Overview of the NTCIR-8 Community QA Pilot Task (Part II): System Evaluation

Tetsuya Sakai* Daisuke Ishikawa† Noriko Kando†
*Microsoft Research Asia †National Institute of Informatics
tetsuyasakai@acm.org, {dais,kando}@nii.ac.jp

ABSTRACT

This paper describes the methods we used for evaluating the runs submitted to the NTCIR-8 Community QA Pilot Task, and report on the official results. Moreover, we also describe a set of more systematic variants of the official evaluation methods, and re-evaluate the runs. For details on the NTCIR-8 Community QA test collection and the task specifications, we refer the reader to Overview Part I [3]. For details on the task participants' approaches, we refer the reader to their papers [4, 6, 13].
Keywords: community QA, evaluation, pyramid.

1. INTRODUCTION

This paper describes the methods we used for evaluating the runs submitted to the NTCIR-8 Community QA Pilot Task, and report on the official results. Moreover, we also describe a set of more systematic variants of the official evaluation methods, and re-evaluate the runs. For details on the NTCIR-8 Community QA test collection and the task specifications, we refer the reader to Overview Part I [3]. For details on the task participants' approaches, we refer the reader to their papers [4, 6, 13].

Table 1 shows the four teams that participated in the NTCIR-8 pilot task. A total of 13 runs were evaluated officially. These runs include three baseline runs generated by the organisers:

BASELINE-1 which ranks answers at random (in fact, it preserves the order of the answers in the formal run data set, as the data set already contains answers shuffled at random);

BASELINE-2 which ranks answers in decreasing order of answer length (designed to approximate the comprehensiveness of answers); and

BASELINE-3 which ranks answers by timestamp (assuming that fresh answers are better than old ones).

The remainder of this paper is organised as follows. Section 2 describes the evaluation methods we used for producing the official results for the NTCIR-8 pilot task. Section 3 reports on the official results and provides some analysis based on them. Section 4 discusses a set of more systematic variants of the official evaluation methods, and re-evaluates the runs. Finally, Section 5 concludes this paper and discusses future directions.

NTCIR-8 Workshop Meeting, 2010, Tokyo, Japan.

Table 1: Participating teams.

team name	organisation	#runs
ASURA	National Institute of Informatics	2
BASELINE	Organisers	3
LILY	Shirayuri College	3
MSRA+MSR	Microsoft Research Asia and Redmond	5

2. OFFICIAL EVALUATION METRICS

Participants were asked to submit runs of the following format:

```
<Q_ID>,<A_ID ranked at 1>,<A_ID ranked at 2>,...
```

where Q_ID and A_ID are questions IDs and answer IDs. Thus, for every question, the participating systems were to rank all the answers in decreasing order of answer quality. As we had 1,500 formal run questions, each run file contains exactly 1,500 lines.

We evaluated (a) the top-ranked answers only, and (b) the entire ranked lists of answers, using different evaluation metrics and "gold standards" as described below.

2.1 Using the "Best" Answers

As described in Overview Part I [3], the Yahoo Chiebukuro Data contains exactly one "best" answer (BA) for each question, selected by the asker[1]. The selection of a BA can be very subjective (the answer may be what that particular asker likes regardless of its correctness or quality), and there often exist answers that are of highly quality, yet were not chosen as the BA. So the BA data is not necessarily ideal for building a system that selects high-quality answers. On the other hand, the advantage of the BA data is that "it's already there" for every question: no additional manual assessments are required.

The evaluation metric we use based on the BA data is *hit at 1* (or *precision* at 1), which we denote by BA-Hit@1. For a given question, let $I(r) = 1$ if the answer at rank r is "relevant" according to a gold standard data, and let $I(r) = 0$ otherwise. In the case with the BA data, an answer is relevant to the question if and only if it is the BA. Then

$$BA\text{-}Hit@1 = I(1)\ . \quad (1)$$

Hence, Mean BA-Hit@1 is simply the number of questions for which the system correctly detected the BA.

2.2 Using a Pyramid Approach

As was mentioned in Overview Part I [3], we hired four assessors to independently assess every answer for our formal run question

[1]The Yahoo Chiebukuro site also has a voting mechanism for selecting the best answer, but our data contains the asker's best answers only.

Table 2: Mapping relevance patterns to relevance levels for the answers to 1500 formal run questions.

(a) pattern	(b) #answers	(c) level	(d) #answers
AAAA	1301	$L3$	2806
AAAB	1505		
AABB	1525	$L2$	2910
ABBB	1385		
BBBB	1241	$L1$	1677
AAA	2		
AAB	14		
ABB	76		
BBB	231		
AA	1		
AB	7		
BB	105		
A	1	$L0$	50
B	32		
(C's only)	17		
total	7443	total	7443

set containing 1,500 questions. As each assessor labeled each answer with "A" (highly-quality), "B" (medium-quality) or "C" (low-quality), we obtained *relevance patterns* as shown in Table 2 Column (a). For example, "AAAB" means that the answer was rated A by three assessors and rated B by one assessor, and "AAA" means that the answer was rated A by three assessors and rated C by one assessor. The C's are not shown explictly in the table as we do not regard them as votes of confidence. Note the way we arranged the patterns in the table: for now, we assume that the number of positive votes is more important than whether each vote is an A or a B. For example, "BBBB" is placed above "AAA." (However, below we show that these two different patterns are treated equally in our evaluation.)

In order to reduce the problem of evaluating a ranked list of of answer IDs with the above data into that of evaluating ranked lists with graded-relevance evaluation metrics designed for information retrieval, we mapped the relevance patterns into *relevance levels*, $L3$ (highly relevant), $L2$ (relevant), $L1$ (partially relevant) and $L0$ (not relevant), as shown in Table 2 Column (c). This could have been done in several different ways, but we partitioned the relevance patterns so that the total numbers of $L3$-, $L2$- and $L1$-relevant answers are roughly the same. Note that, according to our definition, an $L3$-relevant answer is one that was *rated highly by many assessors*. We shall refer to this resultant gold-standard data set as the *Good Answers* (GA) data. Sakai *et al.* [12] have shown that the effect of using a mapping that is different from Table 2 on system ranking is small. In Section 4, we propose a more systematic way to define the pattern-to-level mapping.

The above method was inspired by the *pyramid method* used in text summarisation and question answering evaluation [7, 8]. The premise of this method is that different people have different views on which answers are correct or not, and the basic idea is to view this fact as a pyramid, where the top represents data with high inter-assessor agreements, while the bottom represents those with low inter-assessor agreements.

Table 3 shows the distribution of $L3, L2, L1$-relevant and $L0$ answers over questions. For example, there are 691 formal run questions with exactly one $L3$-relevant answers. On average, a formal run question has 1.87 $L3$-relevant answers, 1.94 $L2$-relevant answers, 1.12 $L1$-relevant answers, (4.93 "relevant" answers in total,) and 0.03 $L0$ (judged nonrelevant) answers. Hence our task is like document retrieval where almost all documents are at least somewhat relevant, and there are no unjudged documents.

We compared the BA data with the GA data for the 1,500 formal run questions: For 970 questions, the BA is $L3$-relevant; For 399 questions, the BA is $L2$-relevant; For 130 questions, the BA is $L1$-relevant; and for 1 question, the BA is $L0$. Hence, even though a BA reflects an opinion of one asker and may not always be reliable, the BA data for our question set seems reasonably reliable. The main problem with our BA data therefore seems to be *incompleteness*: as Table 2 shows, there are many excellent or very good answers besides the BAs.

Table 3: Distribution of $L3, L2, L1$-relevant and $L0$ answers over questions (GA). The "#q" columns show the number of questions.

#$L3$	#q	#$L2$	#q	#$L1$	#q	#$L0$	#q
1	691	1	417	0	712	0	1463
2	315	0	381	1	404	1	28
0	174	2	277	2	183	2	6
3	132	3	176	3	85	3	2
4	66	4	95	4	45	4	1
5	43	5	54	5	33		
6	29	6	38	6	13		
7	22	7	30	7	8		
8	13	8	12	8	5		
10	6	9	7	9	4		
9	2	12	4	10	3		
14	2	11	3	13	2		
18	1	15	2	11	2		
16	1	10	2	19	1		
15	1	19	1				
12	1	14	1				
11	1						
	1500		1500		1500		1500

The main problem with our BA data therefore seems to be *incompleteness*: as Table 2 shows, there are many excellent or very good answers besides the BAs.

Using the GA data, we compute four different evaluation metrics.

The first is *hit at rank 1* again, where *any* relevant answer ($L3$, $L2$ or $L1$) is counted as relevant:

$$GA\text{-}Hit@1 = I(1) \ . \tag{2}$$

For example, if the 1,500 BAs are regarded as a system output, then, since we have seen that only one of them is nonrelevant $L0$, this "system" achieves $GA\text{-}Hit@1 = 1499/1500 = 0.9993$. However, as GA-Hit@1 wastes the relevance levels, it rewards a system even if it returns an $L1$-relevant answer. Returning such an answer should not be difficult for systems, since the GA data contains as many as $7443 - 50 = 7393$ "right answers" for the 1,500 questions. Hence we define our primary evaluation metric by utilising graded relevance as shown below.

Let $g(r)$ denote the *gain* of the answer at rank r in a system's ranked list: for our GA data, let the gain value be 3 for each $L3$-relevant answer, 2 for each $L2$-relevant answer and 1 for each $L1$-relevant answer. (Sakai *et al.* [12] have shown that the effect of using more extreme choices of gain values on system ranking is small.) Similarly, let $g^*(r)$ denote the gain of the answer at rank r in an *ideal ranked list*, obtained by exhaustively listing up all $L3$-relevant, $L2$-relevant and then $L1$-relevant answers. Then, *normalised gain at 1* is defined as:

$$GA\text{-}nG@1 = g(1)/g^*(1) \ . \tag{3}$$

This is in fact the same as *normalised discounted cumulative gain* [5] at rank 1, since neither discounting nor gain cumulation does not apply at rank 1. For example, for a question that has at least one $L3$-relevant answer, a system that returns an $L3$-relevant answer at rank 1 receives $GA\text{-}nG@1 = 3/3 = 1$, while one that returns an $L1$-relevant answer at rank 1 receives only $GA\text{-}nG@1 = 1/3$. Note that $GA\text{-}Hit@1 = 1$ in either case. Also, using "flat" gain

values (giving 1 to each $L3$, $L2$, $L1$-relevant answer) implies that nG@1 is reduced to Hit@1.

Since multiple high-quality answers are possible with the GA data, we also evaluate the entire ranked lists of answers using two well-studied graded-relevance information retrieval metrics, namely a version of *normalised discounted cumulative gain* (nDCG) [5] and *Q-measure* (Q) [9]. The NTCIR-8 ACLIA IR4QA [11] and GeoTime [1] tasks also use these graded-relevance metrics for document retrieval evaluation.

With the GA data, we define *GA-nDCG* as follows:

$$GA\text{-}nDCG = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \quad (4)$$

where l is a document cut-off value. In our evaluation, we use $l = 20$ for convenience since the number of answers per question lies within $[2, 19]$. But this is actually the same as using $l = 2$ for questions with two answers, while using $l = 20$ for those with twenty answers, and so on, as both the system's ranked list and the ideal ranked list rank *all* answers to a given question and nothing else.

Let $cg(r) = \sum_{i=1}^{r} g(i)$ and $cg^*(r) = \sum_{i=1}^{r} g^*(i)$. These are known as the *cumulative gains* [5] for the system's ranked list and for the ideal one. Moreover, let $C(r) = \sum_{i=1}^{r} I(i)$. Using the GA data, we define GA-Q as follows:

$$GA\text{-}Q = \frac{1}{R} \sum_{r} I(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \quad (5)$$

where R is the total number of relevant ($L3$, $L2$ or $L1$) items, and β is a *persistence parameter* for penalising late arrival of relevant items: we let $\beta = 1$ [10]. Like nDCG, Q quantifies how the system's ranked list deviates from the ideal one.

2.3 Using Each Assessor's Favourites

As we have seen, the GA data was constructed in the spirit of the pyramid method. We tried an additional approach to constructing the gold standard, namely to collect only the "best" answers from each assessor and treat them all as correct. For a given question, if an assessor rated some answers as A, some as B and the others as C, we call those rated A his/her *favourite* answers. (If no answer was rated A, then those rated B are his/her favourite answers.) We can then take the union of each assessor's favourite answers for each topic to form a binary-relevance gold standard. We refer to this data set as *UFA (union of favourite answers)*. Note that an answer that was not any assessor's favourite (even if its relevance pattern was "BBBB") is treated as nonrelevant. Based on the binary-relevance UFA data, we compute hit at rank 1 again, which we denote by UFA-Hit@1.

Moreover, since the asker's BA can also be regarded as his/her "favourite" answer, we can also treat the asker as the fifth assessor and form a union of favourite answers from these five assessors. We refer to this gold standard set as *UFBA (union of favourite and best answers)*. Thus we combine the BA data and the assessments from four assessors. Based on the binary-relevance UFBA data, we compute hit at rank 1 again, which we denote by UFBA-Hit@1.

Table 4 shows the distribution of the number of relevant answers across topics for the UFA and the UFBA data. On average, a formal run question has 4.18 UFA-relevant answers, and 4.22 UFBA-relevant answers.

2.4 Evaluation Tool

All of the above metrics can be computed using a UNIX-based tool called `cqa_eval` that can be download from `http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en.html`. This tool is actually a small patch to a general purpose information retrieval evaluation tool called `ir4qa_eval2`, also available at the same URL. Details can be found in the README files.

We also released four different "qrels" (gold standard) files to the participants, ntcir8cqa-{BA, GA, UFA, UFBA}.qrels, which can be used with the tool.

Table 4: Distribution of relevant answers over questions (UFA and UFBA). The "#q" columns show the number of questions.

#relevant (UFA)	#q	#relevant (UFBA)	#q
2	353	2	358
3	328	3	334
4	207	4	211
5	156	5	157
1	119	6	111
6	107	1	98
7	69	7	68
8	51	8	51
10	21	9	22
9	20	10	21
11	16	11	16
13	14	13	14
12	13	12	13
14	10	14	10
17	6	17	6
19	3	19	3
18	3	18	3
16	3	16	3
15	1	15	1
	1500		1500

3. OFFICIAL RESULTS

3.1 Evaluation with the Full Question Set

Table 5 summarises the results of our formal run evaluation using the 1,500 questions. The runs are arranged in decreasing order of mean GA-nG@1, our primary evaluation metric. For GA-nG@1, we conducted a two-sided sign test for every adjacent run pair in the table, to roughly cluster the runs. We can observe that:

(1) On average, only three runs (MSRA+MSR-{2, 1, 4}) outperform BASELINE-2 (sorting by length) with all metrics[2].

