Are We Making Progress? -An Analysis of NTCIR QAC1 and 2-

Masako Nomoto Yoshio Fukushige Mitsuhiro Sato Hiroyuki Suzuki
Network Systems Development Center, Matsushita Electric Industrial Co., Ltd.
4-5-15, Higashi-Shinagawa, Shinagawa-ku, Tokyo 140-8632 JAPAN
{nomoto.masako, fukushige.yoshio, sato.mitsuhiro, suzuki.suzuki}@jp.panasonic.com

Abstract

This paper tackles issue on comparing evaluation results using multiple QA test collections(NTCIR QAC1 and 2). We identify two features that have moderate correlation with the performance of systems in QAC1 and 2 and evaluate the difficulty of the two test collections using the features. Answer categories of questions also affect the performance of systems. The evaluation results suggest that QAC2 seems to be easier than QAC1 in terms of the features, and we are making progress at least for some categories. We make a proposal for the future QAC tasks, as regards to the data needed for evaluation using multiple test collections.

Keywords: *NTCIR, question answering (QA), test collection, evaluation*

1 Introduction

Question answering (QA) represents an alternative approach to information retrieval(IR), using information extraction(IE) techniques. The development of QA technology has been supported by evaluation workshops in the field of IR, such as TREC and NTCIR.

The TREC question answering tracks, started in 1999 (TREC-8) [1], have focused on English QA. In earlier tracks(TREC 8 and 9), systems were required to return a ranked list of up to five 'text snippets' with document-ids. The score for a system's submission was calculated as a metric called MRR (Mean Reciprocal Rank), or the mean of the individual question's reciprocal ranks. On the other hand, the Main task (or factoids task in the Main task) in TREC 2002 and 2003 QA[2][3], systems were required to return one 'exact answer' per question.

Efforts to evaluate the technology of Japanese QA have been made in a series of the NTCIR question answering challenge(QAC) task, which started as NTCIR-3 QAC1[4] in 2002.

The NTCIR QAC owes much to the TREC. The basic design of Main Task(Subtask1) was similar to that of earlier TRECs; systems returned up to 5 ranked answers for each question, and the result was evaluated using MRR as defined in TREC. The remarkable point in NTCIR QAC is that 'exact answer' was required, which was later incorporated in TREC 2002 QA.

In the second challenge(QAC2), eighteen teams participated, including six teams newly joined. The problem is that teams participated in QAC1 and 2 successively cannot tell their systems improved or not. There is no evaluation metric to compare the difficulty of multiple QA test collections.

The difficulty of a question for a system may be affected by some factors related to the particular system or the test collection. However, if the question is easy or difficult for many systems, it may be caused by some features of the test collection. Such features may serve to compare the relative difficulty of multiple test collections.

Our goal of this study is to answer the question: Are we making any progress? With no existing metric to measure the difficulty of multiple QA test collections, we now need to examine what kinds of questions are easy or difficult for systems, in other words, what kinds of factors of test collections affect the performance of systems, and evaluate the difficulty of two test collections using the factors.

The remainder of this paper is organized as follows. Section 2 gives an overview of NTCIR QAC1 and 2. Section 3 describes an analysis of the QAC1 and 2 test collections. Section 4 makes a proposal for the next QAC tasks, and Section 5 concludes this paper.

	collection		the Main task			
	documents		# of ques	# of teams	evaluation	response for a question
	newspaper	span	(used)	(systems)	metrics	
QAC1	Mainichi	'98 and '99	200(195)	13(15)	MRR	up to five ranked exact answers
QAC2	Mainichi, Yomiuri	'98 and '99	200(195)	18(25)	MRR	up to five ranked pairs of an exact answer and a document ID

Table 1. Overview of NTCIR QAC1 and 2

2 Overview of NTCIR QAC1 and 2

2.1 The Main task

A summary of the NTCIR QAC1 and 2 is given in Table 1[5][4][6].

