A Note on the Reliability of Japanese Question Answering Evaluation

Tetsuya SAKAI

Knowledge Media Laboratory, Toshiba Corporate R&D Center
tetsuya.sakai@toshiba.co.jp

Abstract

This paper compares some existing QA evaluation metrics from the viewpoint of reliability and usefulness, using the NTCIR-4 QAC2 Japanese QA tasks and our adaptations of Buckley/Voorhees and Voorhees/Buckley reliability measurement methods. Our main findings are: (1) The fraction of questions with a correct answer within Top 5 (NQcorrect5) and that with a correct answer at Rank 1 (NQcorrect1) are not as stable as Reciprocal Rank based on ranked lists containing up to five answers. (2) Q-measure, which can handle multiple correct answers and answer correctness levels, is as reliable and useful as Reciprocal Rank, provided that a *mild* gain value assignment is used. Emphasising answer correctness levels tends to hurt stability, while handling multiple correct answers improves it.

1 Introduction

Question Answering (QA) has received a lot of attention in recent years, but researchers are still exploring how best to evaluate QA. Factoid, list and definitional questions were included in a single task at the TREC 2003 track, but different metrics were used for different questions types: the fraction of correct responses was used for the factoid questions (as systems were required to return a single answer for each question), while *instance-based* and *nugget-based* variations of F-measure were used for the list questions and definitional questions, respectively [13]. At the NTCIR Japanese QA tasks (QAC1 and 2), Reciprocal Rank (RR) was used for evaluating ranked lists containing up to five exact answers in Subtask 1 (factoid), while instance-based F-measure was used for evaluating *sets* of exact answers in Subtasks 2 (list) and 3 (context) [3].

This paper discusses the reliability of some existing QA metrics based on the QAC2 task settings. In particular, we focus on Subtask 1 where systems were required to return ranked lists of exact answers, and examine RR, the fraction (or number) of questions with a correct answer within Top 5 (1) of the answer list (NQcorrect5 and NQcorrect1), and recently-proposed Q-measure [7, 8, 9]. In order to compare the reliability and usefulness of these QA metrics, we adopt methods proposed by Buckley and Voorhees [1] and Voorhees and Buckley [12]. In addition, we examine F-measure for QAC2 Subtask 2 using the same methods.

The remainder of this paper is organised as follows. Section 2 describes our adaptations of the reliability measurement methods. Section 3 discusses the reliability and usefulness of RR, NQcorrect5 and NQcorrect1 based on the formal run results at QAC2 Subtask 1. Section 4 compares the reliability and usefulness of Q-measure with the above "official" metrics, using our own set of QAC2 Subtask 1 results. (We cannot compute Q-measure for all of the formal runs, because only the RR values, not the actual system output files, are available to us.) Section 5 briefly discusses the reliability and usefulness of F-measure for comparing the runs submitted to QAC2 Subtask 2. Finally, Section 7 concludes this paper.

2 Algorithms for Measuring Reliability

This section describes our adaptation of two existing methods for studying the reliability of test collections and effectiveness metrics. The first method, proposed by Buckley and Voorhees [1], computes the *minority rate* and the *proportion of ties*, given an effectiveness metric M, a test collection, and a set of *runs* submitted to a particular task defined by the collection. The second method, proposed by Voorhees and Buckley [12], derives the minimum performance difference (measured by M) required in order to conclude that System x is better than System y with a given *confidence level*, given a test collection and a set of runs. (Both the minority rate and the swap rate were originally called the "error rate".)

2.1 Minority Rate / Proportion of Ties

First, we describe the method for computing the minority rate and the proportion of ties. Let S denote a set of systems (i.e. runs) submitted to a particular task, and let x and y denote a pair of systems from S. Let Q denote the entire set of questions (or topics) used in the task, and let c denote a constant. Let $M(Q_i, x)$ denote the value of metric M for System x averaged over a topic set $Q_i (\subset Q)$. Then, using the algorithm shown in Figure 1, the minority rate and the proportion of ties of M, given a *fuzziness value* f, can be computed as:

for each pair of runs $x, y \in S$
  for each trial from 1 to 1000
    select $Q_i \subset Q$ at random s.t. $|Q_i| == c$;
    $margin = f * \max(M(x, Q_i), M(y, Q_i))$;
    if($|M(x, Q_i) - M(y, Q_i)| < margin$)
      $EQ_M(x, y) + +$
    else if($M(x, Q_i) > M(y, Q_i)$)
      $GT_M(x, y) + +$
    else
      $GT_M(y, x) + +$;