(2) On average, MSRA+MSR-2 is the top performer in terms of GA-{nG@1, nDCG, Q} and {UFA, UFBA}-Hit@1.

(3) On average, MSRA+MSR-5 is the top performer in terms of BA-Hit@1 and GA-Hit@1.

(4) In terms of GA-nCG@1,
ASURA-1 >> MSRA+MSR-3;
MSRA+MSR-5 >> BASELINE-3 (sorting by freshness);
BASELINE-3 >> BASELINE-1 (random); and
BASELINE-1 >> LILY-3,
where "$X >> Y$" means "X significantly outperforms Y at the $\alpha = 0.01$ level".

The anomalous BA-Hit@1 value for MSRA+MSR-5 deserves an explanation: this is an artefact of our preliminary task design. We released the entire Yahoo! Chiebukuro corpus (with the BA data) to participants, and our formal run questions were drawn from the same corpus. That is, for this pilot task, we did not separate training and test data properly. The MSRA+MSR-5 run took the

[2]MSRA+MSR-4 and BASELINE-2 are equally effective in terms of BA-Hit@1.

Table 5: Mean performances with the full question set (1,500 questions). Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-2	0.4980	0.9967	**0.9211**	**0.9747**	**0.9690**	**0.9767**	**0.9807**
MSRA+MSR-1	0.4980	0.9967	0.9203	0.9741	0.9682	0.9753	0.9793
MSRA+MSR-4	0.4847	0.9973	0.9202	0.9745	0.9688	0.9733	0.9767
BASELINE-2	0.4847	0.9953	0.9170	0.9735	0.9680	0.9680	0.9753
ASURA-2	0.4840	0.9953	0.9166	0.9742	0.9689	0.9713	0.9793
ASURA-1	0.4813	0.9940	0.9140**	0.9734	0.9680	0.9680	0.9773
MSRA+MSR-3	0.4813	0.9960	0.8956	0.9679	0.9609	0.9580	0.9647
MSRA+MSR-5	**0.7773**	**0.9987**	0.8863**	0.9604	0.9499	0.9507	0.9733
BASELINE-3	0.3820	0.9940	0.8213**	0.9460	0.9359	0.9000	0.9113
BASELINE-1	0.2713	0.9920	0.7751**	0.9311	0.9169	0.8533	0.8607
LILY-3	0.1767	0.9887	0.6883	0.9142	0.9002	0.7733	0.7847
LILY-2	0.1767	0.9887	0.6883	0.9191	0.9081	0.7733	0.7847
LILY-1	0.1767	0.9887	0.6883	0.9096	0.8927	0.7733	0.7847

Table 6: Mean performances with the good question set (1,429 questions). Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-2	0.5003	0.9979	**0.9228**	**0.9753**	**0.9696**	**0.9769**	**0.9804**
MSRA+MSR-4	0.4899	0.9986	0.9221	0.9752	0.9695	0.9734	0.9762
MSRA+MSR-1	0.5003	0.9979	0.9220	0.9748	0.9688	0.9755	0.9790
BASELINE-2	0.4913	0.9958	0.9179	0.9738	0.9683	0.9685	0.9755
ASURA-2	0.4906	0.9958	0.9173	0.9744	0.9691	0.9713	0.9790
ASURA-1	0.4871	0.9944	0.9140**	0.9734	0.9680	0.9678	0.9769
MSRA+MSR-3	0.4822	0.9972	0.8981	0.9688	0.9619	0.9587	0.9650
MSRA+MSR-5	**0.7782**	**0.9993**	0.8889**	0.9614	0.9510	0.9503	0.9727
BASELINE-3	0.3835	0.9944	0.8240**	0.9468	0.9367	0.8985	0.9090
BASELINE-1	0.2708	0.9923	0.7741**	0.9309	0.9165	0.8495	0.8572
LILY-3	0.1777	0.9902	0.6880	0.9145	0.9003	0.7698	0.7817
LILY-2	0.1777	0.9902	0.6880	0.9194	0.9082	0.7698	0.7817
LILY-1	0.1777	0.9902	0.6880	0.9098	0.8926	0.7698	0.7817

following general approach: given a test question, retrieve similar questions from the training data, and utilise the features of the BAs of these similar questions to estimate the BA for the test question [13]. Thus, MSRA+MSR-5 inadvertently used the actual BA data directly for the formal run, and as a result managed to return the BA at rank 1 for as many as 1,493 questions out of the 1,500. Note that this is a flaw in our task design, and not the task participant's fault. We should remember however that the performance values of MSRA+MSR-5 do not represent those in a practical setting: in the real world, newly posted questions are often not identical to known questions already tagged with a BA.

If we ignore MSRA+MSR-5, it can be observed in Table 5 that BA-Hit@1 in fact agrees well with GA-nCG@1, our primary metric. Figure 1, which visualises Table 5, shows that in fact all of our metrics generally agree well with GA-nCG@1 in ranking systems.

Table 5 contains significance test results for only GA-nG@1 as this is the sort key, but Sakai *et al.* [12] report on some more: Let "$X > Y$" mean "X significantly outperforms Y at the $\alpha = 0.05$ level". According to BA-Hit@1,
ASURA-1 > BASELINE-3;
BASELINE-3 > BASELINE-1; and
BASELINE-1 > LILY-3.
According to GA-nDCG,
ASURA-2 >> MSRA+MSR-1;
MSRA+MSR-1 > BASELINE-2; and
BASELINE-2 > ASURA-1.
Moreover,
ASURA-1 >> MSRA+MSR-3;
MSRA+MSR-3 >> BASELINE-3;
BASELINE-3 >> BASELINE-1;
BASELINE-1 >> LILY-2;
LILY-2 >> LILY-3; and
LILY-3 >> LILY-1.
Acccording to GA-Q,
ASURA-2 > MSRA+MSR-2 even though the former slightly underperforms the former on average (0.9689 vs 0.9690): ASURA-2 outperforms MSRA+MSR-2 for 327 questions while MSRA+MSR-2 outperforms ASURA-2 for only 274 questions. Moreover,
MSRA+MSR-4 >> MSRA+MSR-1;
MSRA+MSR-1 > BASELINE-2;
ASURA-1 >> MSRA+MSR-3;
MSRA+MSR-3 >> BASELINE-3;
BASELINE-3 >> BASELINE-1;
LILY-2 >> LILY-3; and
LILY-3 >> LILY-1.
Thus GA-nDCG and GA-Q detect more significant differences than the "@1" metrics, by observing the entire ranked lists.

As for GA-Hit@1, it failed to detect any significant differences and therefore is not very useful for system comparison.

We shall report on more significance test results in Section 4 based on alternative gold-standard data sets.

3.2 Evaluation with the Good Question Set

As was mentioned in Overview Part I [3], the four assessors independently assessed the quality of each *question* as well, since, generally speaking, not all questions posted to Yahoo! Chiebukuro are useful. We filtered out questions that did not receive an "AAAA" relevance pattern: that is, we devised a "good" question set where every question was rated A by all four assessors. As it turned out, we retained as many as 1,429 questions: only 71 questions were

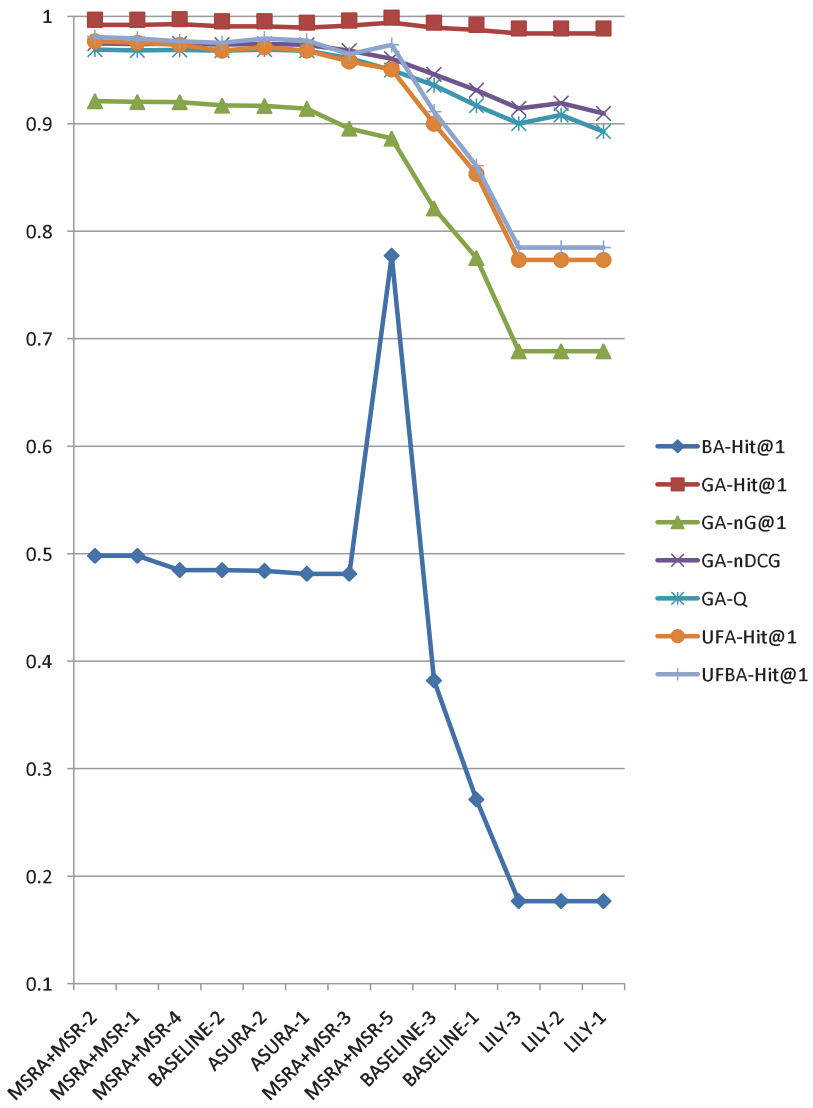

Figure 1: Mean performances with the full question set (1,500 questions). Runs are sorted by GA-nG@1.

filtered out. Hence, our full formal run question set is basically a sample of good questions (though we shall discuss an anomalous example in Section 3.4). This may be partly because the Yahoo! Chiebukuro site has a mechanism for automatic question filtering: questions that do not receive any answers for a period of seven days are automatically deleted [2].

Table 6 summarises the results of our evaluation with the good question set, in a way similar to Table 5. It can be observed that the results are very similar to those with the full question set. The ranks of MSRA+MSR-{1, 4} with GA-nCG@1 have swapped but they are basically ties according to both question sets. The significance test results with GA-nCG@1 for the adjacent run pairs are also identical. Since the effect of low-quality questions appears to be very small for our formal run question set, we shall use the full question set for further analysis.

3.3 Per-category Evaluation

As was mentioned in Overview Part I [3], our formal run question set covers 14 categories as defined by the Yahoo! Chiebukuro service. To examine whether systems' performances differ across categories, we also computed the performance means over each category.

Table 7 shows the performances averaged over 42 BUSINESS questions in a way similar to previous tables. It can be observed that all systems failed to outperform BASELINE-2 (sorting by length), and therefore that systems are not good at handling BUSINESS questions in our data set. It can also be observed that using graded relevance is useful: for example, while MSRA+MSR-3 and BASELINE-3 (sorting by freshness) are considered equal in terms of GA-Hit@1, the former is in fact significantly better than the latter in terms of GA-nG@1. That is, MSRA+MSR-3 is better at returning a *highly* relevant answer at rank 1 than BASELINE-3.

Tables 8 and 9 show the performances averaged over 36 MANNERS questions and 63 NEWS questions, respectively. Again, all systems failed to outperform BASELINE-2 for these question categories. Note also that GA-Hit@1 is too crude for the MANNERS

questions: in Table 8, GA-Hit@1 is 1 for *all* systems. Moreover, {UFA, UFBA}-Hit@1 are also too generous in Table 8: according to these metrics, all runs except BASELINE-3 and the LILY runs are perfect. Again, the GA-nG@1 column shows that using graded relevance is more informative.

The aforementioned three tables suggest that participatants may need to work harder on the BUSINESS, MANNERS and NEWS categories at least.

Tables 10-12 show the performances averaged over 120 EDUCATION questions, 154 HEALTH questions and 135 INTERNET questions. It can be observed that the general trends with these categories are similar to those with the full question set. Note that GA-Hit@1 fails to distinguish between runs again for the INTERNET questions in Table 12, despite the fact that statistically significant differences exist in terms of GA-nG@1: for example, BASELINE-3 (sorting by length) is significantly better than BASELINE-1 (random) at $\alpha = 0.01$.

Tables 13 and 14 show the performances averaged over 89 SCHOOL questions and 120 LOVE questions, respectively. For these two categories, it can be observed that MSRA+MSRA-4 does well: for the SCHOOL category, it is the top performer in terms of all metrics except BA-Hit@1; for the LOVE category, it is the top performer in terms of GA-{nG@1, nDCG, Q} and {UFA, UFBA}-Hit@1.

Tables 15 and 16 show the performances averaged over 38 CAREER questions and 136 LIFEGUIDE questions, respectively. For these two categories, it can be observed that GA-nG@1 agrees with BA-Hit@1, ranking the aforementioned anomaly MSRA+MSR-5 at the top. Also, the GA-{nDCG, Q} values suggest that the ASURA runs handle the task of ranking all answers (as opposed to returning exactly one answer) quite well.

Tables 17 and 18 show the performances averaged over 120 SPORTS questions and 167 ENTERTAINMENT questions, respectively. It can be observed that ASURA-2 handles the SPORTS category well: it is the top performer according to GA-{nG@1, nDCG, Q} and {UFA, UFBA}-Hit@1. (Again, all runs are equal for this category according to GA-Hit@1.) ASURA-2 also does well for ENTERTAINMENT: although MSRA+MSR-4 and MSRA+MSR-2 are the top performers according to GA-nG@1, ASURA-2 is the top performer according to GA-{Hit@1, nDCG, Q} and {UFA, UFBA}-Hit@1.

Finally, Tables 19 and 20 show the performances averaged over 58 TRAVEL questions and 222 YAHOO questions, respectively.