The basic design of Main task is common to QAC1 and 2. MRR was used as the formal evaluation metric. Questions were manually developed, of which those having more than one correct answer in documents were used for evaluation.

The main differences between QAC1 and QAC2 are listed as follows:

- In the Main task of QAC1, answer strings were not necessarily extracted from documents in the test collection, though the IDs of documents from which answer strings could be extracted were given in the answer data set. In QAC2, pairs of an answer string and the ID of the corresponding document were used for evaluation.
- A new set of news articles was added to document data in QAC2.

Evaluation results, the answer set, and the statistical data on questions, as shown in Table 2 were delivered to participants.

2.2 Evaluation results

The following summarizes evaluation results of Main task in QAC1 and 2.

	QAC1	QAC2
average of MRR	0.310	0.363
SD	0.150	0.145

The averaged MRR of systems in QAC2 is a little higher than that of QAC1. The standard deviation(*SD hereafter*) in QAC1 and 2 are almost the same.

Figure 1 and 2 show the overall performance of systems participated in Main task of QAC1 and 2, respectively. In QAC2, the mode is higher and the data skews to the left.

However, we cannot conclude that systems improved from these data, for the relative difficulty of the two test collections is not known. Let us leave the question open for now, and move on to the analysis of the test collections.

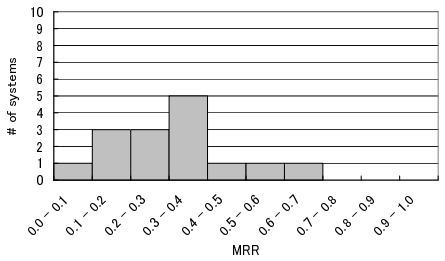


Figure 1. Performance of systems(QAC1)

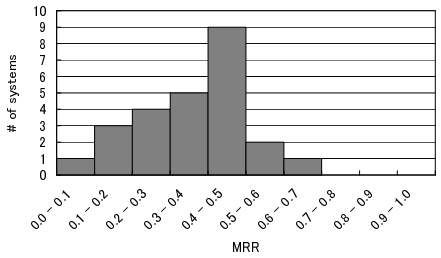


Figure 2. Performance of systems(QAC2)

3 Analysis of the QAC1 and 2 test collections

In this section, we will compare the question-wise performance of systems and investigate what factors of test collections caused the good or poor performance. Features of test collection, supposed to be related to modules of typical QA systems, and answer categories are tested if they affected the performance of systems in QAC1 and 2.

To see how difficult or easy each question of the test collection is for systems, 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[7], given as the following:

$$RR(AVG) = AvgSys5 * \#Sys5/\#SysAll \quad (1)$$

where, #SysAll is number of systems, 15 in QAC1, and 25 in QAC2, respectively, and AvgSys5 and #Sys5 are defined as in Table 2. In the following, MRR(AVG), the averaged RR(AVG)s for a set of

data	description
#Sys1	the number of systems that returned the correct answer in the first place
#Sys5	the number of systems that returned the correct answer in up to the fifth place
AvgSys5	the average of the RRs of the systems that obtained more than zero in RR

Table 2. Statistical data on the Main task questions in QAC1 and 2

questions, refers to the averaged performance of all the systems, which will be supplemented by the data of #Sys5, or the ratio of #Sys5.

3.1 Question-wise performance of systems

The following summarizes question-wise performance of systems in QAC1 and 2:

	QAC1	QAC2
average of RR(AVG)	0.303	0.363
SD	0.204	0.230

In QAC2, the MRR(AVG) slightly improved and the standard deviation is a little larger than QAC1, which means the difficulty level varies among questions in QAC2.

Figure 3 shows the question-wise comparison of performance of systems. The number of questions in QAC1 decreases steeply at RR(AVG) 0.6 or more, but in QAC2, considerable number of questions got more than 0.6. Is this caused by some features of the test collection or does it mean that the systems improved? To answer these questions, we go through an analysis of what caused good or poor performance of systems.