Figure 1. The algorithm for computing $EQ_M(x, y)$,$GT_M(x, y)$ and $GT_M(y, x)$.

$$MinorityRate_M = \frac{\sum_{x,y \in S} \min(GT_M(x, y), GT_M(y, x))}{\sum_{x,y \in S}(GT_M(x, y) + GT_M(y, x) + EQ_M(x, y))} \quad (1)$$

$$PropTies_M = \frac{\sum_{x,y \in S} EQ_M(x, y)}{\sum_{x,y \in S}(GT_M(x, y) + GT_M(y, x) + EQ_M(x, y))} \quad (2)$$

The minority rate is an estimate of the chance of reaching a wrong conclusion about a pair of runs using a given metric, while the proportion of ties reflects its discrimination power. Thus, for a good performande metric, both of these values should be small. However, from the algorithm, it is clear that $GT_M(x, y) + GT_M(y, x) + EQ_M(x, y) = 1000$ for each run pair, and that a larger fuzziness value yields larger $EQ_M(x, y)$ values, and therefore a larger proportion of ties and a smaller minority rate. That is, there is a trade-off between these two statistics. Buckley and Voorhees [1] have used a fixed fuzziness value ($f = 0.05$), but as this may imply different trade-offs for different metrics, we prefer to vary the fuzziness value ($f = 0.01, 0.02, \ldots, 0.10$) and draw *minority-rate / proportion-of-ties* curves for comparing the stability of different metrics.

2.2 Swap Rate / Performance Differences

The minority rate method described above starts by deciding on the fuzziness value (i.e. how much relative difference should be regarded as negligible), and then establishes a relationship between the minority rate and the proportion of ties, *assuming* that $\min(GT_M(x, y), GT_M(y, x))$ represents errors. The second method may be more intuitive and practical, as it starts by setting the confidence level of a conclusion, and then directly measures the absolute difference required to reach the conclusion, using *disjoint* pairs of topic sets.

for each pair of runs $x, y \in S$
  for each trial from 1 to 1000
    select $Q_i \subset Q$ and $Q'_i \subset Q$ s.t.
      $Q_i \cap Q'_i == \phi$ and $|Q_i| == |Q'_i| == c$;
    $d_M(Q_i) = M(x, Q_i) - M(y, Q_i)$;
    $d_M(Q'_i) = M(x, Q'_i) - M(y, Q'_i)$;
    $counter(BIN(d_M(Q_i))) + +$;
    if(($d_M(Q_i) * d_M(Q'_i) < 0$) or
      ($d_M(Q_i) == 0$ and $d_M(Q'_i) \neq 0$) or
      ($d_M(Q_i) \neq 0$ and $d_M(Q'_i) == 0$))
      $swap\_counter(BIN(d_M(Q_i))) + +$;
for each bin b
  $swap\_rate(b) = swap\_counter(b)/counter(b)$;

Figure 2. The algorithm for computing the swap rates.

Moreover, it is possible to discuss the practical usefulness of metrics based on this method.

Let d denote a performance difference between two systems. The second method begins by defining 21 *performance difference bins*, where the first bin represents performance differences such that $0 \leq d < 0.01$, the second bin represents those such that $0.01 \leq d < 0.02$, and so on, and the last bin represents those such that $0.20 \leq d$. Let $BIN(d)$ denote a mapping from a difference d to one of the 21 bins where it belongs. The algorithm shown in Figure 2 calculates a *swap rate* for each bin [12]. Our test is stricter than the original one by Voorhees and Buckley, in that our "swap" includes cases in which one of $d_M(Q_i)$ and $d_M(Q'_i)$ is zero. This is because Voorhees and Buckley's original test, which increments the swap counter only when one of the differences is positive and the other is negative, tends to underrate the swap rates for near-zero bins as the differences are actually quite often zero. (We have verified that this modification gives graphs that look more stable, but do not affect our conclusions.)

Because Q_i and Q'_i must be disjoint, they can only be up to half the size of the original topic set Q. (Voorhees and Buckley have used *extrapolation* for larger topic set sizes, but we stick to the statistics *actually measured* in our study, as our objective is to compare the reliability of different metrics under the same conditions.) Given a confidence level (e.g. 95%), we can plot the swap rate (i.e. 1 minus the confidence level) against the performance difference bins, so that the mininum difference required to reach a conclusion about system comparisons can be obtained [12, 13, 14, 15].