As we have seen, the binary relevance metrics {GA, UFA, UFBA}-Hit@1 are too crude for discussing the differences between systems. Hereafter, we shall focus on BA-Hit@1 and GA-{nCG@1, nDCG, Q}.

3.4 Question Hardness

In this section, we examine how each evaluation metric ranks *questions* in terms of hardness, where question hardness is defined in terms of the per-question performance averaged across runs. For averaging, we exclude MSRA+MSR-5 as this is an outlier as was explained in Section 3.1. Thus, for example, *average BA-Hit@1* for each question is computed using the 12 submitted runs.

Since we have 1,500 questions, for each metric, we define *easy* questions as those with the top 500 highest average performance values, and *hard* questions as those with the lowest 500 values. The remaining 500 questions are the *medium* questions.

Figures 2-5 show the distribution of easy/medium/hard questions for each question category for BA-Hit@1 and GA-{nG@1, nDCG, Q}. It can be observed that easy/hard questions according to BA are quite different from those according to GA. More specifically,

- According to BA-Hit@1, INTERNET, LIFEGUIDE, SPORTS and TRAVEL categories often contain easy questions, and LOVE often contains hard questions.
- According to GA-nG@1, LIFEGUIDE, LOVE and SCHOOL categories often contain easy questions, and ENTERTAINMENT, NEWS, TRAVEL and YAHOO often contain hard questions.
- According to GA-nDCG, LIFEGUIDE, LOVE and SCHOOL categories often contain easy questions, and NEWS, TRAVEL and YAHOO often contain hard questions. The results with GA-Q are similar.

Thus, "LOVE is hard according to BA, but easy according to GA; and TRAVEL is easy according to BA, but difficult according to GA". Why? We hypothesize that many LOVE questions are "social," eliciting many diverse answers expressing different views. The asker subjectively chooses only one of them as the BA, but some other answers are equally valid (i.e. highly relevant in terms of GA). Hence the BA is difficult for the system to identify. The first example shown in Figure 6 seems to support this hypothesis: this is a LOVE question, selected here because its average BA-Hit@1 is 0 while its average GA-nCG@1 is 1[3]. In this example, the asker is planning to propose to his girlfriend and is looking for the right words to say. The best answer says that nothing is better than the asker's own sincere words.

As for TRAVEL, we are not so sure: the other two examples in Figure 6 are TRAVEL questions, for which average BA-Hit@1 outperformed average GA-nG@1[4]. Question 6079055 is a completely silly question, asking "Fukushima Prefecture is in which prefecture?". (This question, for some reason, was rated AAAA by the four assessors.) And the best answer is equally silly: "Don't you know? Fukushima Prefecture is in Aichi Prefecture!". Whereas, Question 642255 may partially explain why "TRAVEL is easy according to BA." The asker is looking for a URL that contains information on a certain big wheel, and the answerer *copy and pastes the entire question* before providing the URL in his answer. If systems utilise the similarity between the question and the answer for answer selection, this best answer should be relatively easy to find. But further analysis on the discrepancies between different question hardness rankings is required.

Figures 7 and 8 visualise the correlation between the question ranking by average BA-Hit@1 and that by average GA-{nG@1, nDCG}. Note that since BA-Hit@1 is a binary flag and the average is taken over 12 runs, the possible values of average BA-Hit@1 are $0, 1/12, 2/12, \ldots, 1$. It can be observed that:

- Questions that are easy according to BA are also easy according to GA. (High average BA-Hit@1 values imply high average GA-{nG@1, nDCG} values.)
- Questions that are hard according to BA may not necessarily be hard according to GA. (Low average BA-Hit@1 values do not say anything about average GA-{nG@1, nDCG} values.)

The first observation suggests that if systems have found the BA for a question, they would also have found highly-relevant GA's successfully. The second observation suggests that systems may successfully find highly-relevant GA's that are not the BA.

[3]There were six such questions: 5565979, 5442340, 383311, 2939964, 2528961 and 1395988.

[4]Average BA-Hit@1 outperformed average GA-nG@1 for only three questions: 6079055, 642255 and 289580.

Table 7: Mean performances for the 42 BUSINESS questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
BASELINE-2	**0.5750**	**1**	**0.9524**	**0.9843**	**0.9794**	**1**	**1**
ASURA-2	0.5250	**1**	0.9444	0.9825	0.9779	**1**	**1**
ASURA-1	0.5250	**1**	0.9444	0.9827	0.9781	**1**	**1**
MSRA+MSR-4	0.5250	**1**	0.9365	0.9805	0.9771	0.9762	0.9762
MSRA+MSR-5	0.9250	**1**	0.9048	0.9663	0.9577	0.9286	0.9762
MSRA+MSR-2	**0.5750**	0.9762	0.9048	0.9691	0.9688	0.9524	0.9524
MSRA+MSR-1	0.5500	0.9762	0.8889	0.9666	0.9663	0.9524	0.9524
MSRA+MSR-3	0.5250	0.9762	0.8571$*$	0.9566	0.9530	0.9286	0.9286
BASELINE-3	0.3750	0.9762	0.7897	0.9377	0.9309	0.8810	0.8810
BASELINE-1	0.2500	0.9762	0.7857	0.9347	0.9276	0.8095	0.8333
LILY-3	0.3000	**1**	0.7381	0.9325	0.9233	0.8095	0.8333
LILY-2	0.3000	**1**	0.7381	0.9338	0.9254	0.8095	0.8333
LILY-1	0.3000	**1**	0.7381	0.9294	0.9173	0.8095	0.8333

Table 8: Mean performances for the 36 MANNERS questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, adjacent pairs of runs were tested using a two-sided sign test, but none of the differences were statistically significant. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
BASELINE-2	0.6000	**1**	**0.9352**	**0.9797**	0.9749	**1**	**1**
ASURA-1	0.4857	**1**	**0.9352**	0.9789	0.9739	**1**	**1**
ASURA-2	0.4857	**1**	0.9259	0.9788	**0.9752**	**1**	**1**
MSRA+MSR-4	0.5429	**1**	0.9167	0.9770	0.9730	**1**	**1**
MSRA+MSR-2	0.5143	**1**	0.9167	0.9761	0.9719	**1**	**1**
MSRA+MSR-1	0.5143	**1**	0.9167	0.9756	0.9709	**1**	**1**
MSRA+MSR-5	**0.9143**	**1**	0.8889	0.9668	0.9577	**1**	**1**
MSRA+MSR-3	0.4571	**1**	0.8611	0.9627	0.9543	**1**	**1**
BASELINE-3	0.1143	**1**	0.8056	0.9450	0.9317	0.9722	0.9722
LILY-3	0.1429	**1**	0.7963	0.9508	0.9386	0.9722	0.9722
LILY-2	0.1429	**1**	0.7963	0.9529	0.9423	0.9722	0.9722
LILY-1	0.1429	**1**	0.7963	0.9414	0.9235	0.9722	0.9722
BASELINE-1	0.1143	**1**	0.7963	0.9376	0.9195	**1**	**1**

Table 9: Mean performances for the 63 NEWS questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, adjacent pairs of runs were tested using a two-sided sign test, but none of the differences were statistically significant. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
BASELINE-2	0.4340	0.9841	**0.8942**	**0.9634**	**0.9573**	**0.9524**	0.9524
ASURA-2	0.3962	0.9841	0.8757	0.9556	0.9465	0.9206	0.9365
MSRA+MSR-2	0.4717	0.9841	0.8730	0.9541	0.9458	0.9365	0.9524
MSRA+MSR-1	0.4528	0.9841	0.8730	0.9540	0.9452	0.9365	0.9524
MSRA+MSR-4	0.4151	0.9841	0.8704	0.9571	0.9496	**0.9524**	0.9524
ASURA-1	0.4340	0.9841	0.8677	0.9555	0.9457	0.9048	0.9365
MSRA+MSR-3	0.5283	0.9683	0.8492	0.9494	0.9396	0.8889	0.9206
MSRA+MSR-5	**0.8302**	**1**	0.8280	0.9394	0.9282	**0.9524**	**0.9841**
BASELINE-3	0.2830	0.9841	0.7593	0.9193	0.9080	0.9048	0.9206
BASELINE-1	0.2642	0.9524	0.6825	0.9028	0.8917	0.8095	0.8095
LILY-3	0.1132	0.9365	0.6164	0.8908	0.8842	0.7302	0.7302
LILY-2	0.1132	0.9365	0.6164	0.8973	0.8927	0.7302	0.7302
LILY-1	0.1132	0.9365	0.6164	0.8866	0.8772	0.7302	0.7302

Table 10: Mean performances for the 120 EDUCATION questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-2	0.5603	**1**	**0.9278**	**0.9764**	**0.9700**	**0.9750**	**0.9833**
MSRA+MSR-1	0.5603	**1**	0.9250	0.9753	0.9689	**0.9750**	**0.9833**
MSRA+MSR-4	0.5259	**1**	0.9222	0.9743	0.9670	**0.9750**	0.9750
ASURA-1	0.5431	**1**	0.9222	0.9742	0.9684	0.9583	0.9667
ASURA-2	0.5259	**1**	0.9167	0.9744	0.9687	0.9500	0.9583
BASELINE-2	0.5517	**1**	0.9111	0.9719	0.9661	0.9417	0.9500
MSRA+MSR-3	0.5862	**1**	0.8958	0.9684	0.9612	0.9500	0.9583
MSRA+MSR-5	**0.7845**	**1**	0.8833$*$	0.9616	0.9529	0.9250	0.9417
BASELINE-3	0.4569	0.9917	0.8181	0.9452	0.9364	0.8583	0.8750
BASELINE-1	0.3103	0.9917	0.7889$**$	0.9335	0.9214	0.8167	0.8167
LILY-3	0.2069	0.9917	0.6361	0.8996	0.8850	0.6583	0.6833
LILY-2	0.2069	0.9917	0.6361	0.9059	0.8954	0.6583	0.6833
LILY-1	0.2069	0.9917	0.6361	0.8935	0.8751	0.6583	0.6833

Table 11: Mean performances for the 154 HEALTH questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-2	0.5400	**1**	**0.9491**	**0.9841**	**0.9787**	**0.9870**	0.9870
MSRA+MSR-1	0.5400	**1**	**0.9491**	0.9839	0.9785	**0.9870**	0.9870
MSRA+MSR-4	0.5133	**1**	0.9426	0.9830	0.9781	0.9740	0.9740
BASELINE-2	0.4933	0.9935	0.9275	0.9799	0.9769	0.9675	0.9675
ASURA-2	0.4933	0.9935	0.9275	0.9792	0.9757	0.9740	0.9740
MSRA+MSR-3	0.5333	**1**	0.9264	0.9771	0.9697	0.9805	0.9870
ASURA-1	0.5067	0.9935	0.9253	0.9784	0.9747	0.9675	0.9675
MSRA+MSR-5	**0.8400**	**1**	0.9242$**$	0.9738	0.9650	0.9740	**0.9935**
BASELINE-3	0.4600	**1**	0.8626	0.9608	0.9533	0.9351	0.9481
BASELINE-1	0.3067	**1**	0.8247$**$	0.9464	0.9331	0.9026	0.9091
LILY-3	0.0933	0.9935	0.6818	0.9164	0.9004	0.8117	0.8182
LILY-2	0.0933	0.9935	0.6818	0.9202	0.9069	0.8117	0.8182
LILY-1	0.0933	0.9935	0.6818	0.9114	0.8922	0.8117	0.8182

Table 12: Mean performances for the 135 INTERNET questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-2	0.5940	**1**	**0.9370**	**0.9796**	**0.9745**	0.9778	0.9852
MSRA+MSR-1	0.6015	**1**	0.9346	0.9785	0.9730	0.9778	0.9852
MSRA+MSR-4	0.5865	**1**	0.9296	0.9783	0.9730	0.9778	0.9852
BASELINE-2	0.5865	**1**	0.9259	0.9761	0.9692	0.9778	0.9852
ASURA-2	0.5789	**1**	0.9259	0.9784	0.9726	0.9852	**0.9926**
ASURA-1	0.5564	**1**	0.9259	0.9776	0.9713	0.9852	**0.9926**
MSRA+MSR-3	0.6165	**1**	0.9247	0.9756	0.9680	**0.9926**	**0.9926**
MSRA+MSR-5	**0.8571**	**1**	0.9025$*$	0.9664	0.9556	0.9556	0.9778
BASELINE-3	0.4511	**1**	0.8482$**$	0.9539	0.9422	0.9037	0.9111
BASELINE-1	0.3684	**1**	0.7432$*$	0.9270	0.9112	0.7852	0.7926
LILY-3	0.2406	**1**	0.6753	0.9104	0.8919	0.7185	0.7259
LILY-2	0.2406	**1**	0.6753	0.9122	0.8950	0.7185	0.7259
LILY-1	0.2406	**1**	0.6753	0.9064	0.8853	0.7185	0.7259

Table 13: Mean performances for the 89 SCHOOL questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-4	0.5349	**1**	**0.9813**	**0.9896**	**0.9843**	**1**	**1**
BASELINE-2	0.4884	**1**	0.9775	0.9890	0.9837	**1**	**1**
ASURA-2	0.4767	**1**	0.9738	0.9882	0.9827	0.9888	0.9888
ASURA-1	0.4767	**1**	0.9663	0.9871	0.9815	0.9888	0.9888
MSRA+MSR-2	0.5000	**1**	0.9625	0.9862	0.9802	0.9888	0.9888
MSRA+MSR-1	0.5000	**1**	0.9625	0.9848	0.9779	0.9888	0.9888
MSRA+MSR-3	0.4651	**1**	0.9532	0.9841	0.9785	**1**	**1**
MSRA+MSR-5	**0.8605**	**1**	0.9513	0.9790	0.9690	0.9888	**1**
BASELINE-3	0.3721	0.9775	0.8970	0.9657	0.9540	0.9438	0.9438
BASELINE-1	0.3023	**1**	0.8539$**$	0.9535	0.9367	0.9101	0.9101
LILY-3	0.0930	0.9888	0.7584	0.9364	0.9181	0.8539	0.8539
LILY-2	0.0930	0.9888	0.7584	0.9386	0.9221	0.8539	0.8539
LILY-1	0.0930	0.9888	0.7584	0.9340	0.9137	0.8539	0.8539