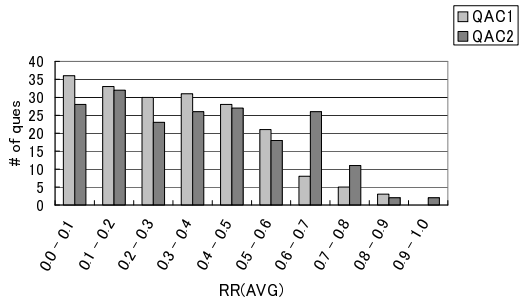


Figure 3. Question-wise comparison of performance of systems

3.2 Causes of difficulty

In this subsection, we will test some features of test collections and answer categories if they cause good or poor performance of systems, and try to evaluate the difficulty of questions.

Our basic ideas are listed as follows:

- Typical QA systems are comprised of basic modules, for example, information retrieval (IR), answer type (or question type) decision, information extraction (IE), and answer selection.
- The performance of systems for questions is affected by features of test collections, each of which is closely related to at least one module of typical QA systems.
- The effect of each feature varies among questions.
- Answer categories, which seem to be related to multiple modules such as answer type decision, answer selection, information retrieval, may also affect the performance of systems, based on our earlier report on QAC1[7].

3.2.1 Testing features of test collections

Based on the basic ideas, we selected four features supposed to be related to at least one module of typical QA system as shown in Table 3. #RD is expected to serve as a potential indicator of how hard a question is, based on our earlier report on QAC1[7]. QLength is supposed to affect the number of keywords extracted from questions and we believe is related to IR and answer selection.

Values of these features are calculated using the answer set, question set, and document data in the test collection. As for document data, we used data of Yomiuri Newspaper (1998 and 1999) provided as part of the test collection and CD-Mainichi Newspaper (1998 and 1999).

Scatter diagrams and the correlation coefficient(r) are compared between the features and the performance measures, namely, RR(AVG), MRR(AVG) and #Sys5.

3.2.2 Results and discussions

Figure 4, and 6 through 9 given in APPENDIX, show the scatter diagrams RR(AVG) versus each of the features. The horizontal axis represents the value of each feature, and the vertical axis represents the RR(AVG) and each dot represents a question.

As can been seen from figure 4(a) and (b), the data on #RD show similar patterns in QAC1 and 2. The larger the number of #RD is, the higher the bottom of RR(AVG) is, and the right bottom of the diagrams are left blank. If a question has

feature	description	supposed modules
#RD	the number of relevant documents for a question	IR
#AS|RD	the total number of answer expressions for a question that appear in relevant documents	answer selection
E(#AS|RD)	the average number of answer expressions for a question that appear in a relevant document	answer selection
QLength	the length of a question	IR, answer selection

Table 3. Tested features of test collections

more than 13 relevant documents, the RR(AVG) is highly likely to be more than 0.5 in QAC1 in most cases, and more than 0.45 in QAC2. In other words, when a question has many relevant documents, #RD seriously affects the performance of systems and make the questions much easier than others.

The data on #AS|RD in figure 4(c) and (d) show similar tendencies to that of #RD. It is very probable that the RR(AVG) is more than 0.5, if answer strings appear more than 35 times in relevant documents in QAC1, and more than 28 times in QAC2. We can say from these data, that #AS|RD, also strongly affect the performance of systems, if answers for a question appear many times in relevant documents. However, we should notice the data on #AS|RD, as well as #RD, imply that RR(AVG) is likely to be affected by some factor other than these.

On the other hand, as revealed by figure 6 through 9 in APPENDIX, the average number of answer expressions for a question that appear in a relevant document (E(#AS|RD)) and the length of a question (QLength) do not affect RR(AVG).

As for correlation coefficient, Table 4 gives correlation coefficient (r)between the features of test collections and performance of systems. In both QAC1 and QAC2, #RD and #AS|RD seem to have moderate correlation with both MRR(AVG) and #Sys5. This suggests that #RD and #AS|RD may affect the performance of QA systems, though they do not fully represent the difficulty of the questions. Notwithstanding these limitations, we will try to use these features as a potential indicator of how hard a question is.