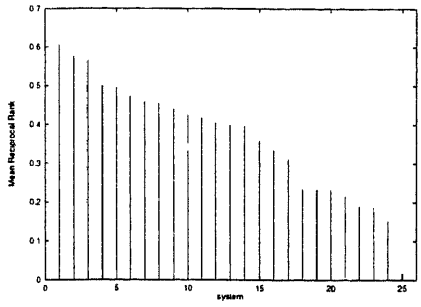

Figure 3. QAC2 Subtask 1 formal run performances (Mean Reciprocal Rank).

3 QAC2 Subtask 1 Formal Runs: Reciprocal Rank

This section examines the results of NTCIR-4 QAC2 Subtask 1 [3], where systems returned a ranked list containing up to five exact answers for each of the 195 formal run questions. Our analyses here are based on the RR values of the 25 systems submitted to this task, which were included in the CD-ROM distributed at the NTCIR-4 Workshop. RR is defined as $1/r'$, where r' is the rank of the first correct response within the answer list, or zero if the answer list does not contain a correct answer at all. Thus, RR is either $1, 1/2, 1/3, 1/4, 1/5$ or 0.

Figure 3 shows the Mean RR (MRR) values for the 25 formal runs. In terms of RR with the Sign Test, 68% of the $25 * 24/2 = 300$ run pairs are significantly different at $\alpha = 0.01$, and an additional 7% are significantly different at $\alpha = 0.05$.

The QAC2 organisers have considered the Fraction (or Number) of Questions with a correct answer within Top 5 (1) (NQcorrect5 and NQcorrect1) as alternative metrics [3]. These metrics are merely *coarse reductions* of RR: NQcorrect5 is 1 if $RR > 0$ and 0 otherwise; NQcorrect1 is 1 if $RR = 1$ and 0 otherwise. Thus, we can derive NQcorrect5 and NQcorrect1 values from the RR values included in the NTCIR-4 CD-ROM. It is self-evident that NQcorrect5 and NQcorrect1 are highly correlated with RR.

As there are 195 questions, up to $c = 97$ questions can be used for the swap rate measurement described in Section 2.2. We therefore tried $c = 97$ and $c = 50$ with both the swap rate and the minority rate measurement. Thus, the questions we address here is: "When using approximately 100 (50) questions for comparing systems submitted to QAC2 Subtask 1, how reliable and useful are RR, NQcorrect5 and NQcorrect1?" Note that the *absolute* values in the results reported below are dependent on the test collection and the set of runs. What we are interested here is how different QA metrics compare to one another in terms of reliability and usefulness under the same conditions.

Figure 4 shows the minority-rate / proportion-of-ties curves for RR, NQcorrect5 and NQcorrect1 when $c = 97$. The horizontal line indicates when the minority rate is 1%. These results suggest that, not surprisingly, RR is more stable than NQcorrect5, which in turn is more stable than NQcorrect1.

Figures 5 and 6 plot the swap rates againt the performance difference bins when $c = 97$ and $c = 50$, respectively. The horizontal line indicates when the swap rate is 5% (i.e. 95% confidence level). Again, RR appears to be more stable than the other two.

Table 1 summarises Figures 5 and 6 from a practical point of view: For example, if you are using RR with only 97 questions to compare the QAC2 Subtask 1 formal runs, you should look for a performance difference of at least 0.12 in order to conclude that System x is better than System y with 95% confidence. As the maximum MRR observed among the 50,000 values (1000 trials for 25 systems, each with 2 disjoint topic sets) is 0.7156, this translates to a relative difference of at least 17%. Of the 300,000 comparisons (1000 trials for 300 system pairs), 59.8% actually have this difference. Thus, the last column of this table represents the discrimination power, or *usefulness* of a metric. It can be observed that, for discriminating the systems submitted to QAC2 Subtask 1, using NQcorrect5 instead of RR may suffice. This suggests that most of the differences among the submitted systems arise from whether they managed to include a correct answer somewhere in the list, rather than *where* in the list the correct answer was.

The table also shows that NQcorrect1 (i.e. looking at the first response only) is less useful. One way to improve the situation with NQcorrect1 is to use a very large topic set (e.g. hundreds of questions), as the TREC 2003 QA track did with factoid questions [13].

To sum up, RR is more stable than NQcorrect5 and NQcorrect1, although NQcorrect5 may suffice for discriminating the systems submitted to QAC2 Subtask 1.

