.7% of the questions, an answer is one of the following categories:

ARTIFACT_AND_OTHER, PERSON, LOCATION, NUMBER, ORGANIZATION

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

(basic categories)
ORGANIZATION, EVENT, LIVING_THINGS
(sub categories)
ORGANIZATION:SPORTS, LIVING_THINGS:PLANTS, ARTIFACT_AND_OTHER:PRODUCT_CLASS, ARTIFACT_AND_OTHER:OTHER_NON_NE, ORGANIZATION:OTHER

Easy categories for AVG were:

(basic categories)
PERSON, LOCATION
(sub categories)
NUMBER:MONEY,LOCATION:CAPITAL, LOCATION:COUNTY, TIME:DATE

The MRR(AVG) of QAC2 as a whole is better than that of QAC1. In most categories, also, the MRR(AVG) of QAC2 improves on that of QAC1. Exceptions are the followings:

(basic categories)
ORGANIZATION, ARTIFACT_AND_OTHER
(sub categories)
ARTIFACT_AND_OTHER:PRODUCT_CLASS, ORGANIZATION:SPORTS, LIVING_THINGS:PLANTS,LOCATION:TOWN

5.2 Causes for failure

Below we go through an analysis of what caused failures or poor performance on some of the questions in the test collection.

Let us look at the following features of questions, which we suppose may influence performance of systems:

- **RelD:** the number of relevant documents for a question
- **ARelD:** the total number of answer expressions for a question that appear in a relevant document
- **AvgARelD**: the average number of answer expressions for a question that appear in a relevant document
- **QLen:** the length of a question

Table 7 shows the correlation coefficient between the values of each of the above features of QAC test collection and Sys5, MRR(AVG), AvgSys5:

Table 7. correlation coefficient between features of test collections and Sys5, MRR(AVG), AvgSys5

	QAC2				QAC1
	RelD	ARelD	Avg ARelD	QLEN	RelD
Sys5	0.440	0.501	0.213	-0.122	0.535
MRR (AVG)	0.429	0.504	0.196	-0.174	0.511
AvgSys5	0.231	0.266	0.062	-0.219	0.251

As for QAC2 test collection, there is positive correlation between RelD and Sys5, RelD and MRR(AVG), ARelD and Sys5, ARelD and MRR(AVG). The same is true for RelD and Sys5, and RelD and MRR(AVG) of QAC1.

This shows that the number of relevant documents (RelD) has an impact on performance of QA systems, and thus serves as a potential indicator of how hard a question is. ARelD might also be used as an indicator, for the distribution pattern of correlation coefficent on ARelD across Sys5, MRR(AVG), and AvgSys5 is similar to that of RelD.

As for AvgSys5, there is weak postive correlation between RelD and AvgSys5, and ARelD and AvgSys5, not so obvious as Sys5 or MRR(AVG).

Figure 2, 3, 4, 5 shows the scatter diagram between each of the above features of test collection and Sys5 in QAC2.

The scatter diagrams between RelD and Sys5, and between ARelD and Sys5 show similar tendencies: The higher the value of RelD or ARelD is, the higher the bottom of Sys5 is.

On the other hand, Figure4 and 5 say the values of AvgARelD and QLen do not affect Sys5.

Figure 6 shows the scatter diagram between RelD and Sys5 in QAC1.

The diagrams between RelD and Sys5 in QAC1 and 2 (Figure 2 and 6) ,and between ARelD and Sys5 in QAC2 (Figure 3) all display similar patterns to reinforce the validity of RelD and ARelD as an indicator of hardness of a question.

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 discussions in the previous sections.

6.1 Measure for evaluating the difficulty of test collections

We need measures for evaluating the difficulty of the test collections. Without that, we can not correctly compare the results of evaluation using different test collections. The existing values such as MRR, Sys1, Sys5, AvgSys5, MRR(AVG), are all affected by the performance of systems, which are not stable, and are not appropriate for the evaluation of test collection itself.

6.2 Task definition

The range of the target categories of potential answer strings are not specified in the task definition of QAC1 and 2. This might be caused by the fact that partifipants do not share common answer categories, but we should know, for example, what kinds of non-NEs can be answers for questions in QAC.

One possibility is to separate questions into a few groups, for example, NEs, and non-NEs. It allows participants to evaluate their system's performance more precisely.