Table 14: Mean performances for the 120 LOVE questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-4	0.3243	0.9917	**0.9403**	**0.9791**	**0.9703**	**1**	**1**
MSRA+MSR-2	0.3514	0.9917	**0.9403**	0.9785	0.9693	**1**	**1**
MSRA+MSR-1	0.3514	0.9917	**0.9403**	0.9780	0.9682	**1**	**1**
MSRA+MSR-3	0.3243	**1**	0.9333	0.9761	0.9658	0.9750	0.9750
BASELINE-2	0.3874	0.9833	0.9208	0.9741	0.9673	0.9833	0.9917
ASURA-2	0.3874	0.9833	0.9208	0.9735	0.9670	0.9833	0.9917
ASURA-1	0.3784	0.9833	0.9181	0.9730	0.9662	0.9833	0.9917
MSRA+MSR-5	**0.8108**	0.9917	0.9167$**$	0.9641	0.9490	0.9750	0.9833
BASELINE-3	0.2703	**1**	0.8681	0.9572	0.9417	0.9667	0.9750
BASELINE-1	0.1441	0.9917	0.8139$**$	0.9432	0.9262	0.9500	0.9583
LILY-3	0.0721	0.9833	0.7445	0.9316	0.9135	0.9333	0.9417
LILY-2	0.0721	0.9833	0.7445	0.9390	0.9257	0.9333	0.9417
LILY-1	0.0721	0.9833	0.7445	0.9286	0.9087	0.9333	0.9417

Table 15: Mean performances for the 38 CAREER questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-5	**0.9143**	**1**	**0.9386**	0.9733	0.9619	**1**	**1**
MSRA+MSR-3	0.5429	**1**	0.9211	0.9755	0.9686	0.9737	0.9737
BASELINE-2	0.4857	**1**	0.9211	0.9767	0.9707	**1**	**1**
ASURA-1	0.4857	**1**	0.9211	**0.9778**	**0.9730**	**1**	**1**
MSRA+MSR-4	0.5429	**1**	0.9123	0.9755	0.9696	0.9737	0.9737
MSRA+MSR-2	0.6000	**1**	0.9123	0.9755	0.9695	0.9737	0.9737
MSRA+MSR-1	0.6000	**1**	0.9123	0.9760	0.9701	0.9737	0.9737
ASURA-2	0.5143	**1**	0.9123$*$	0.9767	0.9720	**1**	**1**
BASELINE-3	0.3714	**1**	0.8290	0.9487	0.9365	0.8947	0.8947
BASELINE-1	0.1429	**1**	0.7456	0.9245	0.9073	0.7895	0.7895
LILY-3	0.1429	0.9737	0.6842	0.9205	0.9080	0.7895	0.7895
LILY-2	0.1429	0.9737	0.6842	0.9203	0.9079	0.7895	0.7895
LILY-1	0.1429	0.9737	0.6842	0.9066	0.8864	0.7895	0.7895

Table 16: Mean performances for the 136 LIFEGUIDE questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-5	**0.7794**	**1**	**0.9436**	0.9753	0.9650	**0.9853**	**0.9926**
MSRA+MSR-4	0.4853	**1**	0.9351	0.9778	0.9711	0.9706	0.9706
MSRA+MSR-1	0.5074	**1**	0.9326	0.9773	0.9700	0.9779	0.9779
BASELINE-2	0.4706	**1**	0.9326	0.9776	0.9704	0.9779	0.9779
MSRA+MSR-2	0.4926	**1**	0.9302	0.9770	0.9698	0.9779	0.9779
ASURA-2	0.5221	**1**	0.9277	**0.9781**	**0.9721**	0.9779	0.9779
ASURA-1	0.5000	**1**	0.9277	0.9773	0.9708	0.9779	0.9779
MSRA+MSR-3	0.4853	**1**	0.9179$**$	0.9744	0.9666	**0.9853**	0.9853
BASELINE-3	0.4118	**1**	0.8493	0.9578	0.9472	0.9338	0.9412
BASELINE-1	0.3162	0.9926	0.8382$**$	0.9507	0.9368	0.8897	0.8971
LILY-3	0.2132	0.9853	0.7316	0.9271	0.9122	0.7647	0.7721
LILY-2	0.2132	0.9853	0.7316	0.9306	0.9183	0.7647	0.7721
LILY-1	0.2132	0.9853	0.7316	0.9233	0.9060	0.7647	0.7721

Table 17: Mean performances for the 120 SPORTS questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
ASURA-2	0.5370	**1**	**0.9500**	**0.9825**	**0.9777**	**0.9917**	**0.9917**
ASURA-1	0.5463	**1**	0.9472	0.9810	0.9755	**0.9917**	**0.9917**
BASELINE-2	0.5648	**1**	0.9403	0.9777	0.9718	0.9750	0.9750
MSRA+MSR-4	0.5556	**1**	0.9361	0.9793	0.9748	0.9833	0.9833
MSRA+MSR-2	0.5556	**1**	0.9319	0.9780	0.9734	0.9833	0.9833
MSRA+MSR-1	0.5463	**1**	0.9319	0.9780	0.9732	0.9833	0.9833
MSRA+MSR-5	**0.8241**	**1**	0.8958	0.9618	0.9509	0.9667	0.9750
MSRA+MSR-3	0.5000	**1**	0.8903$*$	0.9683	0.9628	0.9417	0.9417
BASELINE-3	0.3981	**1**	0.8403$*$	0.9511	0.9432	0.9000	0.9000
BASELINE-1	0.2593	**1**	0.7847$**$	0.9295	0.9135	0.8667	0.8750
LILY-3	0.2315	**1**	0.6903	0.9136	0.8989	0.7667	0.7750
LILY-2	0.2315	**1**	0.6903	0.9189	0.9075	0.7667	0.7750
LILY-1	0.2315	**1**	0.6903	0.9076	0.8897	0.7667	0.7750

Table 18: Mean performances for the 167 ENTERTAINMENT questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-4	0.5226	0.9940	**0.9172**	0.9708	0.9645	0.9701	0.9760
MSRA+MSR-2	0.5419	0.9940	**0.9172**	0.9721	0.9665	0.9641	0.9701
ASURA-2	0.5419	**1**	0.9132	**0.9732**	**0.9676**	**0.9820**	**0.9880**
MSRA+MSR-1	0.5355	0.9940	0.9122	0.9701	0.9642	0.9641	0.9701
BASELINE-2	0.5548	**1**	0.9082	0.9710	0.9655	0.9581	0.9701
ASURA-1	0.5419	0.9940	0.8972	0.9704	0.9653	0.9641	0.9701
MSRA+MSR-5	**0.6710**	0.9940	0.8603	0.9536	0.9442	0.9401	0.9701
MSRA+MSR-3	0.4774	0.9940	0.8583$**$	0.9590	0.9519	0.9341	0.9401
BASELINE-3	0.4065	0.9880	0.7864	0.9345	0.9226	0.8802	0.8922
BASELINE-1	0.3097	0.9940	0.7465$*$	0.9246	0.9111	0.8443	0.8443
LILY-3	0.2452	0.9940	0.6816	0.9108	0.8986	0.7725	0.7784
LILY-2	0.2452	0.9940	0.6816	0.9146	0.9041	0.7725	0.7784
LILY-1	0.2452	0.9940	0.6816	0.9077	0.8931	0.7725	0.7784

Table 19: Mean performances for the 58 TRAVEL questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-3	0.5862	**1**	**0.9080**	0.9681	0.9610	0.9483	0.9828
BASELINE-2	0.6034	0.9828	0.9052	0.9678	0.9624	0.9655	0.9828
MSRA+MSR-1	0.6207	**1**	0.9023	**0.9736**	0.9686	**0.9828**	1
MSRA+MSR-2	0.6034	**1**	0.8966	**0.9736**	**0.9692**	**0.9828**	1
ASURA-2	0.6379	0.9828	0.8908	0.9658	0.9587	0.9483	0.9655
MSRA+MSR-4	0.6207	**1**	0.8879	0.9655	0.9590	0.9655	1
MSRA+MSR-5	**0.8448**	**1**	0.8764	0.9551	0.9439	0.9483	1
ASURA-1	0.6207	0.9655	0.8764$*$	0.9634	0.9573	0.9310	0.9655
BASELINE-3	0.5690	0.9828	0.8046$*$	0.9457	0.9383	0.8793	0.9138
BASELINE-1	0.2759	0.9310	0.6925$**$	0.9047	0.8905	0.7759	0.7931
LILY-3	0.0862	0.9310	0.5948	0.8814	0.8685	0.6897	0.6897
LILY-2	0.0862	0.9310	0.5948	0.8875	0.8800	0.6897	0.6897
LILY-1	0.0862	0.9310	0.5948	0.8767	0.8611	0.6897	0.6897

Table 20: Mean performances for the 222 YAHOO questions. Runs are sorted by GA-nG@1. For the GA-nG@1 column, "$**$" indicates that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.01$). Note that statistical significance is not transitive. The highest performance in each column is shown in bold.

RunID	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q	UFA-Hit@1	UFBA-Hit@1
MSRA+MSR-1	0.3521	0.9955	**0.8799**	**0.9635**	**0.9590**	0.9595	0.9640
MSRA+MSR-2	0.3521	0.9955	0.8784	**0.9635**	**0.9590**	**0.9685**	0.9730
ASURA-1	0.3521	0.9910	0.8701	0.9616	0.9581	0.9414	0.9685
MSRA+MSR-4	0.3662	0.9955	0.8686	0.9617	0.9579	0.9459	0.9505
ASURA-2	0.3662	0.9910	0.8679	0.9614	0.9578	0.9414	0.9685
BASELINE-2	0.3333	0.9910	0.8649	0.9599	0.9554	0.9414	0.9640
MSRA+MSR-3	0.3333	0.9910	0.8484$**$	0.9548	0.9505	0.9279	0.9414
MSRA+MSR-5	**0.5822**	**1**	0.7951	0.9353	0.9245	0.8829	0.9279
BASELINE-3	0.2911	0.9955	0.7508	0.9251	0.9172	0.8333	0.8559
BASELINE-1	0.2160	**1**	0.7252$**$	0.9143	0.9012	0.8108	0.8288
LILY-3	0.2019	**1**	0.6667	0.9041	0.8946	0.7252	0.7568
LILY-2	0.2019	**1**	0.6667	0.9134	0.9089	0.7252	0.7568
LILY-1	0.2019	**1**	0.6667	0.8994	0.8871	0.7252	0.7568

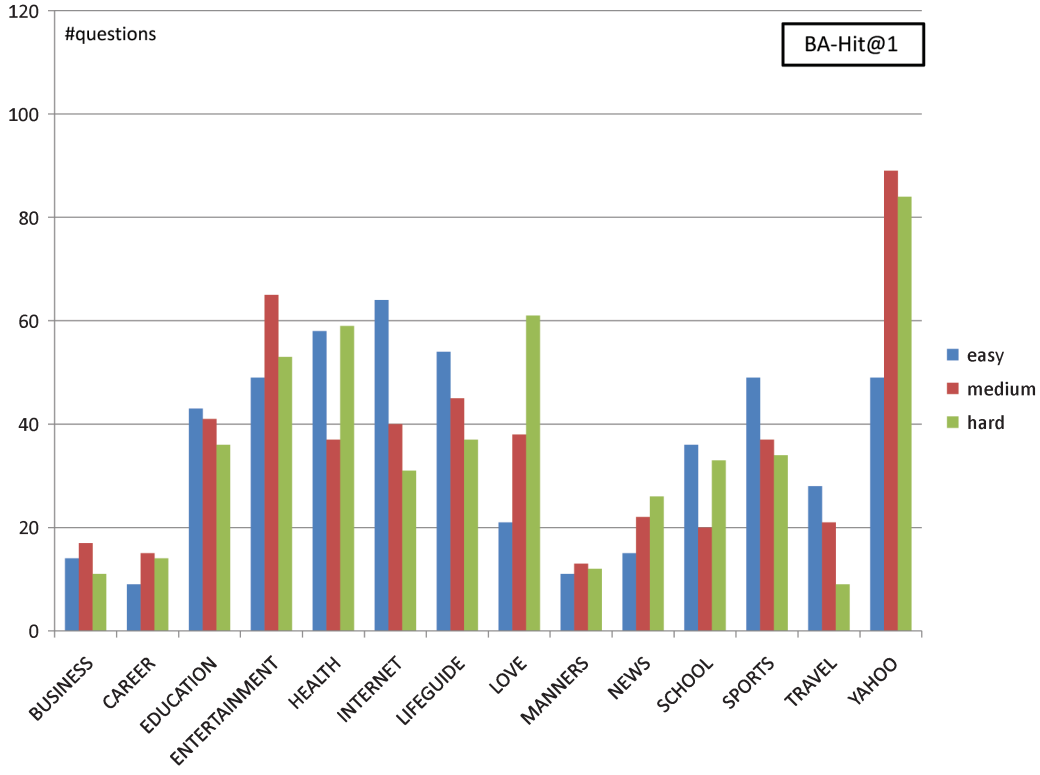

Figure 2: Distribution of easy/medium/hard questions according to BA-Hit@1.

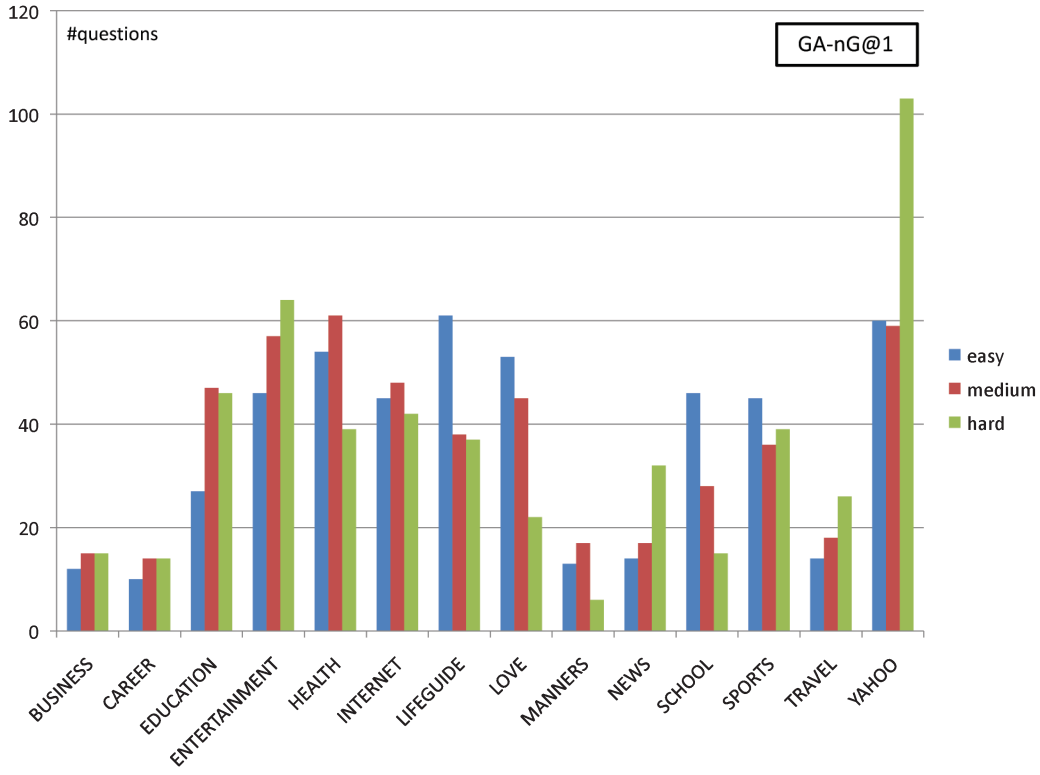

Figure 3: Distribution of easy/medium/hard questions according to GA-nG@1.