4 QAC2 Subtask 1 ASKMi Runs: Q-measure and Reciprocal Rank

Section 3 used the formal runs results at QAC2 Subtask 1 to verify that RR is more reliable than NQcorrect1 and NQcorrect5. This section examines Q-measure [7, 8, 9] using similar methods, but use *our own* set of runs generated for QAC2 Subtask 1, because only the RR values are available for the formal runs submitted to QAC2 Subtask 1, and we cannot calculate Q-measure values for all of the submitted runs. As our runs are generated using a *single* system (though with substantially different parameter settings), the absolute values (e.g. the minority rate and the swap rate) obtained here may be of little value. Note also that Section 3 used 25 runs, while here we use only 10 runs, which inevitably

Table 1. The sensitivity of metrics at 95% confidence level for QAC2 Subtask 1 formal runs.

metric	absolute diff required	max performance among 50,000 values	relative diff required	#comparisons (out of 300,000) with required diff
(a) $c = 97$				
RR	0.12	0.7156	17%	59.8%
NQcorrect5	0.14	0.8454	17%	59.2%
NQcorrect1	0.14	0.6289	22%	49.1%
(ii) $c = 50$				
RR	0.17	0.8006	21%	46.1%
NQcorrect5	0.20	0.9200	22%	46.0%
NQcorrect1	0.20	0.7600	26%	34.1%

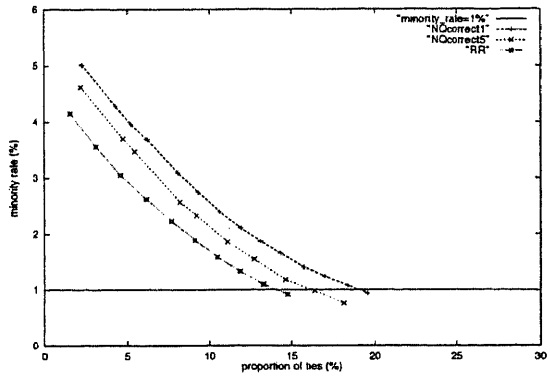

Figure 4. Minority rate / proportion of ties for QAC2 Subtask 1 formal runs ($c = 97$).

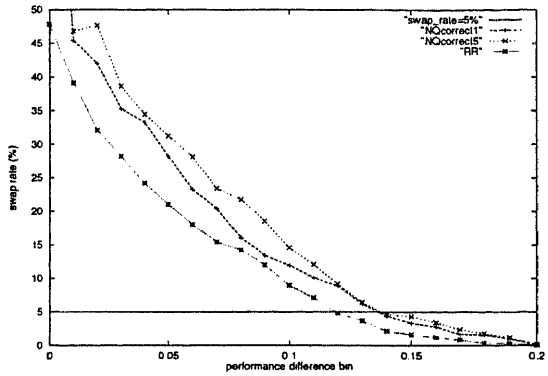

Figure 5. Swap rates / performance differences for QAC2 Subtask 1 formal runs ($c = 97$).

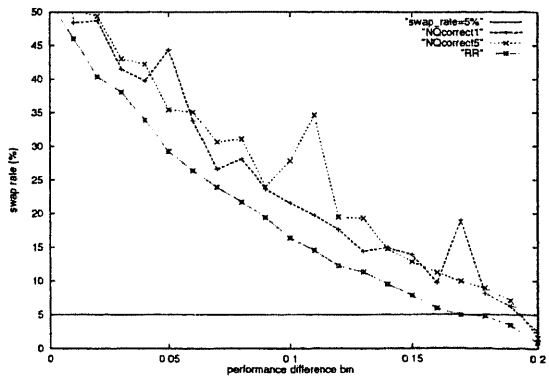

Figure 6. Swap rates / performance differences for QAC2 Subtask 1 formal runs ($c = 50$).

makes the results less stable. Nevertheless, this experimental setting should suffice for comparing different metrics from the viewpoint of reliability and usefulness.

Q-measure is basically an *Information Retrieval* (IR) metric based on graded relevance [9]. However, by assigning *correctness levels* to answer strings (just like assigning *relevance levels* to documents in IR) and defining *answer equivalence classes* for penalising inclusion of duplicates in the ranked list, Q-measure can be applied to QA evaluation [7, 8]. Thus, unlike RR, Q-measure can handle multiple correct answers *and* answer correctness levels. The question is how Q-measure compares to RR in terms of reliability and usefulness. The Appendix contains the formal definition of Q-measure as an IR metric.