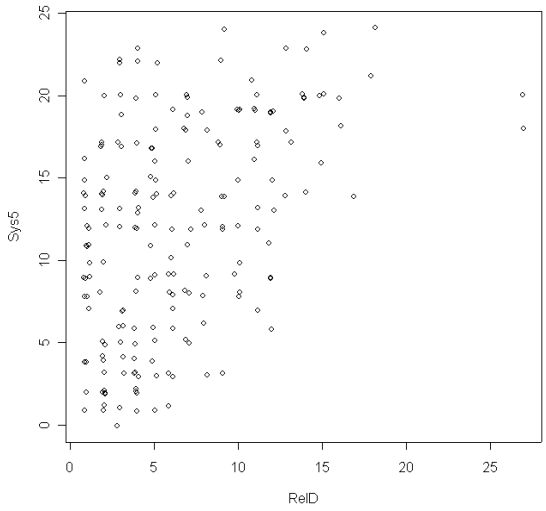

Figure2. Scatter diagram between RelD and Sys5

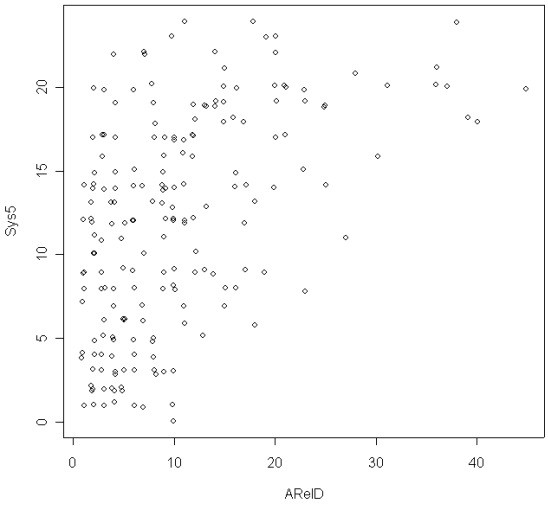

Figure3. Scatter diagram between ARelD and Sys5

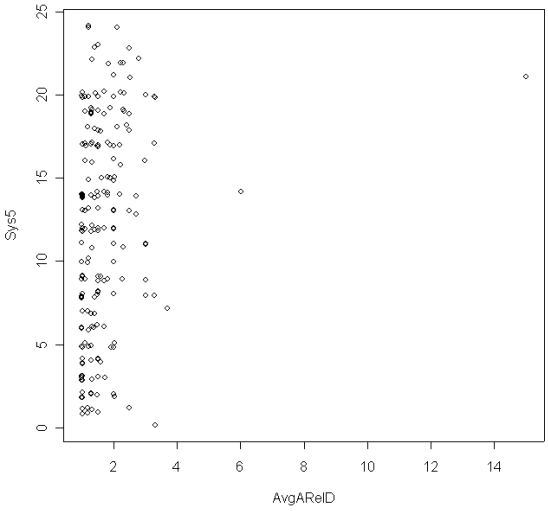

Figure4. Scatter diagram between AvgARelD and Sys5

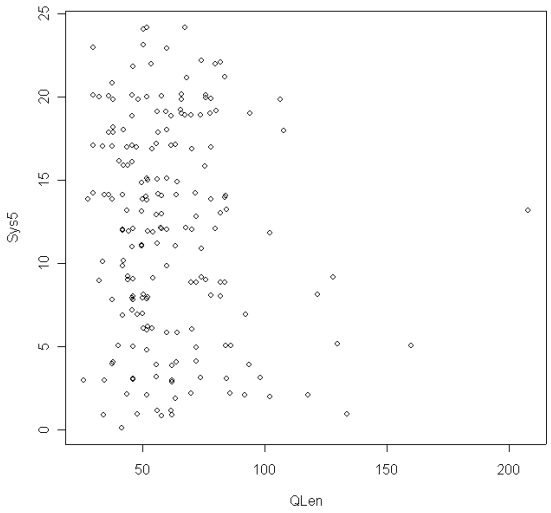

Figure5. Scatter diagram between QLen and Sys5

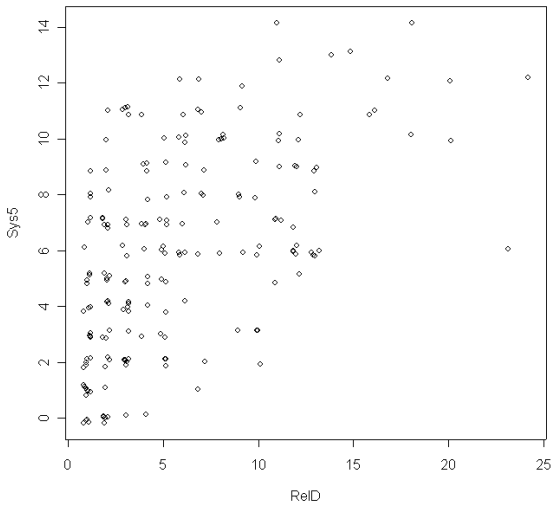

Figure6. Scatter diagram between RelD and Sys5 in QAC1

7. Conclusions

We analyzed the result of NTCIR4 QAC2 subtask 1.

Failures of our system, NE extraction module, answer type decision module, and answer selection module are classified for each answer category. The result proved that the distribution of failures among the modules shows different patterns for each of the major categories.

As for NE extraction module, we should prepare finer grained classification scheme to reduce failures due to the lack of the appropriate category. Also, we should reconsider the range of the target NE categories.

The results of experiments on answer selection showed that the change of settings slightly improves the result, but we need a new method.

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 measure for evaluating the difficulty of test collection, and task definition, based on the analysis of the test collection.

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 (QAC2) Question answering evaluation at NTCIR Workshop3", Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, 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] M. Nomoto, M. Sato, and H. Suzuki, "NTCIR-3 QAC Experiments at Matsushita", Proceedings of the Third NTCIR Workshop Meeting on Research in Information Retrieval, Automatic Text Summarization and Question Answering, 2003.

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

[7] J. Fukumoto, T. Kato, and F. Masui, "Question Answering Challenge for Five ranked answers and List answers - An Overview of NTCIR4 QAC2 Subtask 1 and 2 -", Working Notes of the Fourth NTCIR Workshop Meeting, to be published in Tokyo, 2004.