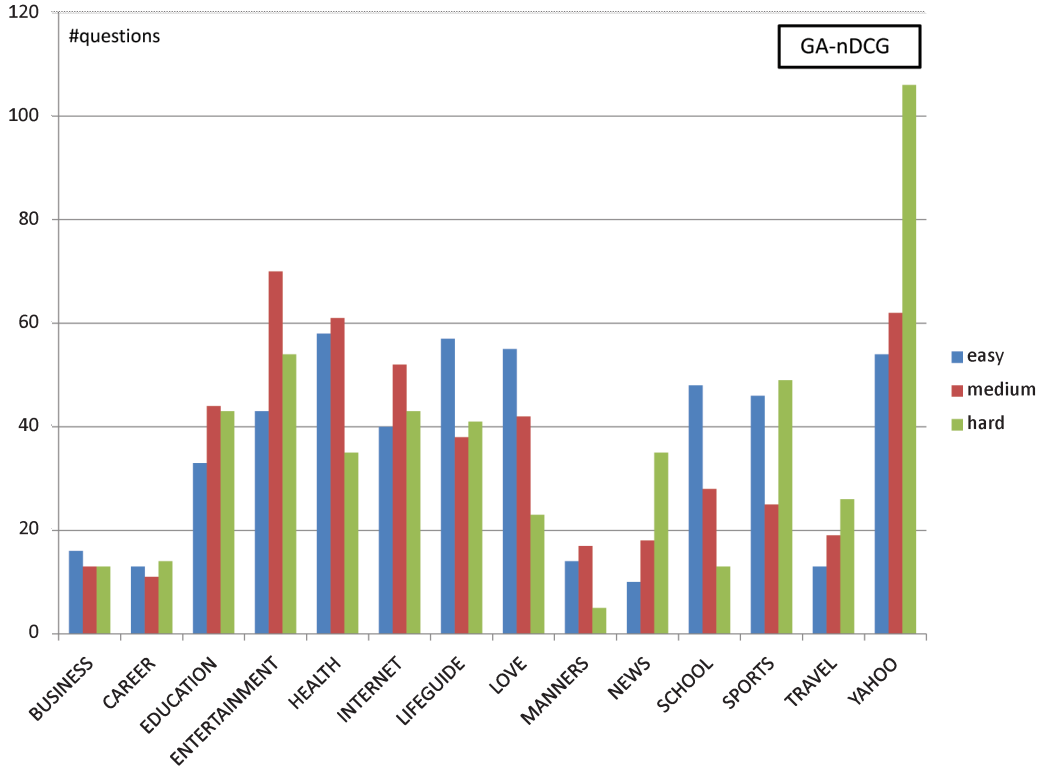

Figure 4: Distribution of easy/medium/hard questions according to GA-nDCG.

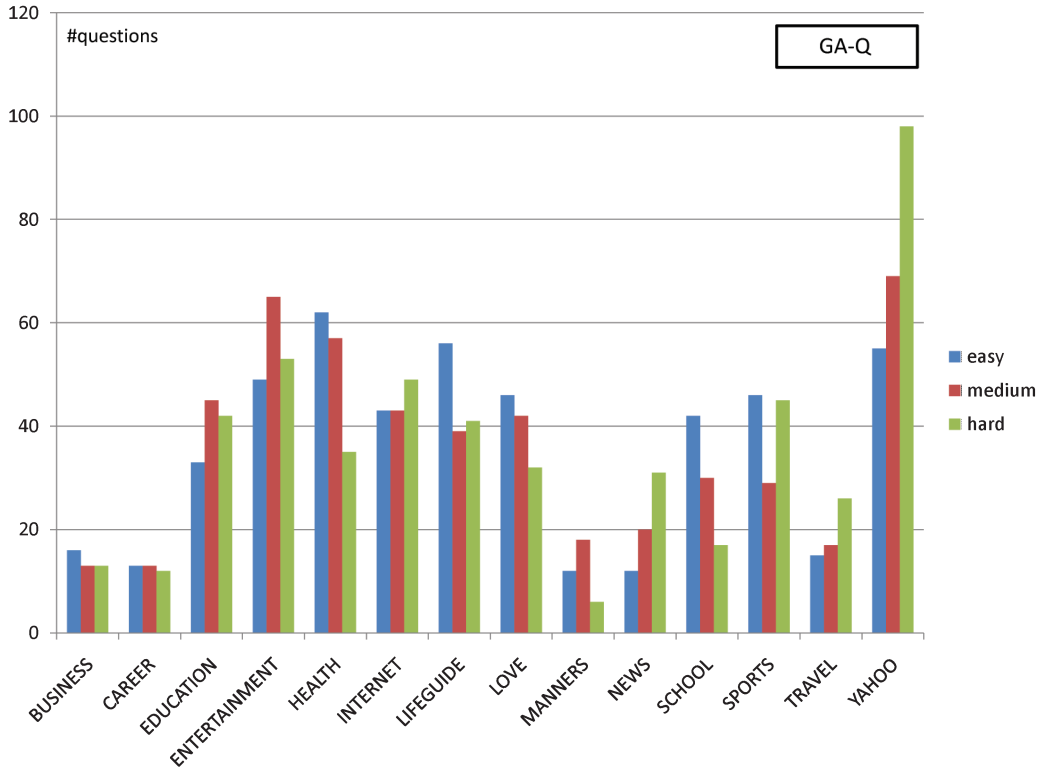

Figure 5: Distribution of easy/medium/hard questions according to GA-Q.

<Q_ID> 5565979 </Q_ID>
<TOPCATEGORY_LABEL> love </TOPCATEGORY_LABEL>
<NUM_ANSWERS> 9 </NUM_ANSWERS>
<QUESTION_TEXT>
明日プロポーズをします！でも言葉に悩んでます！正直にストレートに言うつもりです！
でも言葉が出てきません！
</QUESTION_TEXT>

BEST ANSWER:
<A_ID> 20176136 </A_ID>
<USER_ID> 351653 </USER_ID>
<ANSWER_TEXT>
そんな大事な言葉は、他人の言葉より、心からの自分の言葉で勝負しろ！ 整った言葉じゃなくたっていいんだ。 お前の言葉で言ったらいい。 今日一日、一生懸命考えてみろ。
そんな一日を過ごすのも幸せな事だぞ！
</ANSWER_TEXT>

<Q_ID> 6079055 </Q_ID>
<TOPCATEGORY_LABEL> travel </TOPCATEGORY_LABEL>
<NUM_ANSWERS> 5 </NUM_ANSWERS>
<QUESTION_TEXT>
地理にウトイ私です。福島県ってなに県にあるん ですか、オネガイ‥ 名古屋市・１９才で～す
</QUESTION_TEXT>

BEST ANSWER:
<A_ID> 22024934 </A_ID>
<ANSWER_TEXT>
名古屋市在住なのに知らないんですか!? 福島県は愛知県北東部にあるんですよ！
自分の住んでる県の地理ぐらいは勉強しておかないと…恥かきますよ！
</ANSWER_TEXT>

<Q_ID> 642255 </Q_ID>
<TOPCATEGORY_LABEL> travel </TOPCATEGORY_LABEL>
<NUM_ANSWERS> 6 </NUM_ANSWERS>
<QUESTION_TEXT>
名古屋の方に質問です。栄のど真ん中に観覧車が作りかかってるんです！ いつできるんですか？ その他この観覧車について、詳しいこと教えて下さい。
参考ＵＲＬあれば教えて下さい！
</QUESTION_TEXT>

BEST ANSWER:
<A_ID> 3458905 </A_ID>
<ANSWER_TEXT>
名古屋の方に質問です。栄のど真ん中に観覧車が作りかかってるんです！ いつできるんですか？ その他この観覧車について、詳しいこと教えて下さい。
参考ＵＲＬあれば教えて下さい！
回答 http://www.chunichi.co.jp/kodomo/all/kdm040421T1335.html
</ANSWER_TEXT>

Figure 6: Examples of LOVE and TRAVEL questions with their best answers.

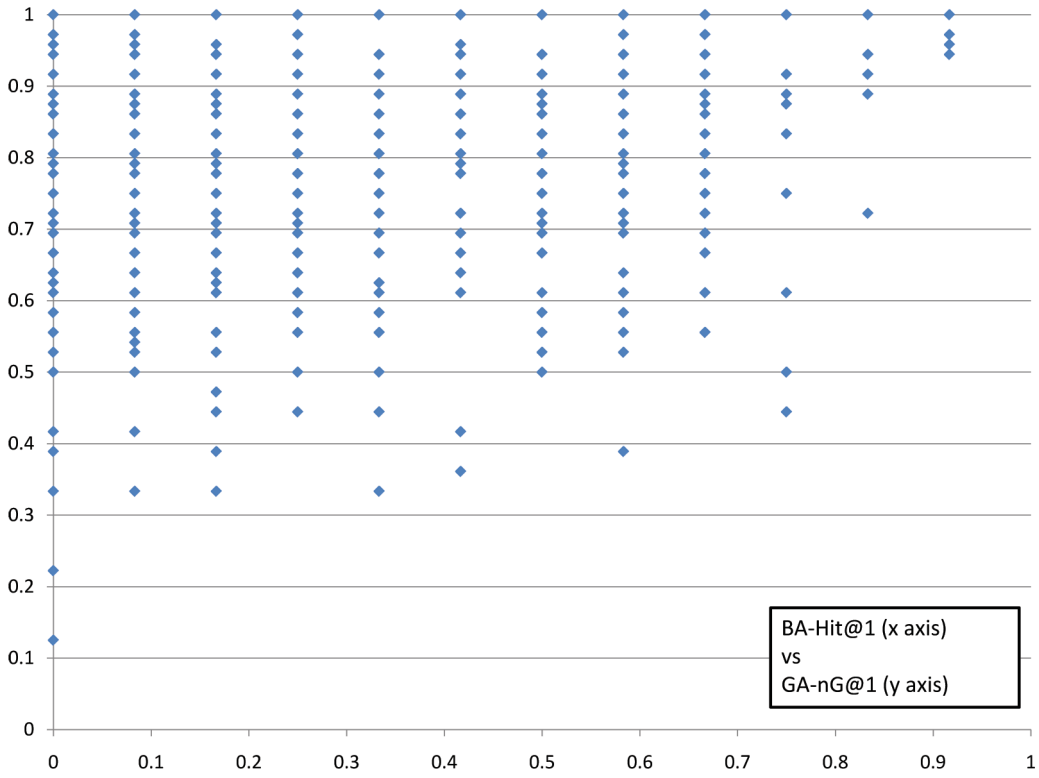

Figure 7: Question ranking correlation between BA-Hit@1 and GA-nG@1.

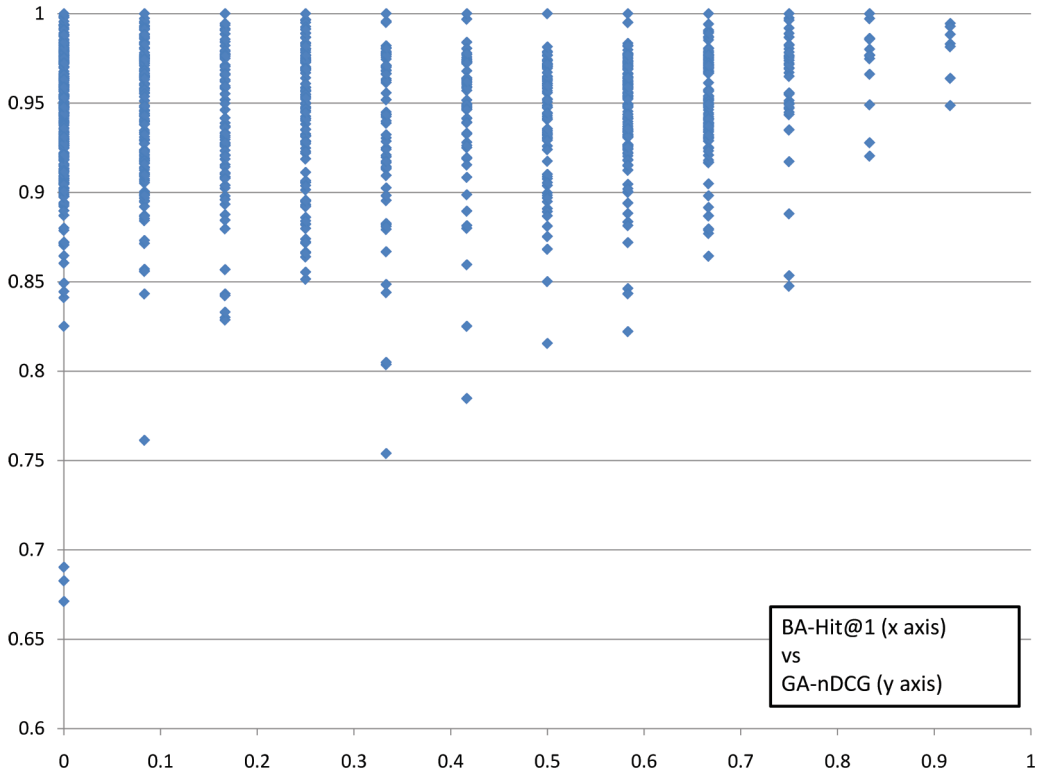

Figure 8: Question ranking correlation between BA-Hit@1 and GA-nDCG.

