NTCIR-4 QAC Experiments at Matsushita
-Analysis of QAC1/QAC2 test collections-

The 4th NTCIR4 Workshop Meeting
June 2nd-4th, 2004
NII

Masako Nomoto, Yoshio Fukushige, Mitsuhiro Sato, Hiroyuki Suzuki
Network Systems Development Center,
Matsushita Electric Industrial Co.,Ltd.

Motivation

- **Problems with evaluation across multiple test collections**
- **Goal of this study**

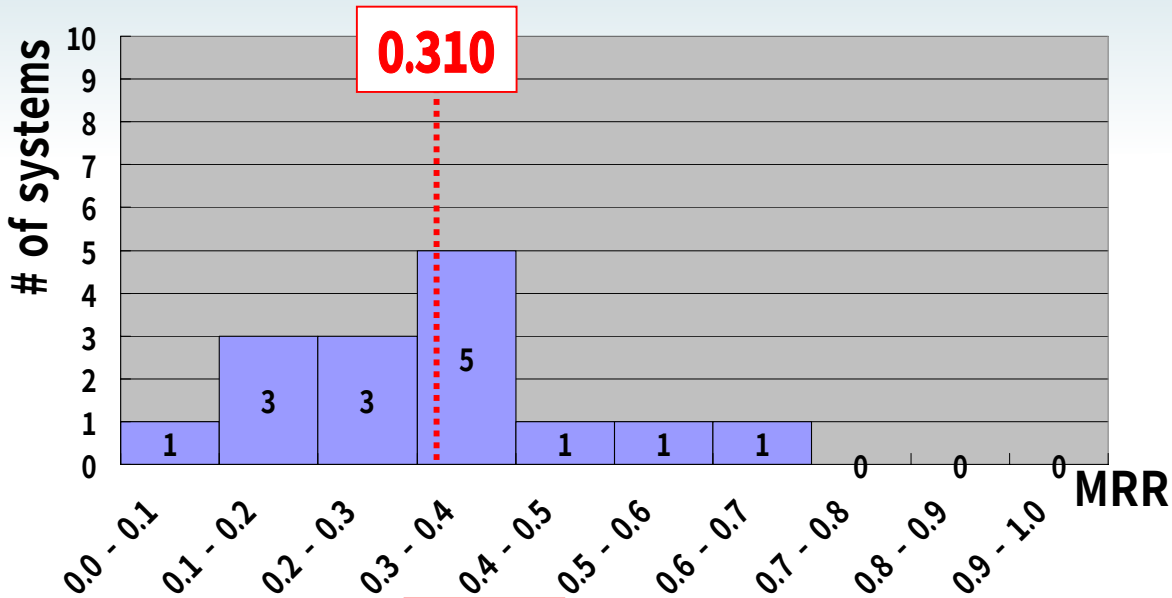
  to answer the following questions:

  - ***Are we making any progress?***
  - ***What kinds of questions are difficult or easy to answer correctly for QA systems?***
  - ***What kinds of features of a test collection affect the difficulty of QA?***

Performance of Systems in Subtask1 of QAC1/QAC2

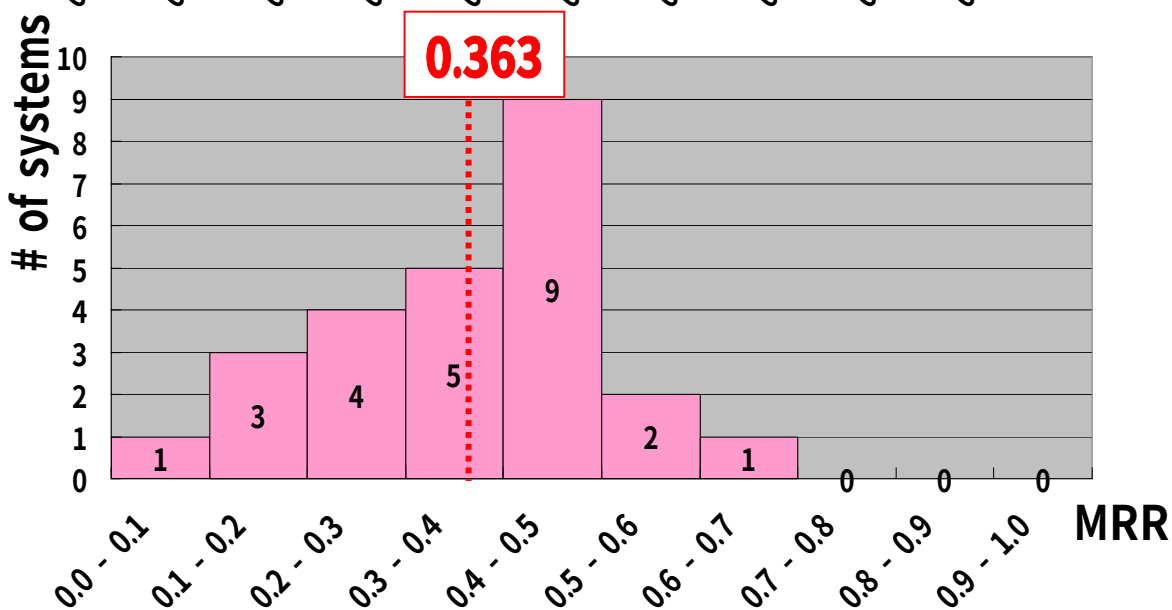
QAC1

of systems: 15
# of questions: 195
average of MRR: 0.310
variance: 0.022
standard deviation: 0.150



QAC2

of systems: 25
# of questions: 195
average of MRR: 0.363
variance: 0.021
standard deviation: 0.145