The ASKMi Japanese QA system [6, 7] was used to generate the following 10 runs (in order of descreasing MRR) for the QAC2 Subtask 1 with 195 questions:

1. An oracle run, with *correct answer types* and *correct supporting documents* given. This is the run TSB-A+OAT+OSD described in [7].

2. An oracle run, with *correct supporting documents* given. This is TSB-A+OSD described in [7].

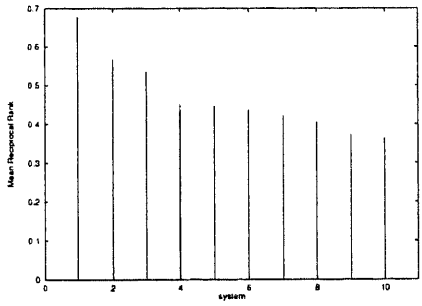

Figure 7. QAC2 Subtask 1 ASKMi performances (Mean Reciprocal Rank).

3. An oracle run, with *correct answer types* given. This is TSB-A+OAT described in [7].

4. A formal run actually submitted to QAC2, called TSB-A [7]. Thus this is one of the 25 runs used in Section 3.

5. This is the same as TSB-A, except that the document score parameter P_d was set to 0 instead of 1 [6, 7, 10]. That is, this run *ignored the document scores completely*: the answer scores were calculated based only on distances between answer candidates and query terms within each document.

6. This is the same as TSB-A, except that the Okapi/BM25 parameter b was set to 0 instead of the default value 0.75 for document retrieval [6]. This means that *document length normalisation was switched off*.

7. This is the same as TSB-A, except that *the Answer Formulator*, which tries to erase duplicates in the answer list [6], was switched off. This run is similar to the noAF run described in [8].

8. This is the same as TSB-A, except that *Pseudo-Relevance Feedback* (PRF) was activated [6], using the top 10 documents and 40 expansion terms. As the number of "relevant" (i.e. supporting) documents is generally small in QA test collections, PRF actually hurts document retrieval performance and the final QA performance.

9. This is the same as TSB-A, except that the candidate score parameter P_c was set to 0 instead of the default value 0.1 [6, 7, 10]. This means that *the distances between answer candidates and query terms were ignored completely*: only the document scores contributed to the final answer scores (*cf.* Run No. 5).

10. This is the same as TSB-A, except that top 50 documents, instead of the default 10, are used for extracting candidate answers.

Figure 7 shows the MRR values of these 10 ASKMi runs evaluated using the official answer file `QAC2formalAnsTask1_040308`. In terms of RR with the Sign Test, 69% of the $10 * 9/2 = 45$ run pairs are significantly different at $\alpha = 0.01$, and an additional 13% are significantly different at $\alpha = 0.05$. Note that this result is quite similar to that for the actual submitted runs (See Section 3). Thus, although this experiment uses runs generated by a single system, it may be a reasonable mimic of a true QAC2 Subtask 1 environment as the runs are actually quite different from each other.

Figure 8 shows the minority-rate / proportion-of-ties curves for RR, NQcorrect5, NQcorrect1 and Q-measure, calculated based on the 10 ASKMi runs. The Q-measure values are computed based on answer correctness levels and equivalence classes that we devised for QAC2 as described in [7, 10]. By default, Q-measure uses the gain values of 3, 2 and 1 for *S-correct*, *A-correct* and *B-correct* answers, respectively [8]. However, in order to separate the effect of introducing correctness levels from that of handling multiple correct answers, we have also tried the "flat" gain value assignment, i.e., $gain(S) = gain(A) = gain(B) = 1$ (See the Appendix), and this is denoted by "Q1:1:1". In addition, we have tried using $gain(S) = 2$, $gain(A) = 1.5$ and $gain(B) = 1$, denoted by "Q2:1.5:1", which represents a gain value assignment in between the default and the flat one. It can be observed that:

- Q2:1.5:1 and RR are equally stable. Moreover, Q1:1:1 is more stable than these two, while Q-measure (with default gain values) is less so. This means that too much emphasis on the answer correctness levels hurts stability, while handling multiple correct answers improves stability.

- Again, NQcorrect5 and NQcorrect1 are not as stable as RR.

- The minority rates computed based on the 10 ASKMi runs is generally lower than those computed based on the actual submitted runs shown in Figure 4. Thus the ASKMi runs are generally easier to discriminate than the submitted runs.