Figures 9 and 10 visualise the correlation between the question ranking by average GA-nG@1 and that by average GA-{nDCG, Q}. Figure 11 shows a similar graph for average GA-nDCG and average GA-Q. It can be observed that:

- The average GA-nDCG and GA-Q rankings are highly correlated with the average GA-nG@1 ranking, but the former shows a higher correlation. (Kendall's rank correlation: 0.810 vs 0.727.)
- The average GA-nDCG and GA-Q rankings are highly correlated with each other. (Kendall's rank correlation: 0.876.)

In short, these graded-relevance GA metrics agree reasonably well on which questions are hard and which are not.

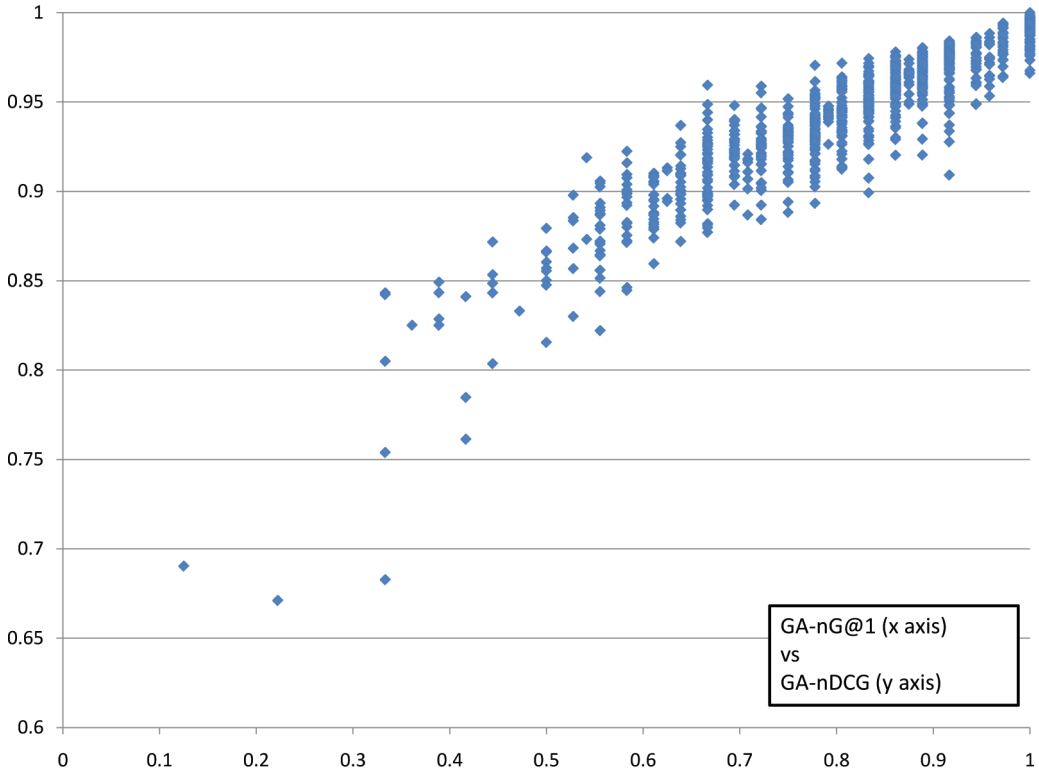

Figure 9: Question ranking correlation between GA-nG@1 and GA-nDCG.

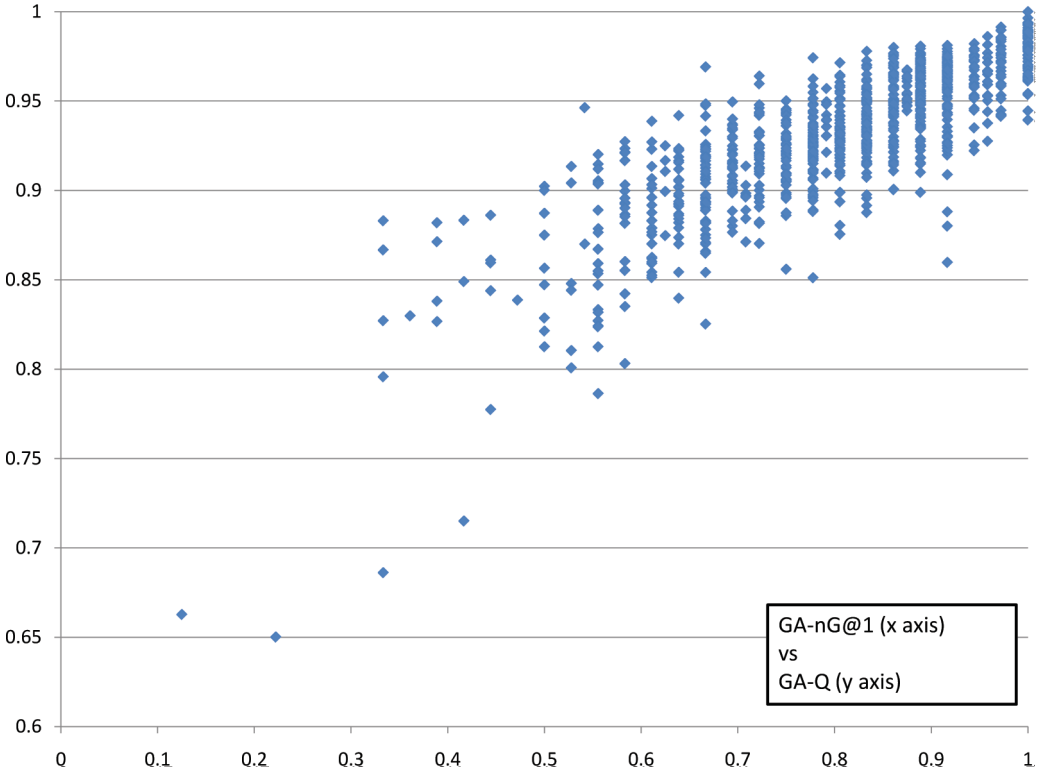

Figure 10: Question ranking correlation between GA-nG@1 and GA-Q.

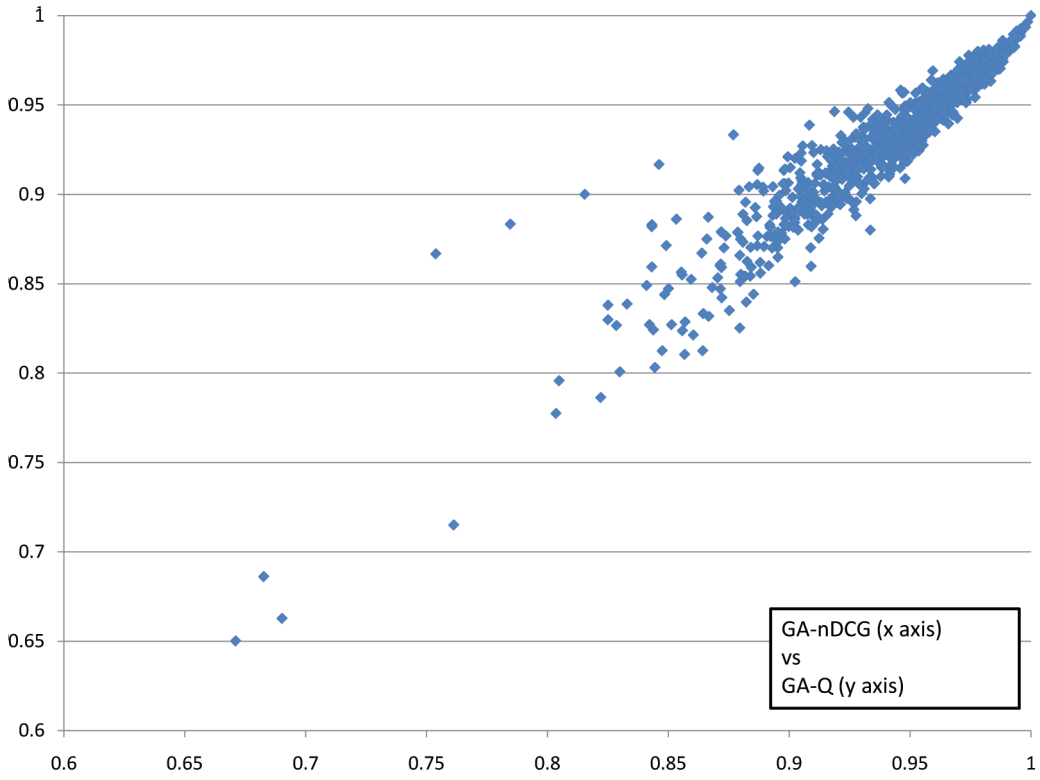

Figure 11: Question ranking correlation between GA-nDCG and GA-Q.

4. NEW PYRAMID EXPERIMENTS

For computing our official results based on GA, we relied on Table 2. However, how to map each relevance pattern to a relevance level was decided in a rather *ad hoc* manner. (Sakai [12] tried one variant of this table.) After releasing the official results to participants, we devised a very simple, more systematic method to define the mapping.

As was described earlier, each answer received four labels, each either A, B or C. Now, let us define a *judgment weight* for each label: we give 2, 1, and 0 in this paper, respectively. Then we can compute a weight for each relevance pattern by summing the judgments weights, for example, $4 * 2 = 8$ for "AAAA."

Table 21 shows a new mapping table based on the judgment weights. Note that we have increased the number of relevance levels from 4 to 9 (including $L0$)[5]. We refer to this new gold-standard data as "GAW" (Good Answers based on judgment Weights). Hence the corresponding metrics we examine are called GAW-{nC@1, nDCG, Q}. In this experiment, gain values of 1-8 were assigned to relevance levels $L1$-$L8$, respectively.

We also included *human performances* in our new experiment. Judges J1, J2, J3 and J4 were each regard as a system, which returns all answers rated A first, then all rated B, and then finally all rated C for each question. In addition, the BA data was also regarded as a system which always returns the BA for each question and nothing else. When treated as a system, they will be denoted by BA, J1 and so on. Note that evaluating BA with nDCG and Q is not very fair, as it suffers heavily from poor recall. However, we include these results for completeness.

Table 22 summarises the results of our new experiments based on the GAW data. We compared the system rankings (excluding the human performances) with the GA-based rankings derived from Table 5, and the rank changes are indicated by "GA↑1" and so on. For example, when the ranking based on GAW-nCG@1 is compared to that based on GA-nCG@1, ASURA-2 moved up one rank, while BASELINE-2 moved down one rank. Recall that BA "returns" only one answer per question so its GAW-{nDCG, Q} values are naturally extremely low. It can be observed from the table that:

(i) The GAW-based rankings are generally similar to the GA-based ones. There are occasional rank swaps, but the performance differences for these cases are generally very small.

(ii) The human performances vary very widely. For example, while J3 is on average the top performer in terms of all three metrics, J4 and J1 underperform even the length-based BASELINE-2. (These two judges significantly underperform BASELINE-2 at $\alpha = 0.01$ for all three metrics.) This is quite surprising, given that GAW was constructed based on the assessments of all four judges.

(iii) The system runs MSRA+MSRA-{2, 1, 4} and ASURA-{2, 1} lie between the above two extreme cases of human performances. In this respect, they are "doing as well as humans."

(iv) J2 is effective in terms of GAW-nG@1, but not necessarily in terms of GAW-{nDCG, Q}. For example, in terms of GAW-nG@1, J2 outperforms BASELINE-2 for 324 questions, and underforms it for 277 questions (although this is statistically not significant); whereas, in terms of GAW-{nDCG, Q}, J2 significantly underperforms BASELINE-2 at $\alpha = 0.01$ (although mean GAW-nDCG ranks J2 above BASELINE-2 in the table).

[5]The aforementioned version of the evaluation tool that we released does not support nine relevance levels. We plan to release the latest version in the near future.

Table 21: Mapping relevance patterns to relevance levels based on judgment weights (A:B=2:1).

(a) pattern (weight)	(b) #answers	(c) level	(d) #answers
AAAA (8)	1301	$L8$	1301
AAAB (7)	1505	$L7$	1505
AABB (6)	1525	$L6$	1527
AAA (6)	2		
ABBB (5)	1385	$L5$	1399
AAB (5)	14		
BBBB (4)	1241	$L4$	1318
ABB (4)	76		
AA (4)	1		
BBB (3)	231	$L3$	238
AB (3)	7		
BB (2)	105	$L2$	106
A (2)	1		
B (1)	32	$L1$	32
(C's only)	17	$L0$	17
total	7443	total	7443

Table 23: Mean performances with the full question set based on the BA data (with J1-J4). "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	BA-Hit@1
BA	1.0000$**$
MSRA+MSR-5	0.7773$**$
MSRA+MSR-2	0.4980
MSRA+MSR-1	0.4980
MSRA+MSR-4	0.4847
BASELINE-2	0.4847
ASURA-2	0.4840
MSRA+MSR-3	0.4813
ASURA-1	0.4813†
J3	0.4353
J2	0.4187
BASELINE-3	0.3820
J1	0.3280$*$
J4	0.3020
BASELINE-1	0.2713$**$
LILY-3	0.1767
LILY-2	0.1767
LILY-1	0.1767

For completeness, we present the BA-Hit@1 values of all runs (including human performances) in Table 23. Note that, unlike Table 5, this new table shows runs sorted by BA-Hit@1. Similar to the GAW-based results, J3 and J2 seem to do much better than J1 and J4 in terms of returning the BA. Moreover, it can be observed that many systems outperform humans in terms of BA.

Figure 12 compares the GAW-{nDCG, Q} and BA-Hit@1 rankings with the GAW-nG@1 ranking. The runs on the vertical axis have been sorted by mean GAW-nG@1. It is clear that the four judges do not do well when evaluated with BA. (The performances of BA (as a system output) are omitted as their GAW-{nDCG, Q} values are extremely low and would hurt the clarity of the graphs.)