The values of these features are higher in QAC2, and the questions in QAC2 seem to be easier than QAC1. In contrast, most of the correlation coefficient between these features and performance measures (#Sys5, MRR(AVG)) decreased in QAC2, suggesting that they are also affected by other factors.

3.2.3 Testing answer categories

In this subsection, we will test answer categories if they affect performance of systems and evaluate the difficulty of questions for each answer category.

We formulated a classification scheme for answers so as to cover Main task1 questions having at least one answer in QAC1 and 2, based on the scheme we used for the analysis of QAC1[7]. We defined 42 categories, comprised of 9 basic categories and 33 sub categories. They include the following 8 categories defined in IREX NE task[8]:Organization, Person, Location, Artifact, Date, Time, Money, Percent. Notice that the category Artifact_and_other:Other, used for the analysis of QAC1, is now obsolete and separated into Other_ne and Other_non_ne in our scheme.

Each question was manually classified into one of the categories the answer of which belongs to. 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 Number:Number. In other cases, answers that convey less detailed meaning are selected.

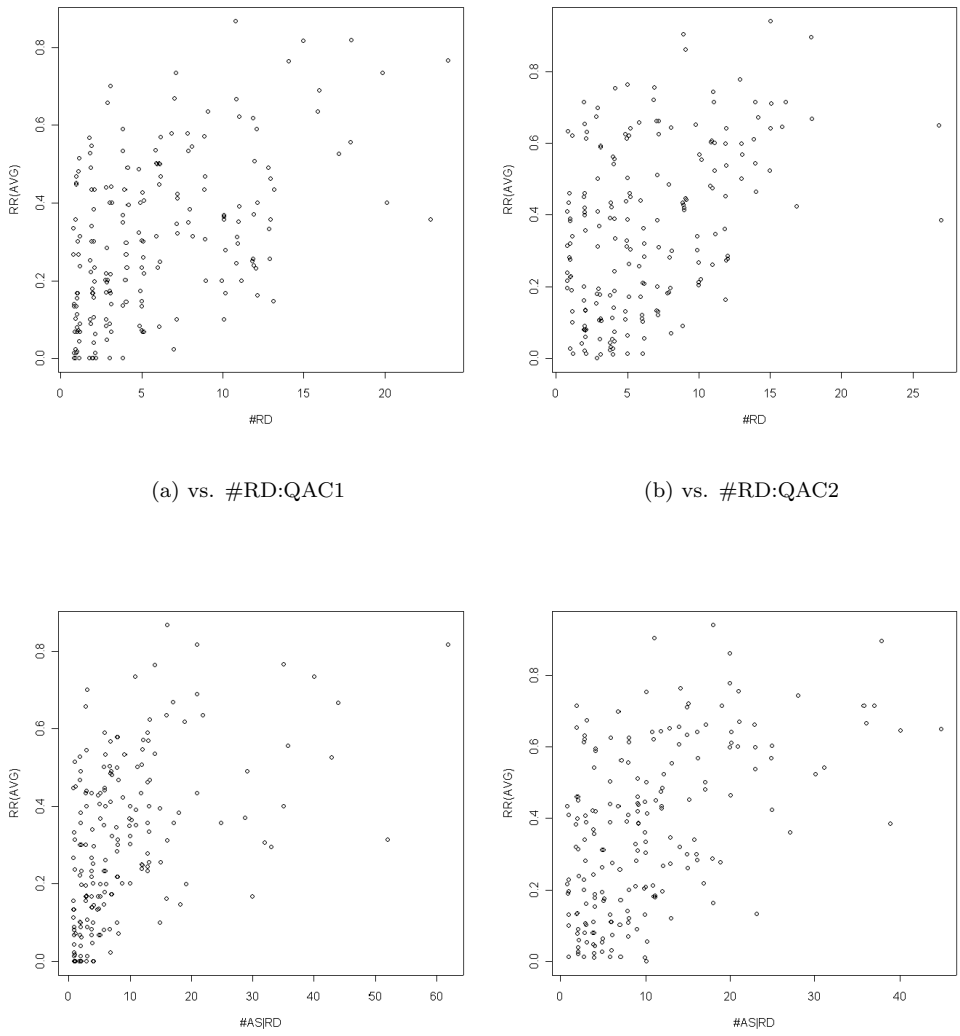
We compared the performance of systems and evaluated the difficulty of questions based on the features (#RD and #AS|RD) for each answer category.