Measuring performance of systems

for a question

- *RR(AVG)* : the average of the RR(reciprocal rank)s of all the systems

  $RR(AVG) = AvgSys5 * M(Sys5)) / M(SysAll)$

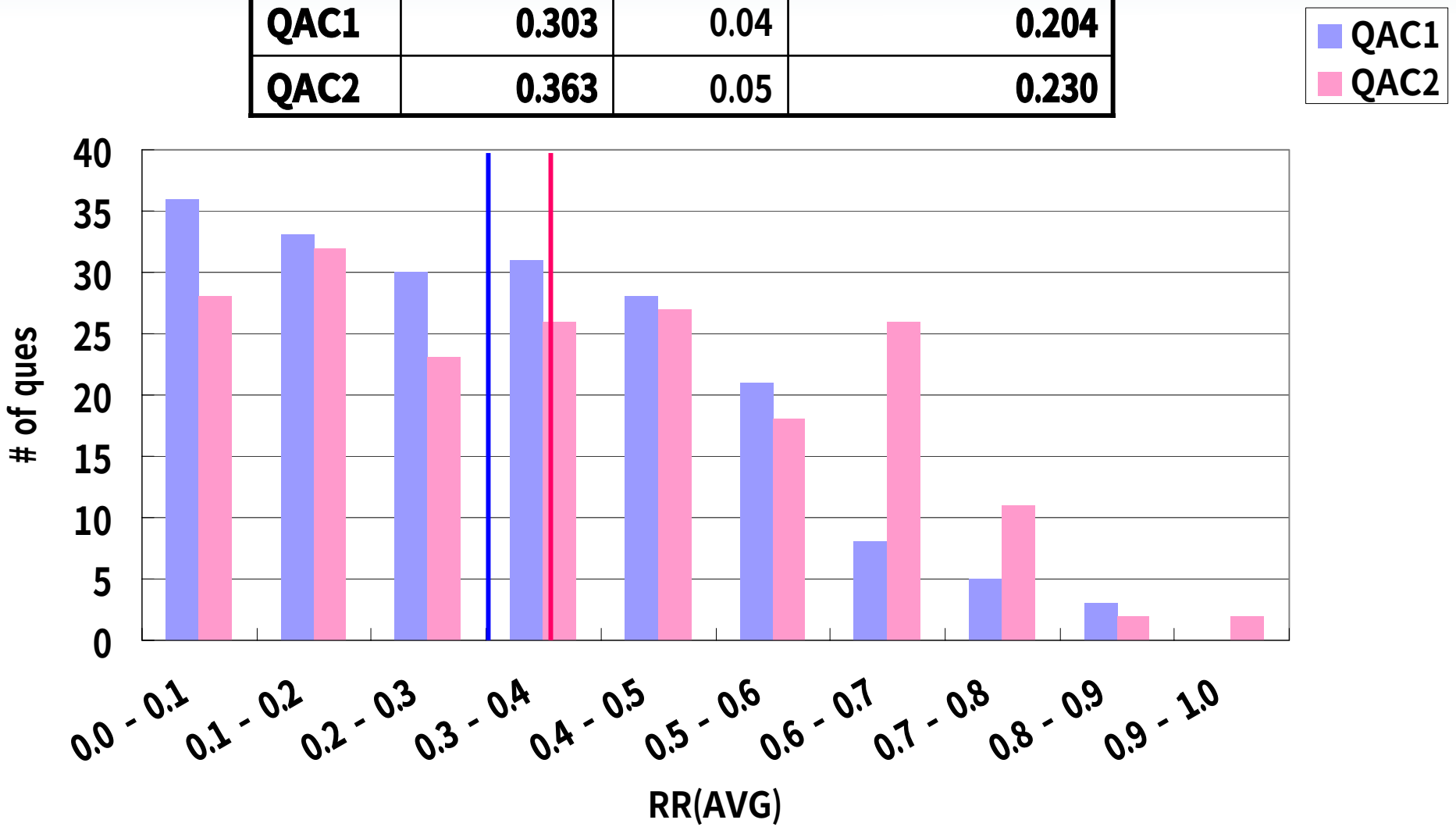
  - *AvgSys5* : The average of the RRs of the systems that obtained more than zero in RR.
  - *N(Sys5)* : the number of systems that returned the correct answer in up to the fifth place
  - *M(SysAll)*: the number of all the systems participated

for a set of questions

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

Question-wise comparison of performance of systems

	MRR(AVG)	variance	standard deviation
QAC1	0.303	0.04	0.204
QAC2	0.363	0.05	0.230



Testing Features of test collections

- **Selection of features**
  - supposed to affect the difficulty of questions for some modules of typical QA system
  - calculated automatically with questions, answer strings, and documents in the test collection
- **Methods**
  - Scatter diagrams and correlation coefficient between the features of test collections and performance measures(N(Sys5), RR(AVG))