We have shown that the submitted runs MSRA+MSRA-{2, 1, 4} and ASURA-{2, 1} do as well as humans based on the GAW data. However, we may be overestimating the human performances,

Table 22: Mean performances with the full question set based on the GAW data. For each metric, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	GAW-nG@1		GAW-nDCG		GAW-Q
J3	0.9567	J3	0.9857	J3	0.9760
J2	0.9446	MSRA+MSR-2	0.9797	MSRA+MSR-2	0.9688
MSRA+MSR-2	0.9288	MSRA+MSR-4	0.9795**	MSRA+MSR-4 (GA↑1)	0.9687
MSRA+MSR-1	0.9278	J2	0.9794‡	ASURA-2 (GA↓1)	0.9683*
MSRA+MSR-4	0.9276	MSRA+MSR-1 (GA↑1)	0.9791†	MSRA+MSR-1	0.9679
ASURA-2 (GA↑1)	0.9251	ASURA-2 (GA↓1)	0.9790**	ASURA-1 (GA↑1)	0.9674
BASELINE-2 (GA↓1)	0.9242	ASURA-1 (GA↑1)	0.9785	BASELINE-2 (GA↓1)	0.9673**
ASURA-1	0.9238**	BASELINE-2 (GA↓1)	0.9784**	J2	0.9646
MSRA+MSR-3	0.9076	MSRA+MSR-3	0.9744**	MSRA+MSR-3	0.9609**
BA	0.9038	J1	0.9724	J1	0.9573
MSRA+MSR-5	0.8984**	J4	0.9699†	J4	0.9538
J1	0.8916	MSRA+MSR-5	0.9686**	MSRA+MSR-5	0.9505**
J4	0.8814*	BASELINE-3	0.9576**	BASELINE-3	0.9366**
BASELINE-3	0.8460**	BASELINE-1	0.9455**	BASELINE-1	0.9172
BASELINE-1	0.8057**	LILY-2	0.9365**	LILY-2	0.9094**
LILY-3	0.7354	LILY-3	0.9325**	LILY-3	0.9015**
LILY-2	0.7354	LILY-1	0.9291**	LILY-1	0.8944**
LILY-1	0.7354	BA	0.4318	BA	0.2518

as the GAW data aggregates the assessments of all four judges. What if a judge is evaluated based on a gold-standard data to which he/she did *not* contribute? To answer this question, we created four new gold-standard data, by ignoring the assessments of one judge at a time. Thus, we used a table similar to Table 21, but with relevance patterns "AAA", "AAB", "ABB" etc., and relevance levels $L0$ through $L6$. Gain values of 1-6 were assigned to $L1$-$L6$. The gold-standard data thus built without using the assessments from J1 is referred to as LOO1 (Leave One Out - judge 1). The other three data sets are named similarly.

Table 24 shows the distribution of relevant answers over relevance levels for each of our leave-one-out data sets.

Tables 25-28 summarise the results of our experiments using the leave-one-out data. For each metric, the entire ranking has been compared to the case with the GAW data, and the rank changes are indicated by "GAW↑1" and so on. Figures 13, 14 and 15 compare the rankings based on leave-one-out data with those based on GAW for each metric.

1. When evaluated based on leave-one-out data, the judge who has been left out performs relatively poorly (i.e. underperforms many systems). For example, in Table 26 which shows the LOO2-based results, J2 moved down six ranks in the nCG@1 column, and significantly underperforms BASELINE-2 at $\alpha = 0.01$ in terms of LOO2-nC@1; in the same table, J2 moved down three ranks in the Q column, and significantly underperforms J4 at $\alpha = 0.01$ in terms of LOO2-Q. Similarly, in Table 27 which shows the LOO3-based results, J3 moved down five ranks in the nCG@1 column, and significantly underperforms ASURA-2 at $\alpha = 0.01$ in terms of LOO3-nCG@1; it moved down eight ranks in the nDCG column, and significantly underperforms BASELINE-2 at $\alpha = 0.01$ in terms of LOO3-nDCG; it moved down nine ranks in the Q column, and significantly underperforms MSRA+MSR-3 in terms of LOO3-Q (although mean LOO3-Q ranks J3 above MSRA+MSR-3).

2. Apart from the above rank changes, the LOO-based rankings are very similar to the original GAW rankings. It is still true that many systems lie between different human performances.

Table 24: Distribution of answers over relevance levels for the leave-one-assessor-out data.

	LOO1	LOO2	LOO3	LOO4
$L6$	1366	2091	1808	1446
$L5$	1647	2015	2180	1737
$L4$	1963	1574	1689	2077
$L3$	2081	1406	1501	1786
$L2$	272	268	171	280
$L1$	82	70	68	93
$L0$	32	19	26	24
total	7443	7443	7443	7443

Since different judges perform quite differently when viewed as systems (whether the judge's assessments are included in or excluded from the gold-standard data), it is probably a good idea to hire multiple assessors and create a pyramid-based graded-relevance gold-standard data as we did in this paper. Moreover, since our GAW-based results are similar to the official GA-based results, probably the GAW-based approach, which involves nine relevance levels, is preferrable to GA for future community QA task evaluation. This is because (a) the mapping from relevance patterns to levels can be done more systematically than with GA; and (b) GAW retains more fine-grained information than GA through the use of many relevance levels, and this can always be "coarsened" if required. For example, if we decide to use four relevance levels (including $L0$), we can use the nine-level GAW data but assign the gain values as $L1 : L2 : L3 : L4 : L5 : L6 : L7 : L8 = 1 : 1 : 1 : 2 : 2 : 2 : 3 : 3$ and so on. Thus, "qrels" can retain the nine relevance levels, and can be coarsened at the time of computing the evaluation metrics when necessary.

5. CONCLUSIONS

This paper presented an overview of the NTCIR-8 Community QA Pilot Task. We first presented the official results based on the Good Answers (GA) data as well as the Best Answer (BA) data. We also analysed the results for each question category, and examined question hardness, defined as the performance averaged across runs. Moreover, we proposed a more systematic method for aggregating the assessments of multiple assessors, and used the resultant "Good Answers with Weights" (GAW) data to re-evaluate the

Table 25: Mean performances with the full question set based on the LOO1 data. For each metric, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	LOO1-nG@1		LOO1-nDCG		LOO1-Q
J3	0.9565	J3	0.9865**	J3	0.9789**
J2	0.9438**	J2 (GAW↑2)	0.9791†	J2 (GAW↑6)	0.9664‡
MSRA+MSR-2	0.9162	MSRA+MSR-2 (GAW↓1)	0.9763	MSRA+MSR-4	0.9660
MSRA+MSR-1	0.9151	MSRA+MSR-4 (GAW↓1)	0.9761	MSRA+MSR-2 (GAW↓2)	0.9659
MSRA+MSR-4	0.9149	MSRA+MSR-1	0.9756	ASURA-2 (GAW↓1)	0.9653
BASELINE-2 (GAW↑1)	0.9124	ASURA-2	0.9754*	MSRA+MSR-1 (GAW↓1)	0.9650
ASURA-2 (GAW↓1)	0.9115	BASELINE-2 (GAW↑1)	0.9753	BASELINE-2	0.9649
ASURA-1	0.9108**	ASURA-1 (GAW↓1)	0.9750**	ASURA-1 (GAW↓2)	0.9645**
MSRA+MSR-3	0.8965	MSRA+MSR-3	0.9711**	MSRA+MSR-3	0.9587**
BA	0.8918	J4 (GAW↑1)	0.9684	J4 (GAW↑1)	0.9547
MSRA+MSR-5	0.8851**	MSRA+MSR-5 (GAW↑1)	0.9650**	MSRA+MSR-5 (GAW↑1)	0.9483**
J4 (GAW↑1)	0.8729**	J1 (GAW↓2)	0.9582	J1 (GAW↓2)	0.9399
J1 (GAW↓1)	0.8440	BASELINE-3	0.9532**	BASELINE-3	0.9344**
BASELINE-3	0.8302**	BASELINE-1	0.9398	BASELINE-1	0.9140
BASELINE-1	0.7867**	LILY-2	0.9312**	LILY-2	0.9076**
LILY-3	0.7158	LILY-3	0.9269**	LILY-3	0.8993**
LILY-2	0.7158	LILY-1	0.9234**	LILY-1	0.8924**
LILY-1	0.7158	BA	0.4335	BA	0.2506

Table 26: Mean performances with the full question set based on the LOO2 data. For each metric, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	LOO2-nG@1		LOO2-nDCG		LOO2-Q
J3	0.9724**	J3	0.9885**	J3	0.9805**
MSRA+MSR-2 (GAW↑1)	0.9293	MSRA+MSR-2	0.9789	MSRA+MSR-4 (GAW↑1)	0.9685
MSRA+MSR-4 (GAW↑2)	0.9281	MSRA+MSR-4	0.9788**	MSRA+MSR-2 (GAW↓1)	0.9684†
MSRA+MSR-1	0.9276	MSRA+MSR-1 (GAW↑1)	0.9783‡	ASURA-2	0.9681**
ASURA-2 (GAW↑1)	0.9255	ASURA-2 (GAW↑1)	0.9783*	MSRA+MSR-1	0.9675
ASURA-1 (GAW↑2)	0.9247	ASURA-1 (GAW↑1)	0.9780	ASURA-1	0.9674
BASELINE-2	0.9246**	BASELINE-2 (GAW↑1)	0.9777**	BASELINE-2	0.9670**
J2 (GAW↓6)	0.9139	J1 (GAW↑2)	0.9758‡	J1 (GAW↑2)	0.9633‡
MSRA+MSR-3	0.9086**	MSRA+MSR-3	0.9739**	MSRA+MSR-3	0.9614**
J1 (GAW↑2)	0.9079‡	J4 (GAW↑1)	0.9734	J4 (GAW↑1)	0.9601**
BA (GAW↓1)	0.9054	J2 (GAW↓7)	0.9700†	J2 (GAW↓3)	0.9521
MSRA+MSR-5 (GAW↓1)	0.9007**	MSRA+MSR-5	0.9682**	MSRA+MSR-5	0.9513**
J4	0.8985**	BASELINE-3	0.9573**	BASELINE-3	0.9378**
BASELINE-3	0.8502**	BASELINE-1	0.9461**	BASELINE-1	0.9205*
BASELINE-1	0.8123**	LILY-2	0.9369**	LILY-2	0.9126**
LILY-3	0.7427	LILY-3	0.9332**	LILY-3	0.9053**
LILY-2	0.7427	LILY-1	0.9301**	LILY-1	0.8991**
LILY-1	0.7427	BA	0.4293	BA	0.2525**

submitted runs as well as the human judges themselves. We have observed that manual answer assessments vary considerably across judges, and that using our pyramid approach that constructs graded-relevance answer data based on multiple judges is useful for finding differences between systems.

Our GAW-based results show that the best runs outperform some humans, but underperform others. Morevoer, the simple length-based baseline is surprisingly hard to beat. Can systems do better? We would like to tackle these questions together with the participants in the near future, by means of a better, practical task design (e.g. separate test data from training data!) and evaluation methods including, but not limited to, the GAW approach.

6. ACKNOWLEDGMENTS

We would like to thank the NTCIR-8 community QA participants for their efforts and the advisors (Kazuko Kuriyama and Yohei Seki) as well as Chin-Yew Lin and Eugene Agichtein for their advice. Yahoo! Chiebukuro Data provided to National Institute of Informatics by Yahoo Japan Corporation was used in the implementation of this research.

7. REFERENCES

[1] Gey, F., Larson, R., Kando, K., Machado-Fisher, J. and Sakai, T.: NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search, *NTCIR-8 Proceedings*, to appear (2010).

[2] Ishikawa, D., Kuriyama, K., Seki, Y. and Kando, N.: Investigation of Possibility of Best-Answer Estimation in Q&A Site (in Japanese), *IPSJ SIG Technical Report*, 2010-FI-97 / 2010-NL-195, No. 8 (2010).

[3] Ishikawa, D., Sakai, T. and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task (Part I): The Test Collection and the Task, *NTCIR-8 Proceedings*, to appear (2010).

[4] Ishikawa, D.: ASURA: A Best-Answer Estimation System for NTCIR-8 CQA Pilot Task, *NTCIR-8 Proceedings*, to appear (2010).

[5] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based

Table 27: Mean performances with the full question set based on the LOO3 data. For each metric, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	LOO3-nG@1		LOO3-nDCG		LOO3-Q
J2 (GAW↑1)	0.9643**	J2 (GAW↑3)	0.9846*	J2 (GAW↑7)	0.9741
MSRA+MSR-2 (GAW↑1)	0.9285	MSRA+MSR-2	0.9793	MSRA+MSR-2	0.9700
MSRA+MSR-1 (GAW↑1)	0.9283	MSRA+MSR-4	0.9790	ASURA-2 (GAW↑1)	0.9697
MSRA+MSR-4 (GAW↑1)	0.9266	ASURA-2 (GAW↑2)	0.9790*	MSRA+MSR-4 (GAW↓1)	0.9694
ASURA-2 (GAW↑1)	0.9265**	MSRA+MSR-1	0.9789**	MSRA+MSR-1	0.9693
J3 (GAW↓5)	0.9261†	J1 (GAW↑4)	0.9788‡	ASURA-1	0.9688**
ASURA-1 (GAW↑1)	0.9245	ASURA-1	0.9784	J1 (GAW↑3)	0.9683‡
BASELINE-2 (GAW↓1)	0.9234**	BASELINE-2	0.9780**	BASELINE-2 (GAW↓1)	0.9683**
J1 (GAW↑3)	0.9157‡	J3 (GAW↓8)	0.9772	J4 (GAW↑2)	0.9653
MSRA+MSR-3 (GAW↓1)	0.9102	J4 (GAW↑1)	0.9765†	J3 (GAW↓9)	0.9645‡
BA (GAW↓1)	0.9085**	MSRA+MSR-3 (GAW↓2)	0.9747**	MSRA+MSR-3 (GAW↓2)	0.9633**
J4 (GAW↑1)	0.9071‡	MSRA+MSR-5	0.9706**	MSRA+MSR-5	0.9553**
MSRA+MSR-5 (GAW↓2)	0.9057**	BASELINE-3	0.9597**	BASELINE-3	0.9423**
BASELINE-3	0.8550**	BASELINE-1	0.9490*	BASELINE-1	0.9256
BASELINE-1	0.8188**	LILY-2	0.9416**	LILY-2	0.9192**
LILY-3	0.7602	LILY-3	0.9382**	LILY-3	0.9128**
LILY-2	0.7602	LILY-1	0.9352**	LILY-1	0.9068**
LILY-1	0.7602	BA	0.4277	BA	0.2541

Table 28: Mean performances with the full question set based on the LOO4 data. For each metric, "$*$" and "$**$" indicate that a run significantly outperforms the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively). Whereas, "†" and "‡" indicate that a run significantly *underperforms* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

	LOO4-nG@1		LOO4-nDCG		LOO4-Q
J3	0.9543	J3	0.9857*	J3	0.9776
J2	0.9406	MSRA+MSR-2	0.9789**	MSRA+MSR-2	0.9697
MSRA+MSR-2	0.9243	J2 (GAW↑1)	0.9788‡	MSRA+MSR-4	0.9696
MSRA+MSR-4 (GAW↑1)	0.9241	MSRA+MSR-4 (GAW↓1)	0.9787	ASURA-2	0.9692*
MSRA+MSR-1 (GAW↓1)	0.9235	MSRA+MSR-1	0.9784†	MSRA+MSR-1	0.9689†
ASURA-2	0.9202	ASURA-2	0.9780	BASELINE-2 (GAW↑1)	0.9684
BASELINE-2	0.9195	BASELINE-2 (GAW↑1)	0.9774	ASURA-1 (GAW↓1)	0.9682**
ASURA-1	0.9184**	ASURA-1 (GAW↓1)	0.9773**	J2	0.9659
MSRA+MSR-3	0.8989*	MSRA+MSR-3	0.9723**	MSRA+MSR-3	0.9605**
BA	0.8932	J1	0.9706	J1	0.9574
MSRA+MSR-5	0.8857**	MSRA+MSR-5 (GAW↑1)	0.9653**	MSRA+MSR-5 (GAW↑1)	0.9487**
J1	0.8819**	J4 (GAW↓1)	0.9553†	BASELINE-3 (GAW↑1)	0.9365*
BASELINE-3 (GAW↑1)	0.8339*	BASELINE-3	0.9546**	J4 (GAW↓2)	0.9364**
J4 (GAW↓1)	0.8317**	BASELINE-1	0.9413**	BASELINE-1	0.9165
BASELINE-1	0.7912**	LILY-2	0.9303**	LILY-2	0.9073**
LILY-3	0.7104	LILY-3	0.9258**	LILY-3	0.8987**
LILY-2	0.7104	LILY-1	0.9218**	LILY-1	0.8908**
LILY-1	0.7104	BA	0.4354**	BA	0.2516

Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422-446 (2002).