3.2.4 Results and discussions

Table 5 gives the distribution of questions and the performance of systems for each answer category. Each subcategory name is preceded by a colon (:). The distribution patterns of questions across categories are alike both in QAC1 and QAC2. Most of the questions are classified into one of the IREX-based categories.

MRR(AVG) of IREX-based basic categories are more than 0.2, which means that averaged system can return at least one correct answer for an averaged question of the category up to the fifth rank. In QAC2, the performance of systems, both in MRR(AVG) and ratio of #Sys5, improved as a whole and for most of the answer categories. This may be caused by progress of systems or some factor of the test collections. Notice that both MRR(AVG) and #Sys5/#SysAll show similar tendencies across answer categories in QAC1 and 2. This implies that answer categories seem to affect performance of systems. We should also notice that MRR(AVG) has strong correlation with (#Sys5)/(#SysALL); Actually, the correlation coefficient between them across basic categories are 0.979 in QAC1 and 0.956 in QAC2, respectively.

Then, let us go back to the question, are we



(a) vs. #RD:QAC1

(b) vs. #RD:QAC2

(c) vs. #AS|RD:QAC1

(d) vs. #AS|RD:QAC2

Figure 4. Scatter diagram between #RD,#AS|RD and RR(AVG)

		QAC1				QAC2			
		#RD	#AS|RD	E(#AS|RD)	QLength	#RD	#AS|RD	E(#AS|RD)	QLength
average		5.7	9.0	1.5	51.2	6.4	10.3	1.7	61.0
SD		4.8	9.7	0.9	18.2	4.7	8.6	1.2	23.8
r	#Sys5	0.536	0.510	0.152	0.011	0.440	0.501	0.213	-0.122
	MRR(AVG)	0.510	0.473	0.136	0.003	0.429	0.504	0.196	-0.174

Table 4. Correlation coefficient between features of test collections and performance of systems

Answer category	QAC1						QAC2					
	# of ques	MRR (AVG)	#RD		#AS|RD		# of ques	MRR (AVG)	#RD		#AS|RD	
			value	r	value	r			value	r	value	r
ARTIFACT_AND_OTHER	44	0.280	6.1	0.430	9.0	0.291	46	0.266	6.2	0.670	9.5	0.603
PERSON	42	0.358	6.4	0.528	11.2	0.593	45	0.492	6.8	0.526	13.2	0.601
LOCATION	33	0.331	5.5	0.573	7.8	0.513	39	0.400	6.2	0.348	10.4	0.399
NUMBER	29	0.282	4.3	0.463	6.1	0.379	20	0.378	6.3	0.357	9.1	0.352
ORGANIZATION	19	0.268	6.4	0.489	11.0	0.579	19	0.262	7.3	0.323	9.9	0.304
TIME	18	0.314	6.2	0.596	9.6	0.443	14	0.384	6.3	0.285	7.9	0.387
LIVING_THINGS	8	0.204	5.3	0.596	8.1	0.805	7	0.231	3.7	0.110	6.9	0.308
EVENT	0	-	-	-	-	-	3	0.205	7.0	-	9.3	-
ASTRO	2	0.161	2.0	-	2.0	-	2	0.345	7.5	-	10.5	-
total	195	0.303	5.7	0.510	9.0	0.473	195	0.363	6.4	0.429	10.3	0.504

Table 6. Features of test collections and performance of systems for each answer category