Figures 9 and 10 plot the swap rates againt the performance difference bins when $c = 97$ and $c = 50$, respectively, calculated based on the 10 ASKMi runs. Table 2 interprets the figures in a way similar to Table 1. By looking at the last column of this table, it can be observed that the discrimination power (i.e. usefulness) of Q-measure is comparable to that of RR, and that NQcorrect5 and NQcorrect1 are not as good.

To sum up, the reliability of Q-measure is probably comparable to RR, provided that too much emphasis on answer correctness levels is avoided. Using answer correctness levels tends to hurt stability, while handling multiple correct answers improves it.

Table 2. The sensitivity of metrics at 95% confidence level for QAC2 Subtask 1 ASKMi runs.

metric	absolute diff required	max performance among 20,000 values	relative diff required	#comparisons (out of 45,000) with required diff
(a) $c = 97$				
Q1:1:1	0.05	0.6967	7%	66.2%
Q2:1.5:1	0.05	0.6890	7%	65.2%
Q-measure	0.05	0.6860	7%	65.1%
RR	0.06	0.7940	8%	64.3%
NQcorrect1	0.09	0.7423	12%	51.0%
NQcorrect5	0.09	0.8866	10%	49.5%
(a) $c = 50$				
Q1:1:1	0.08	0.7774	10%	49.9%
RR	0.09	0.8350	11%	49.6%
Q2:1.5:1	0.08	0.7743	10%	49.1%
Q-measure	0.08	0.7738	10%	48.9%
NQcorrect5	0.12	0.9600	13%	40.3%
NQcorrect1	0.15	0.8000	19%	29.9%

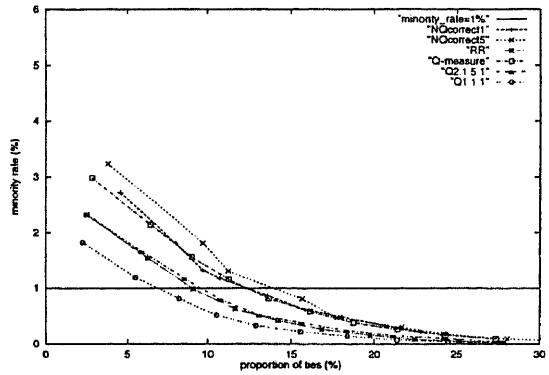

Figure 8. Minority rate / proportion of ties for QAC2 Subtask 1 ASKMi runs ($c = 97$).

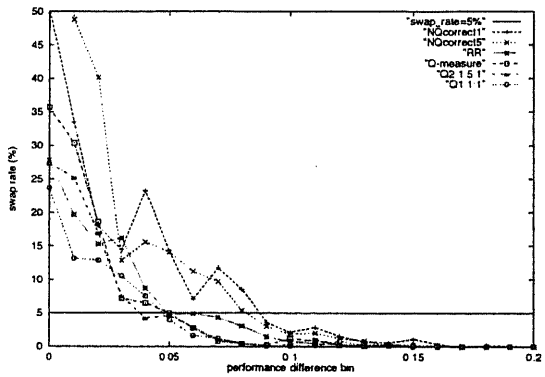

Figure 9. Swap rates / performance differences for QAC2 Subtask 1 ASKMi runs ($c = 97$).

5 QAC2 Subtask 2 Formal Runs: F-measure

We finally take a brief look at the formal runs submitted to QAC2 *Subtask 2*, where systems were required to return *sets* (i.e. unordered list) of answers. The official measure used for ranking systems was instance-based F-measure [13]. That is, inclusion of duplicate answers is penalised, just like Q-measure does with answer equivalence classes for *ranked* lists of answers. For this "list task", we cannot compute any alternative metrics (e.g. [5]) because only the F-measure values, not the actual system output files, are available to us. Thus, we simply measure the reliability and usefulness of F-measure based on the actual runs submitted to QAC2 Subtask 2, just to check that F-measure was adequate for ranking the submitted systems.

Figure 11 shows the Mean F-measure values for the 14 runs submitted to QAC2 Subtask 2. In terms of F-

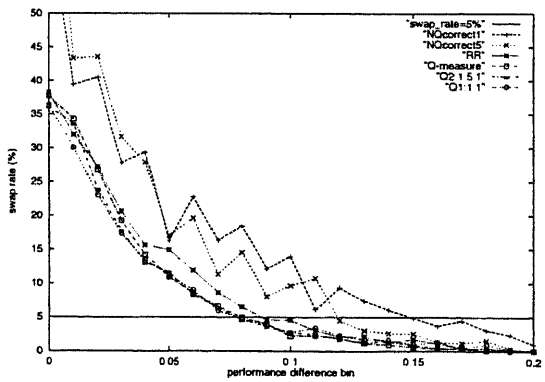

Figure 10. Swap rates / performance differences for QAC2 Subtask 1 ASKMi runs ($c = 50$).