[6] Kuriyama, K.: Best-Answer Selection Using a Machine Learning Tool at NTCIR8 CQA Pilot Task, *NTCIR-8 Proceedings*, to appear (2010).

[7] Lin, J. and Demner-Fushman, D.: Will Pyramids Built of Nuggets Topple Over? *HLT/NAACL 2006 Proceedings*, pp. 383-390 (2006).

[8] Nenkova, A., Passonneau, R. and McKeown, K.: The Pyramid Method: Incorporating Human Content Selection Variation in Sumarization Evaluation, *ACM Transactions on Speech and Language Processing*, Volume 4, Number 2, Article 4 (2007).

[9] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Volume 43, Issue 2, pp.531-548 (2007).

[10] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp. 32-43 (2007). `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/EVIA/1.pdf`

[11] Sakai, T., Shima, H., Kando, N., Song, R., Lin, C.-J., Mitamura, T., Sugimoto, M. and Lee, C.-W.: Overview of NTCIR-8 ACLIA IR4QA, *NTCIR-8 Proceedings*, to appear (2010).

[12] Sakai, T., Ishikawa, D., Seki, Y., Kando, N. and Kuriyama, K.: Selecting Good Answers for Community QA: A Note on Evaluation Methods (in Japanese), *Forum on Information Technology 2010*, to appear (2010).

[13] Song, Y.-I., Liu, J., Sakai, T., Cao, Y., Lin, C.-Y., Wang, X.-J., Feng, G. and Suzuki, H.: Microsoft Research Asia with Redmond at the NTCIR-8 Community QA Pilot Task, *NTCIR-8 Proceedings*, to appear (2010).

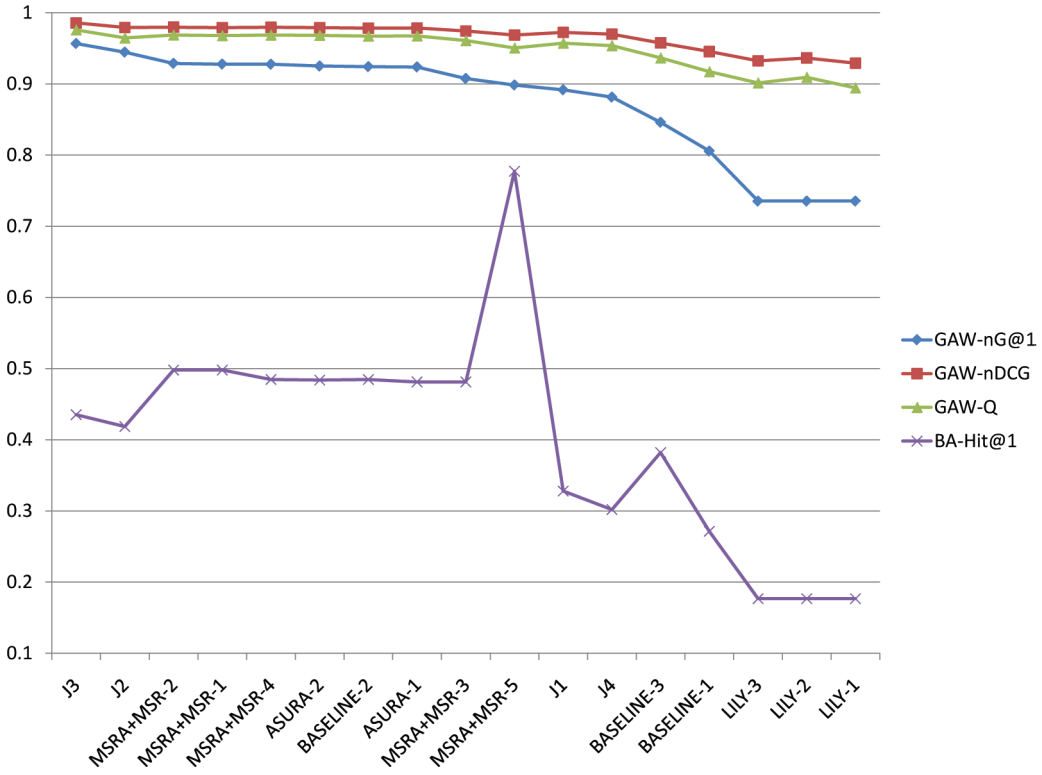

Figure 12: Comparison of system rankings based on GAW-nDCG / GAW-Q / BA-Hit@1 with that based on GAW-nG@1.

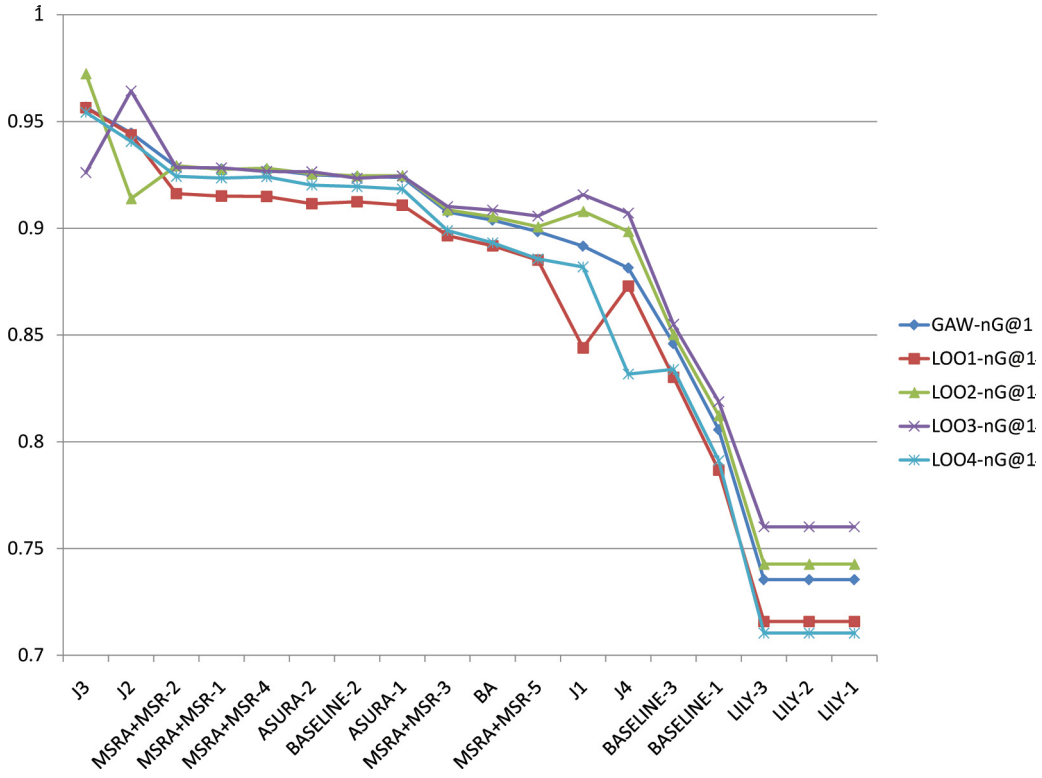

Figure 13: Comparison of system rankings based on the LOO data with that based on the GAW data (nG@1).

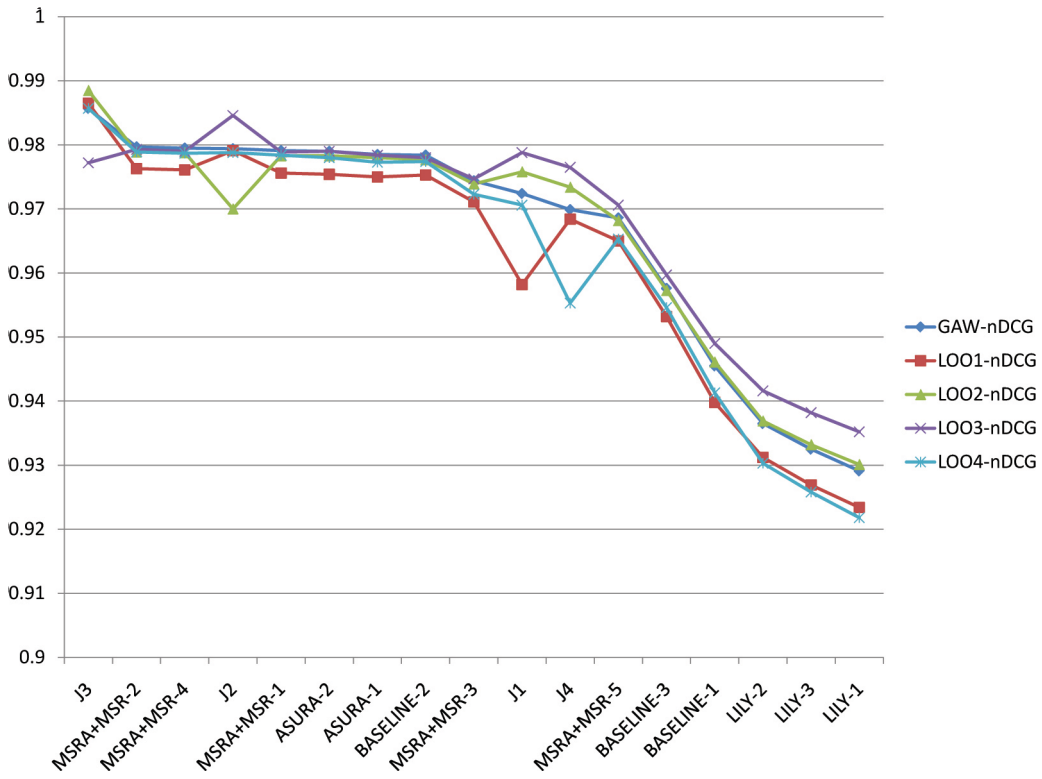

Figure 14: Comparison of system rankings based on the LOO data with that based on the GAW data (nDCG).

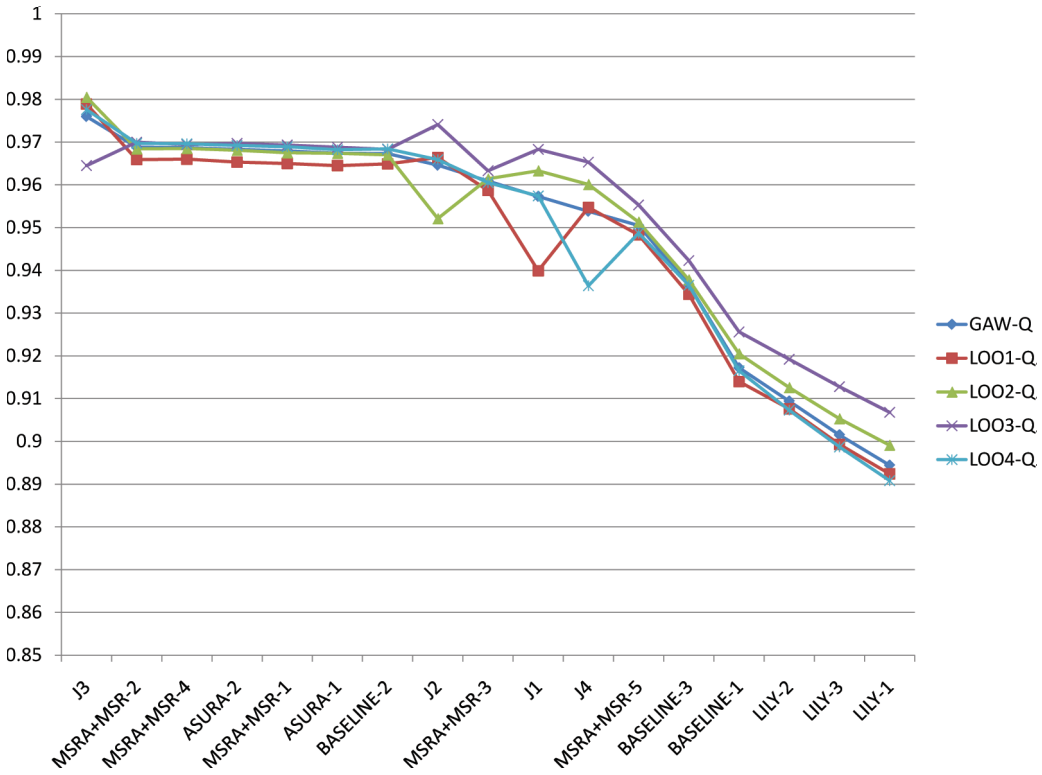

Figure 15: Comparison of system rankings based on the LOO data with that based on the GAW data (Q).