Answer category	QAC1			QAC2		
	# of ques	MRR (AVG)	#Sys5/ #Sys All	# of ques	MRR (AVG)	#Sys5/ #Sys All
ARTIFACT_AND_OTHER	44	0.280	0.370	46	0.266	0.357
:LAW	0	-	-	2	0.202	0.360
:MEDICAL	0	-	-	3	0.232	0.293
:PRIZE	0	-	-	1	0.267	0.400
:PRODUCT_CLASS	6	0.293	0.344	4	0.091	0.160
:PRODUCT_NAME	6	0.325	0.400	0	-	-
:WORK	10	0.317	0.453	8	0.420	0.570
:OTHER	22	0.247	0.330	0	-	-
:OTHER_NE	0	-	-	20	0.316	0.406
:OTHER_NON_NE	0	-	-	8	0.104	0.160
PERSON	42	0.358	0.467	45	0.492	0.604
:FOREIGN	11	0.375	0.448	23	0.494	0.609
:JAPANESE	31	0.352	0.473	22	0.490	0.600
LIVING_THINGS	8	0.204	0.258	7	0.231	0.291
:ANIMAL	1	0.267	0.333	5	0.294	0.464
:PLANTS	5	0.136	0.187	2	0.072	0.100
:OTHER	2	0.345	0.400	0	-	-
ASTRO	2	0.161	0.233	2	0.345	0.420
EVENT	0	-	-	3	0.205	0.360
ORGANIZATION	19	0.268	0.393	19	0.262	0.341
:COMPANY	12	0.265	0.411	6	0.310	0.407
:POLITICS	3	0.400	0.556	3	0.467	0.720
:SPORTS	2	0.212	0.267	1	0.031	0.120
:OTHER	2	0.142	0.167	9	0.188	0.280
LOCATION	33	0.331	0.424	39	0.400	0.527
:COUNTRY	14	0.407	0.510	8	0.512	0.675
:STATE	1	0.434	0.467	3	0.224	0.307
:PREFECTURE	3	0.080	0.156	10	0.461	0.596
:CITY	3	0.323	0.467	6	0.341	0.473
:CAPITAL	3	0.505	0.644	1	0.640	0.800
:TOWN	2	0.278	0.333	2	0.216	0.320
:SPOT	5	0.267	0.347	6	0.331	0.413
:NATURE	2	0.084	0.100	3	0.376	0.507
NUMBER	29	0.282	0.379	20	0.378	0.502
:NUMBER	3	0.200	0.244	0	-	-
:QUANT	21	0.306	0.397	13	0.331	0.452
:MONEY	3	0.298	0.489	2	0.782	0.880
:ORDER	0	-	-	4	0.316	0.460
:PERCENT	2	0.130	0.233	1	0.420	0.560
TIME	18	0.314	0.422	14	0.384	0.523
:DATE	14	0.352	0.486	7	0.504	0.709
:PERIOD	4	0.181	0.200	5	0.277	0.360
:TIME	0	-	-	2	0.229	0.280
total	195	0.303	0.402	195	0.363	0.472

Table 5. # of questions and performance of systems for each answer category

making any progress?

Table 6 shows the features of test collections and their correlation coefficient (r) versus MRR(AVG) for each basic answer category. For most of the basic categories, the value of #RD and #AS|RD increased, though correlation coefficient of the features declined, suggesting the effect of other factors. Questions on Time and Living_things in QAC2 seem to be more difficult than QAC1 judging from the decrease of #RD and #AS|RD. Nevertheless, the MRR(AVG) of these categories is higher in QAC2. Based on these features, it seems that we are making progress at least for Time and Living_things.

We should also notice that the performance for Organization and Artifact_and_other, unlike other basic categories, did not improve. Artifact_and_other in our classification includes miscellaneous questions including those asking various peripheral named entities and common nouns, and let us focus on Organization. Questions on Organization do not seem to be more difficult than other categories judging from the values of #RD and #AS|RD. However, the correlation coefficient of these features versus MRR(AVG) in QAC2 are lower than average, suggesting that they may be seriously affected by other factors.