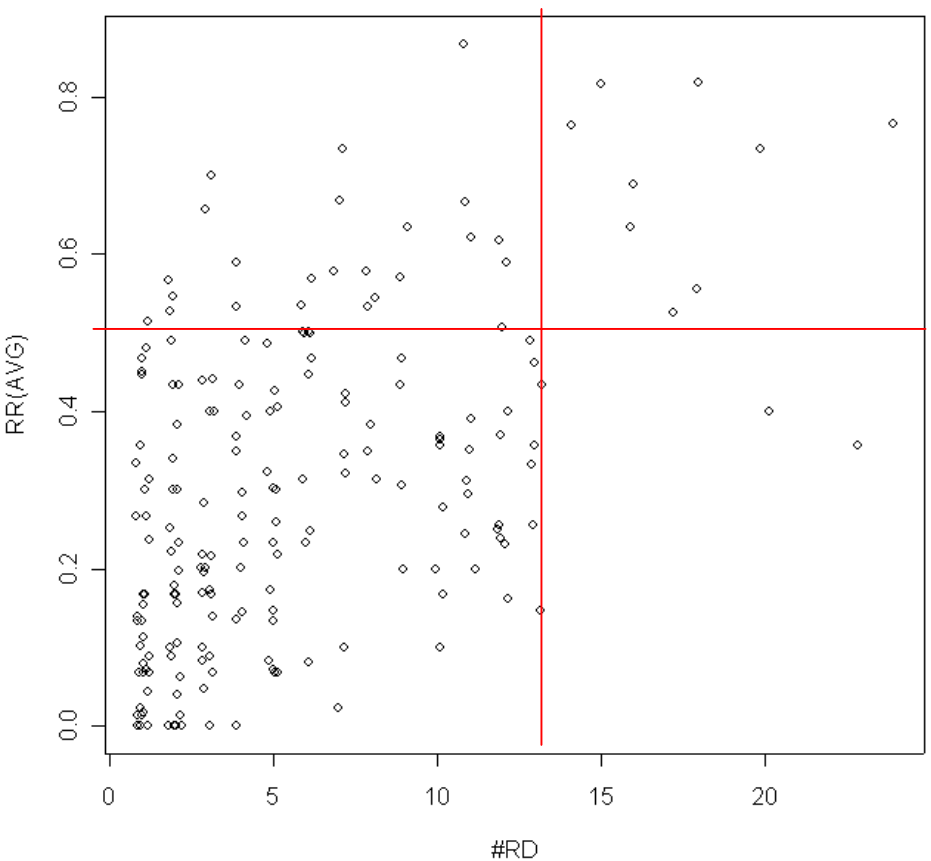
Features of questions

- **#RD:**
  - the number of relevant documents for a question
- **#AS|RD:**
  - the number of answer strings in relevant documents for a question
- **E(#AS|RD):**
  - the averaged number of answer strings in a relevant document for a question
- **QLength:**
  - the length of a question