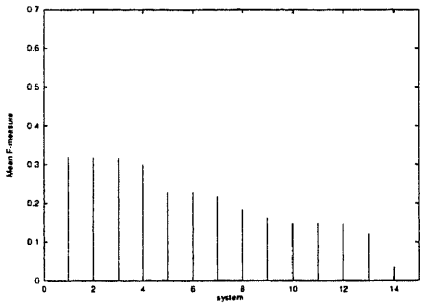

Figure 11. QAC2 Subtask 2 formal run performances (Mean F-measure).

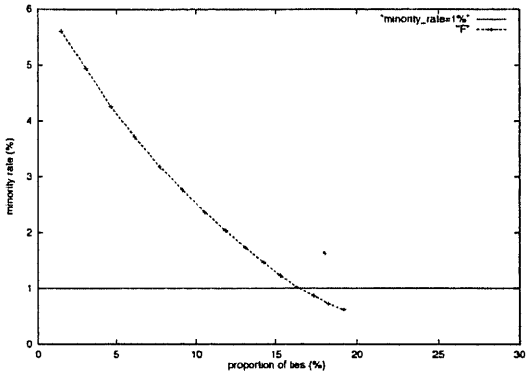

Figure 12. Minority rate / proportion of ties for QAC2 Subtask 2 formal runs ($c = 97$).

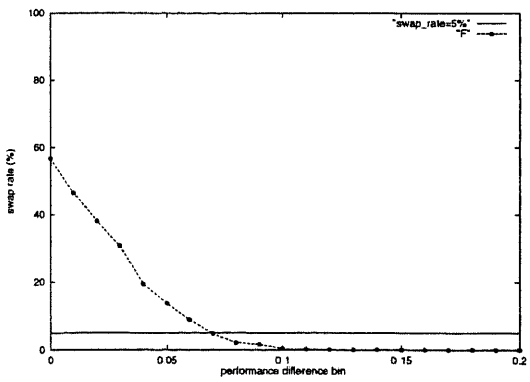

Figure 13. Swap rates / performance differences for QAC2 Subtask 2 formal runs ($c = 97$).

measure with the Sign Test, 63% of the $14 * 13/2 = 91$ run pairs are significantly different at $\alpha = 0.01$, and an additional 15% are significantly different at $\alpha = 0.05$.

Figure 12 shows the minority-rate / proportion-of-ties curve for F-measure based on the 14 submitted runs. Figure 13 plots the swap rate against the performance difference bins when $c = 97$. Table 3 interprets Figure 13: Thus, if the runs submitted to QAC2 Subtask 2 are to be compared using roughly 100 questions, one should look for differences of at least 18% if he wants to be 95% confident. Approximately 60% of the total comparisons have this difference, which is similar to the situation with RR for Subtask 1. As both QAC2 Subtasks 1 and 2 actually used nearly 200 questions, more than 70% of the run pairs may in fact have been distinguishable. However, unlike MRR, it is known that F-measure is extremely sensitive to the changes in the list of correct answers (e.g. addition of newly discovered correct answers) [5], and this remains a serious problem.

6 Related Work

Voorhees [13] has used the swap rate measurement for assessing the reliability of instance-based F-measure for list questions and nugget-based F-measure for definitional questions used at the TREC 2003 QA track. To our knowledge, the present study is the first to address the reliability issues for *Japanese* QA evaluation.

For *IR* evaluation, Buckley and Voorhees [2] and Voorhees [14, 15] further exploited their minority rate and/or swap rate measurement methods. Soboroff [11] used the swap rate measurement for the TREC Web track, where Reciprocal Rank was used for the Known-Item Search task. While these studies considered IR metrics with *binary* relevance only, we are currently investigating the reliability of IR metrics for *graded* relevance, including Q-measure, using the NTCIR *CLIR* task results.