Figure 5 gives the performance of systems for questions asking ORGANIZATION. In QAC1, the large portion of questions lies at the interval of 0.2 - 0.3. On the other hand, the RR(AVG) shows larger variance in case of QAC2.

The features of test collections and the performance of systems for subcategories of Organization is given in Table 7. The MRR(AVG) of these subcategories increased in most subcategories in QAC2. The only exception is a question on SPORTS, which seems to be relatively difficult, judging from low values of #RD and #AS|RD. Note that in QAC1, more than half of the questions on Organization ask Company, a subcategory not so difficult for systems. On the other hand, in QAC2, nearly half of the questions ask

category	QAC1					QAC2				
	# of ques	#RD	#AS|RD	MRR(AVG)		# of ques	#RD	#AS|RD	MRR(AVG)	
				value	*SD*				value	*SD*
ORGANIZATION	19	6.4	11.0	0.268	0.161	19	7.3	9.9	0.262	0.191
:COMPANY	12	5.5	10.3	0.265	0.123	6	9.7	13.8	0.310	0.190
:POLITICS	3	12.7	19.7	0.400	0.236	3	12.7	18.3	0.467	0.112
:SPORTS	2	6.5	10.5	0.212	0.145	1	4.0	6.0	0.031	0.000
:OTHER	2	2.0	2.5	0.142	0.075	9	4.3	5.0	0.188	0.156

Table 7. Features of test collections and performance of systems for subcategories of ORGANIZATION

Other organization, such as names of university, court, music band, union, and so on. And the low MRR(AVG) of this subcategory seems to influence the poor performance of Organization as a whole in QAC2. The results of the above analysis lead us to the conclusion that categories of questions do affect the performance of systems.

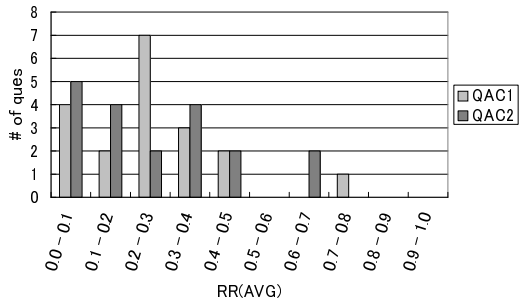


Figure 5. Performance of systems for questions on Organization

4 Proposals for future QA tasks

In this section, we will identify issues the future QAC tasks need to address, based on the discussions in the previous section. We identified two features of test collections which have moderate correlation with the performance of systems in QAC1 and 2. Using these features, we can cut out questions that are supposed to be 'too easy' from the question set in advance.

However, it should be noted that the features do not fully represent the difficulty of answering questions. Each feature was independently applied, though the difficulty of a question may be affected by multiple features. Furthermore, there may be features affecting other modules of QA systems. For example, the difficulty of IE(information extraction) in the QAC2 test collection may be different from that of QAC1.

The problem is that we have no way to test the effect of the features on each module for now. We need data on performance of particular modules of systems to test if candidate features actually affect the systems' performance on the module. An example of the data is the number or ratio of systems that can return at least one relevant document for each question, to be used for the test on IR. Such data will be helpful to detect new features that affect the performance of basic modules of QA systems, and investigate new evaluation measure to compare the difficulty of multiple test collections more precisely using several features. Furthermore, the data may also serve to characterize questions of each test collection, not to mention each system's analysis of evaluation results.

5 Conclusion

We analyzed NTCIR QAC1 and 2 test collections to evaluate the difficulty of the test collections and see if we are making progress. The analysis identified two features of test collections which have moderate correlation with the performance of systems in QAC1 and 2. The performance seems to be also influenced by other factors. Answer categories of questions also affect the performance of systems. The difficulty of the two test collections were compared, and the QAC2 test collection seemed to be easier than QAC1 in terms of the features. We seem to be making progress at least for questions on some answer categories. We also made a proposal for the future QAC tasks in respect of data needed for evaluation using multiple test collections.