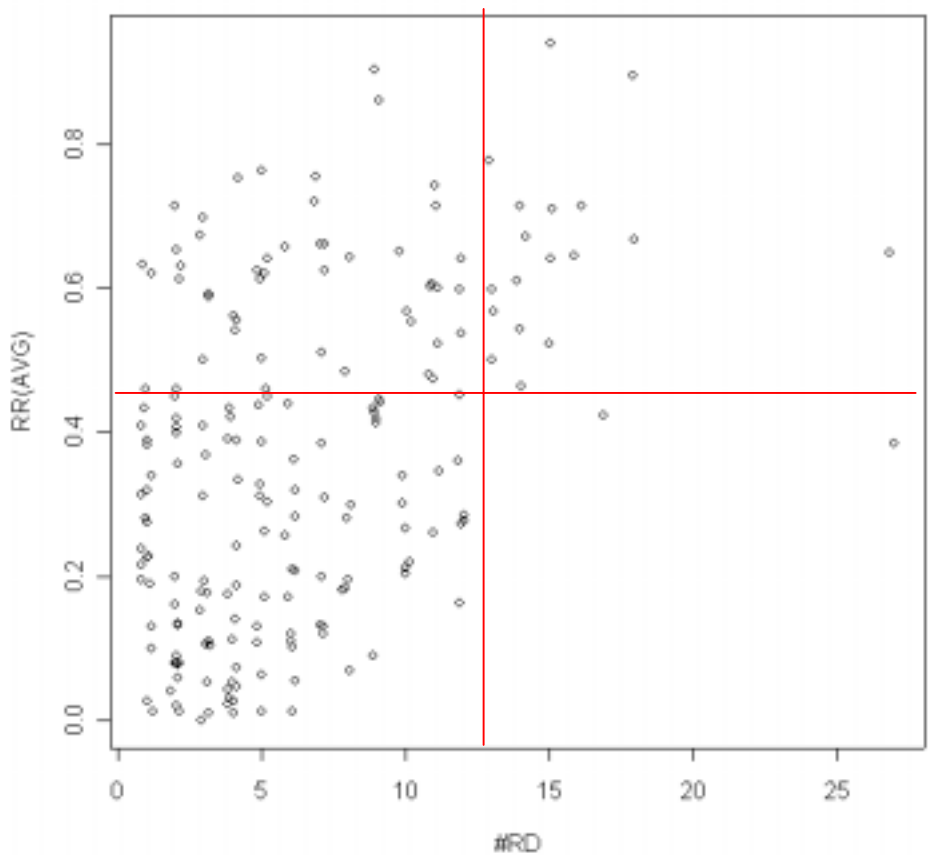
Scatter diagrams: RR(AVG) vs. #RD

#RD: the number of relevant documents for a question

QAC1



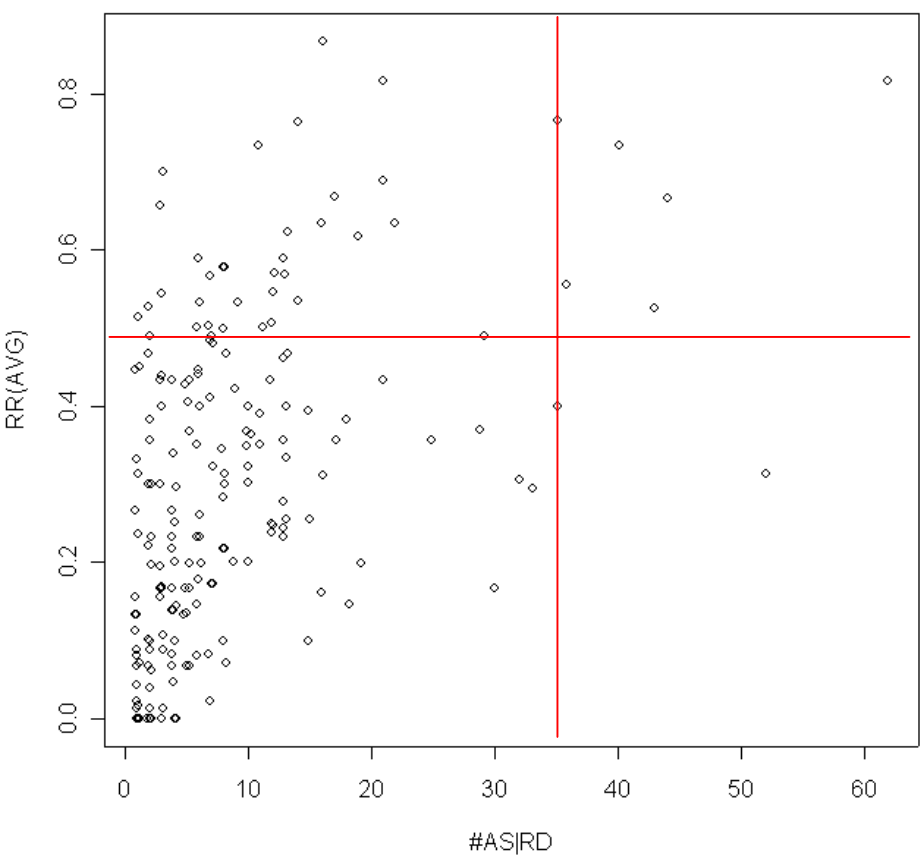
QAC2



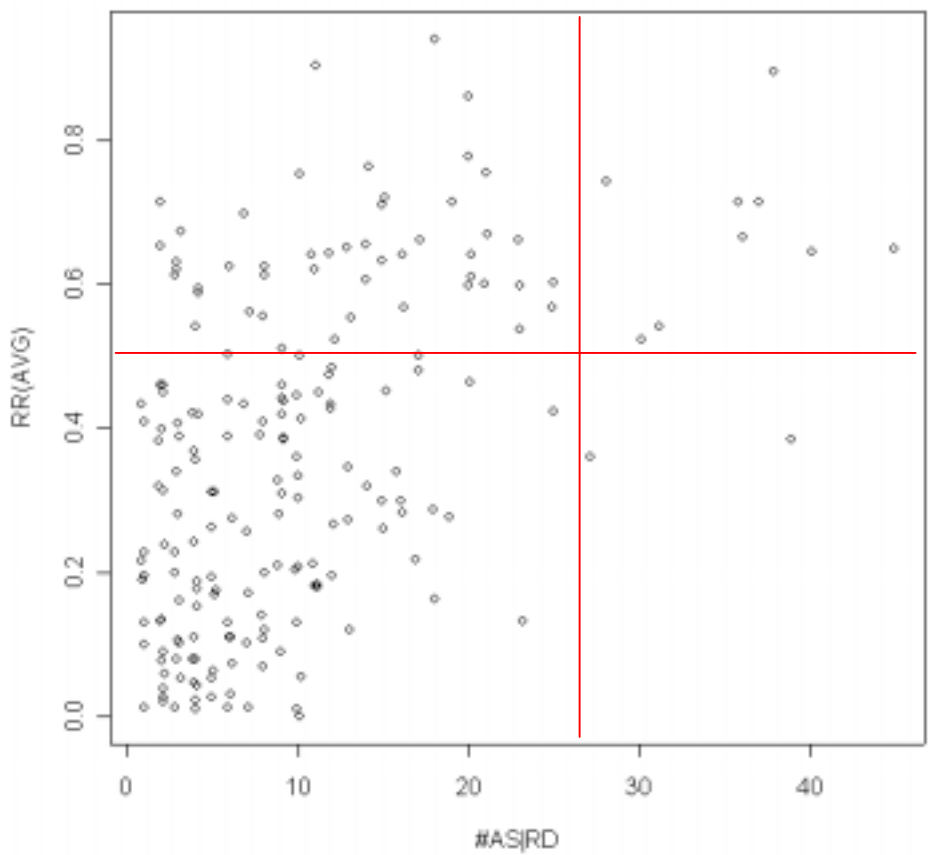
RR(AVG) vs. #AS|RD