7 Conclusions and Future Work

This paper compared existing QA evaluation metrics using the NTCIR-4 QAC2 Japanese QA tasks from the viewpoint of reliability and usefulness. Our main conclusions are:

- The fraction of questions with a correct answer within Top 5 (NQcorrect5) and that with a correct answer at Rank 1 (NQcorrect1) are not as stable as Reciprocal Rank.
- Q-measure, which can handle multiple correct answers and answer correctness levels, is as reliable and useful as Reciprocal Rank, provided that a *mild* gain value assignment is used. Emphasising answer correctness levels tends to hurt stability, while handling multiple correct answers improves it.

Table 3. The sensitivity of F-measure at 95% confidence level for QAC2 Subtask 2 formal runs ($c = 97$).

metric	absolute diff required	max performance among 28,000 values	relative diff required	#comparisons (out of 91,000) with required diff
F-measure	0.07	0.3979	18%	62.2%

Appendix: Q-measure

Let X denote a relevance level, and let $gain(X)$ denote the *gain value* for successfully retrieving an X-relevant document. Further, let L denote the size of a given ranked output and let $X(r)$ denote the relevance level of the document at Rank r ($\leq L$). Then, the *gain at Rank* r is given by $g(r) = gain(X(r))$ if the document at Rank r is relevant, and $g(r) = 0$ if it is non-relevant. The *cumulative gain at Rank* r is given by $cg(r) = g(r) + cg(r-1)$ for $r > 1$ and $cg(1) = g(1)$ [4]. In particular, let $cig(r)$ denote the cumulative gain at Rank r for an *ideal* ranked output. For example, an ideal ranked output for the NTCIR CLIR tasks should have all the *S-relevant* documents at the top, followed by all *A-relevant* documents, followed by all *B-relevant* documents.

We now introduce the *bonused* gain at Rank r, simply given by $bg(r) = g(r) + 1$ if $g(r) > 0$ and $bg(r) = 0$ if $g(r) = 0$. Thus, the system receives an extra reward for finding a relevant document. Then, the *cumulative bonused gain at Rank* r is given by $cbg(r) = bg(r) + cbg(r-1)$ for $r > 1$ and $cbg(1) = bg(1)$. Q-measure is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cbg(r)}{cig(r) + r} \quad (3)$$

where R is the total number of relevant documents and $isrel(r) = 1$ if the document at Rank r is relevant and $isrel(r) = 0$ otherwise.

Q-measure is equal to one if and only if a system output (s.t. $L \geq R$) is an ideal one. It is very highly correlated with TREC Average Precision [8, 9].

References

[1] Buckley, C. and Voorhees, E. M.: Evaluating Evaluation Measure Stability, *ACM SIGIR 2000 Proceedings*, pp. 33-40, 2000.

[2] Buckley, C. and Voorhees, E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2004.

[3] Fukumoto, J., Kato, T. and Masui, F.: Question Answering Challenge for Five Ranked Answers and List Answers – Overview of NTCIR4 QAC2 Subtask 1 and 2 –. *NTCIR-4 Working Notes*, pp.283-290, 2004.

[4] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422-446, 2002.

[5] Kato, T. *et al.*: Characterization of List-Type Question Answering and its Evaluation Measures (in Japanese), *IPSJ SIG Notes*, FI–76–16 / NL–163–16, pp. 115-122, 2004.

[6] Sakai, T. *et al.*: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, *RIAO 2004 Proceedings*, pp. 215-231, 2004.

[7] Sakai, T. *et al.*: Toshiba ASKMi at NTCIR-4 QAC2, *NTCIR-4 Proceedings*, to appear, 2004.

[8] Sakai, T.: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, *NTCIR-4 Proceedings*, to appear, 2004.

[9] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *AIRS 2004 Proceedings*, to appear, 2004.

[10] Sakai, T. *et al.*: High-Precision Search via Question Abstraction for Japanese Question Answering, *IPSJ SIG Notes*, FI-76-19/NL-163-19, pp.139-146, 2004.

[11] Soboroff, I.: On Evaluating Web Search with Very Few Relevant Documents, *ACM SIGIR 2004 Proceedings*, pp. 530-531, 2004.

[12] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *ACM SIGIR 2002 Proceedings*, pp. 316-323, 2002.

[13] Voorhees, E. M.: Overview of the TREC 2003 Question Answering Track, *TREC 2003 Proceedings*, 2004.

[14] Voorhees, E. M.: Overview of the TREC 2003 Robust Retrieval Track, *TREC 2003 Proceedings*, 2004.

[15] Voorhees, E. M.: Measuring Ineffectiveness, *ACM SIGIR 2004 Proceedings*, pp. 562-563, 2004.