References

[1] E. M. Voorhees. The TREC 8 Question Answering Track Report. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 2000.

[2] E. M. Voorhees. Overview of the TREC 2002 Question Answering Track. *Proceedings of the*

Eleventh Text REtrieval Conference (TREC 2002), 2003.

[3] E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, 2004.

[4] J.Fukumoto, T.Kato, and F.Masui. Question Answering Challenge (QAC1): An Evaluation of Question Answering Task at NTCIR Workshop3. *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[5] N. Kando. Overview of the Fourth NTCIR Workshop. *Working Notes of the Fourth NTCIR Workshop Meeting*, 2004.

[6] J. Fukumoto, T. Kato, and F. Masui. Question Answering Challenge for Five ranked answers and List answers - Overview of NTCIR4 QAC2 Subtask 1 and 2 -. *Working Notes of the Fourth NTCIR Workshop Meeting*, 2004.

[7] M. Nomoto, M. Sato, and H. Suzuki. NTCIR-3 QAC Experiments at Matsushita. *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[8] S. Sekine and H.Isahara. IREX Project Overview. *Proceedings of the IREX Workshop*, Sept 1999.

A APPENDIX

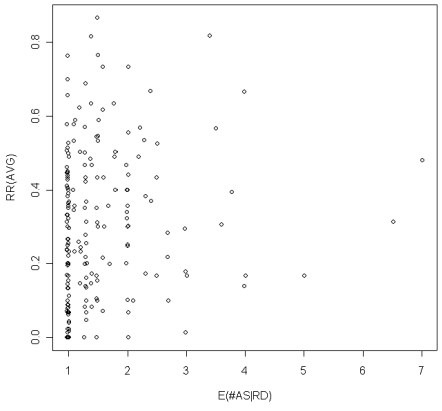


Figure 6. RR(AVG) vs. E(#AS|RD)

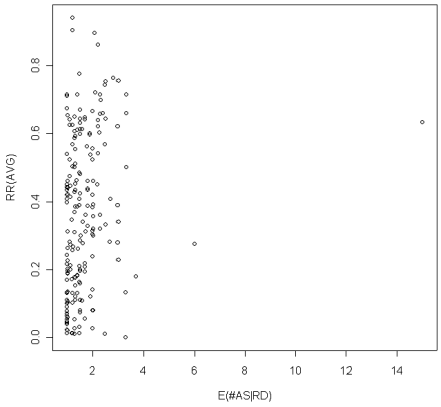


Figure 7. RR(AVG) vs. E(#AS|RD)

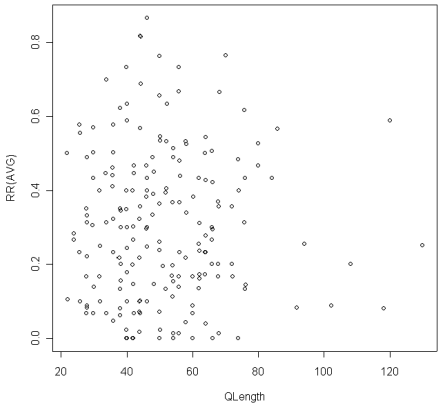


Figure 8. RR(AVG) vs. QLength

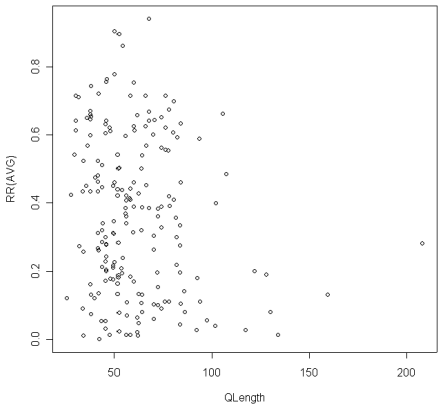


Figure 9. RR(AVG) vs. QLength