#AS|RD: the number of answer strings in relevant documents

QAC1



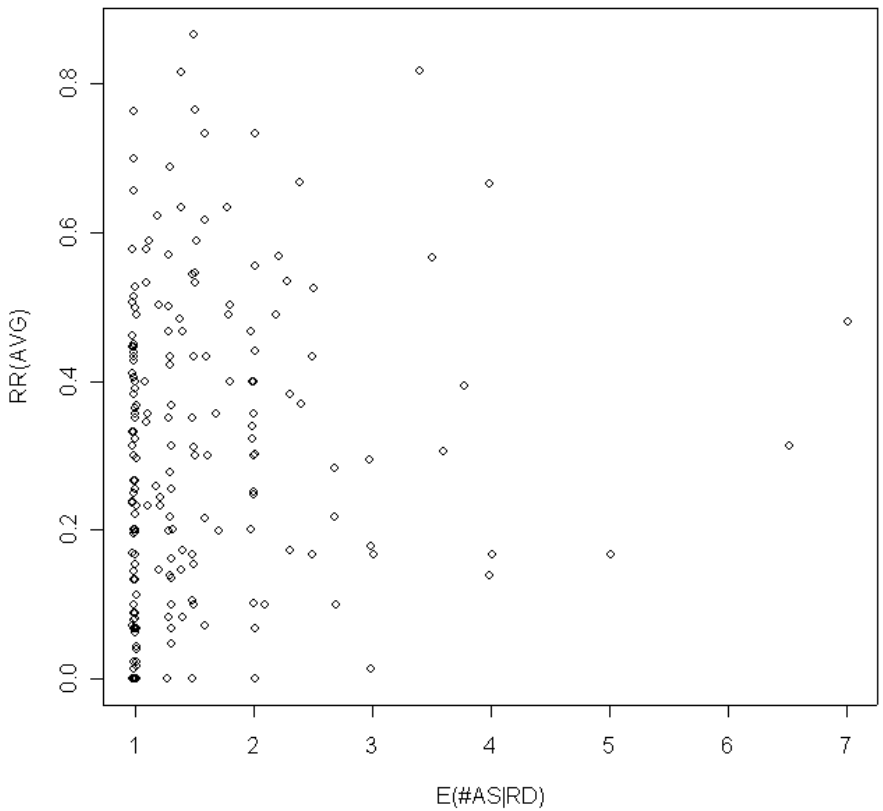
QAC2



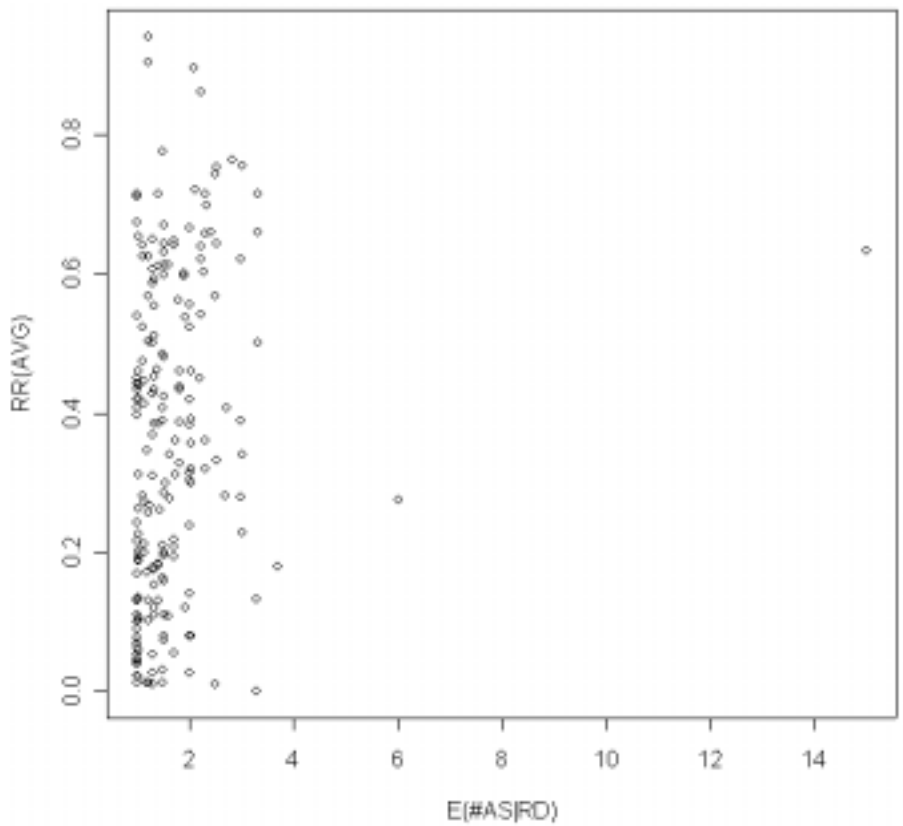
RR(AVG) vs. E(#AS|RD)

E(#AS|RD): the averaged number of answer strings in a relevant document for a question

QAC1



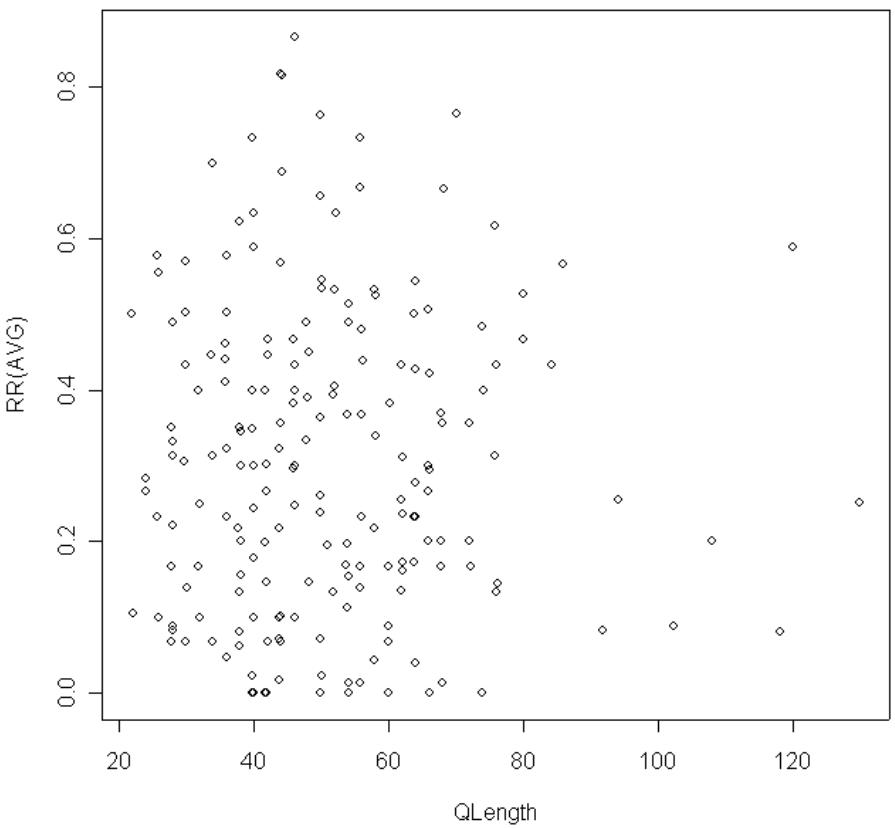
QAC2



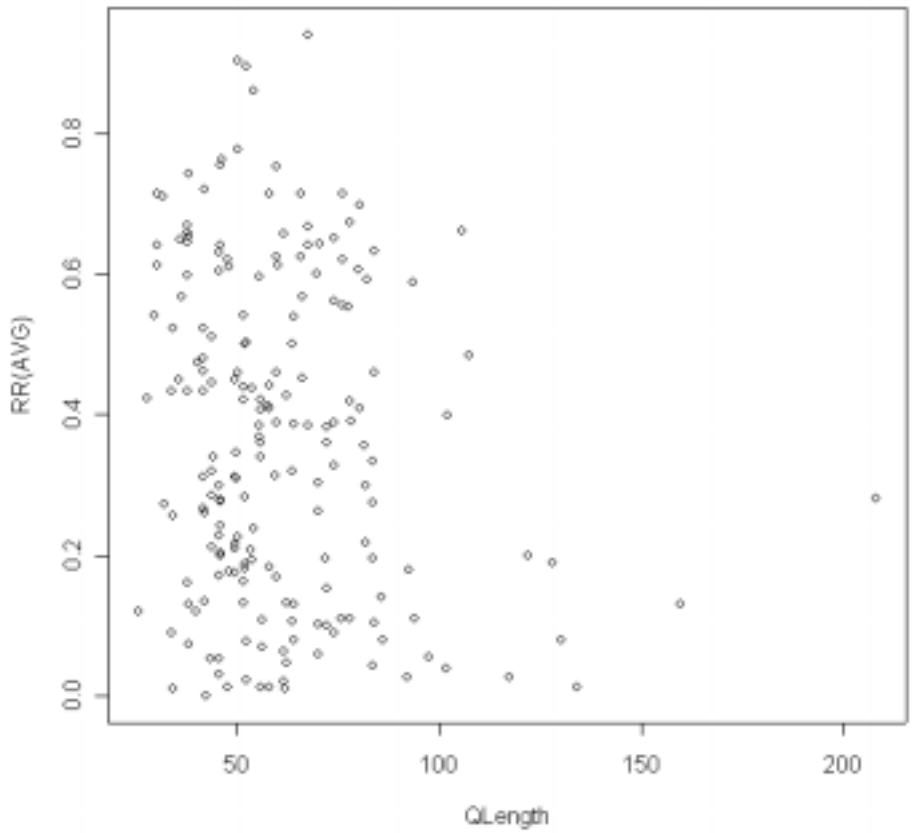
RR(AVG) vs. QLength

QLength: the length of a question

QAC1



QAC2



Correlation coefficient between features of test collections and performance measures

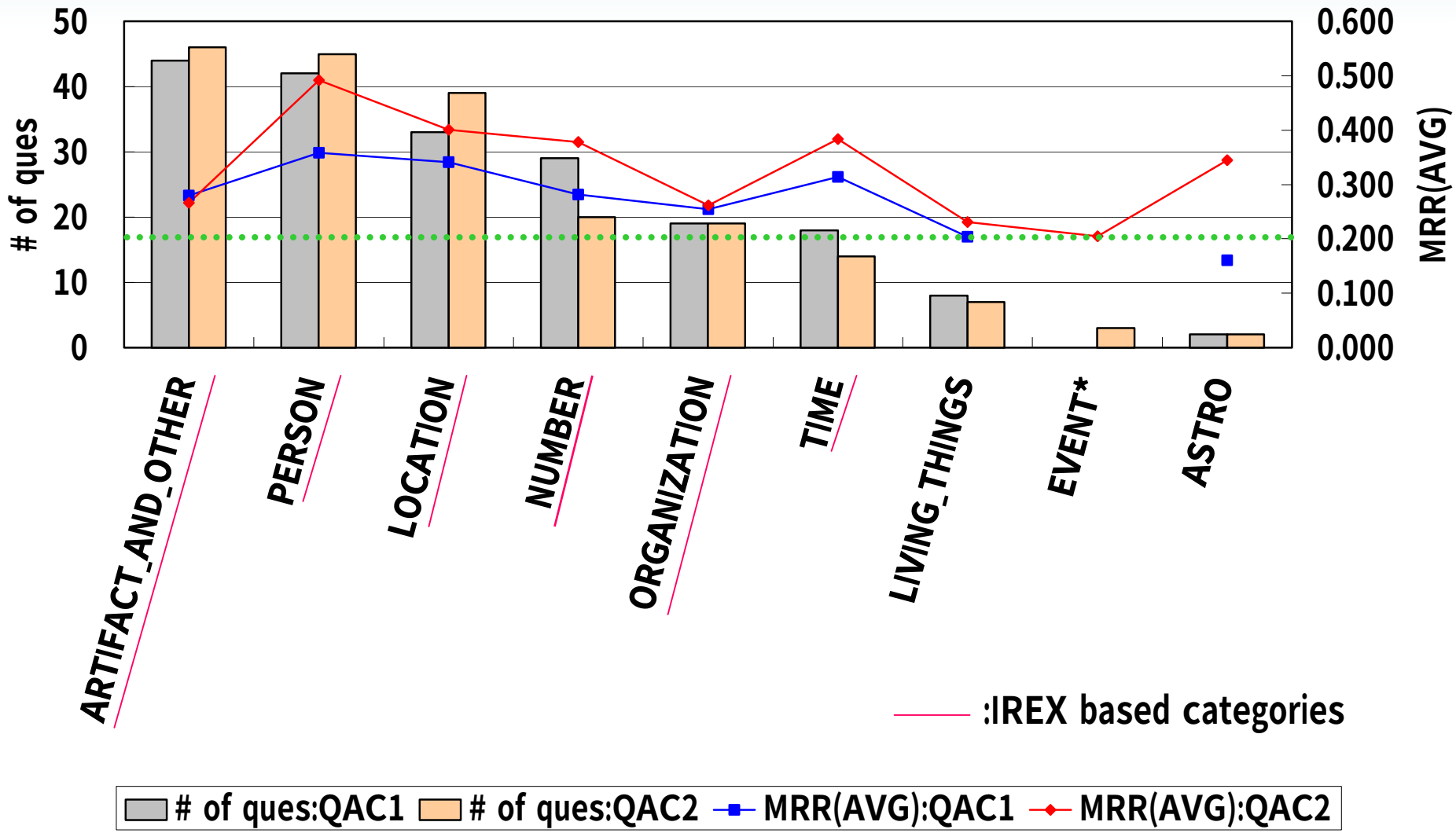
QAC1

		#RD	#AS∣RD	E(#AS∣RD)	QLength
average		5.7	9.0	1.5	51.2
variance		23.0	94.6	0.8	330.5
correlation coefficient	N(Sys5)	0.536	0.510	0.152	0.011
	MRR(AVG)	0.510	0.473	0.136	0.003

QAC2

		#RD	#AS∣RD	E(#AS∣RD)	QLength
average		6.4	10.3	1.7	61.0
variance		22.2	74.0	1.4	567.3
correlation coefficient	N(Sys_5)	0.440	0.501	0.213	-0.122
	MRR(AVG)	0.429	0.504	0.196	-0.174

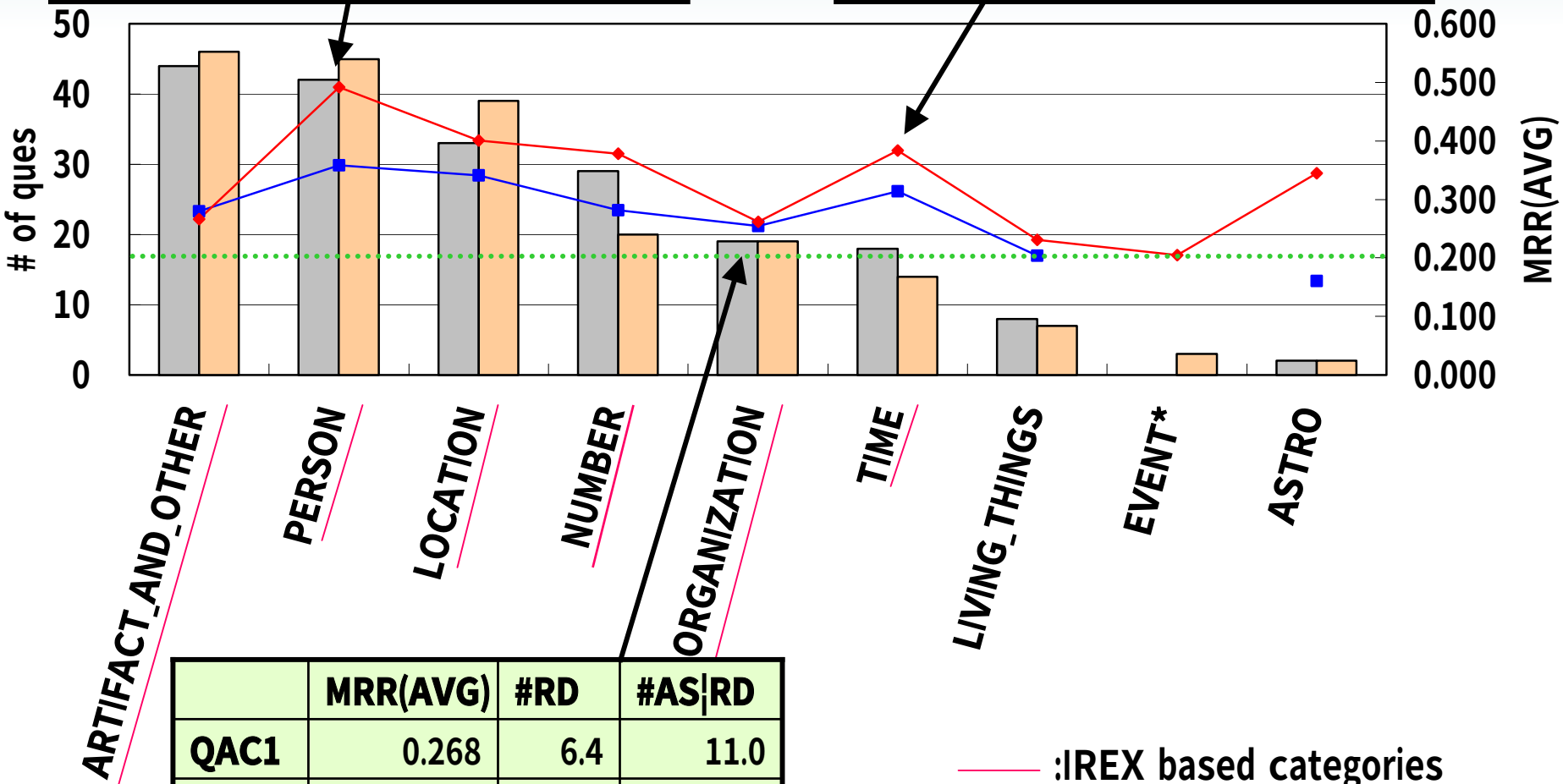
Answer categories and performance of systems



Answer categories and performance of systems

	MRR(AVG)	#RD	#AS|RD
QAC1	0.358	6.4	11.2
QAC2	0.492	6.8	13.2

	MRR(AVG)	#RD	#AS|RD
QAC1	0.314	6.2	9.6
QAC2	0.384	6.3	7.9



	MRR(AVG)	#RD	#AS|RD
QAC1	0.268	6.4	11.0
QAC2	0.262	7.3	9.9

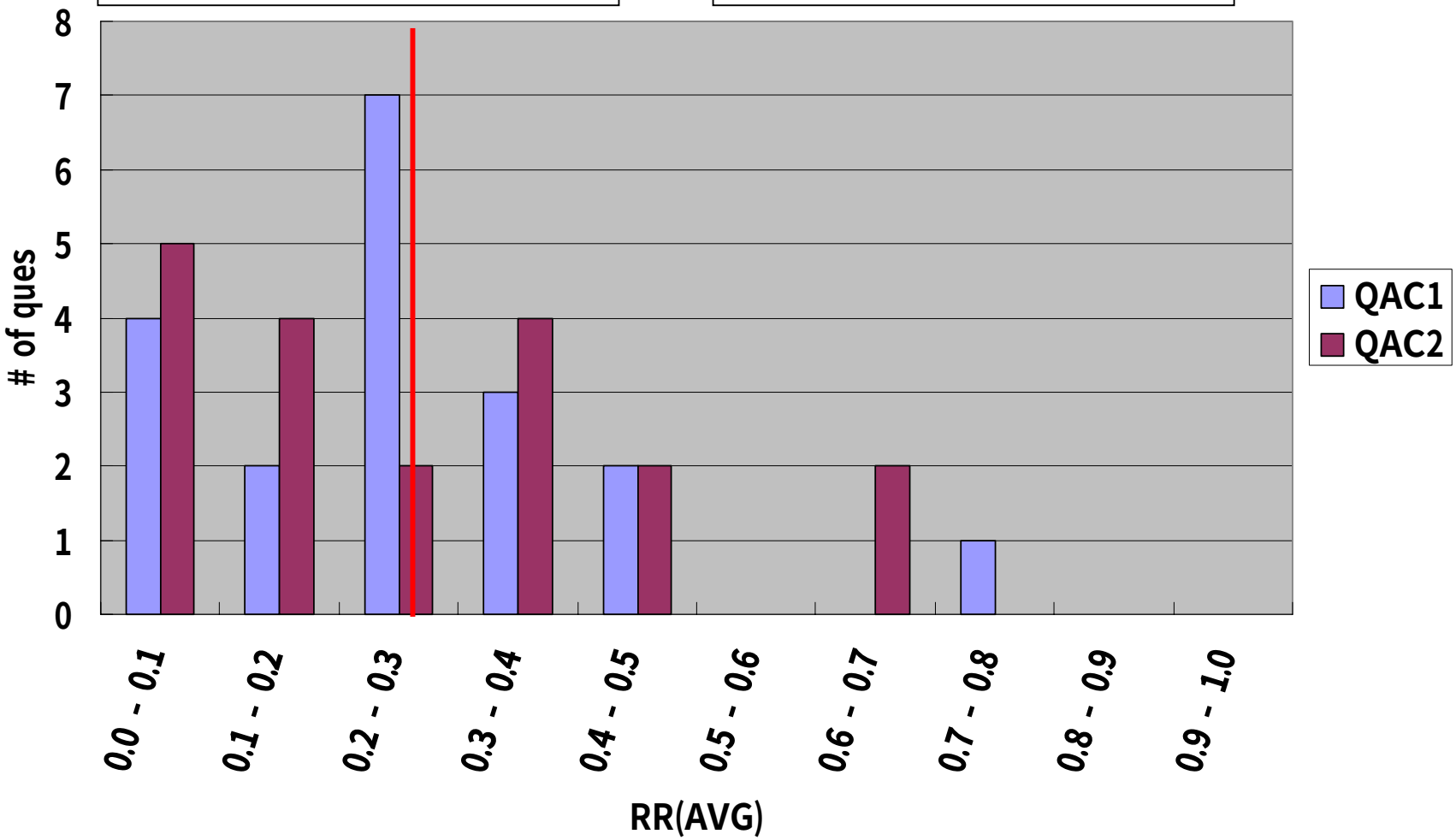
:IREX based categories

# of ques:QAC1 # of ques:QAC2 MRR(AVG):QAC1 MRR(AVG):QAC2

Performance of systems for questions on ORGANIZATION

QAC1
MRR(AVG): 0.268
standard deviation: 0.161

QAC2
MRR(AVG): 0.262
standard deviation: 0.194



Subcategories of ORGANIZATION

QAC1

categories	# of ques	#RD	#AS¦ RD	MRR (AVG)	STD EVP
ORGANIZATION	19	6.4	11.0	0.268	0.161
:COMPANY	**12**	5.5	10.3	0.265	0.123
:POLITICS(*)	3	12.7	19.7	0.400	0.236
:SPORTS	2	6.5	10.5	0.212	0.145
:OTHER	2	**2.0**	**2.5**	**0.142**	0.075

QAC2

categories	# of ques	#RD	#AS¦ RD	MRR (AVG)	STD EVP
ORGANIZATION	19	7.3	9.9	0.262	0.191
:COMPANY	6	9.7	13.8	0.310	0.190
:POLITICS(*)	3	12.7	18.3	0.467	**0.112**
:SPORTS	1	4.0	6.0	0.031	0.000
:OTHER	**9**	4.3	5.0	0.188	0.156

*: names of political parties

Summary

- Analysis of QAC1/QAC2 test collections
  - The questions of QAC2 test collection seem to be easier than those of QAC1 at least in terms of IR and answer selection.
  - We seem to be making progress, at least for questions on some answer categories.
  - The features of questions (#RD, #AS¦RD) have moderate correlation with the performance of systems.
  - Answer categories of questions also affect the performance of systems.