Proceedings of NTCIR-4, Tokyo, April 2003 - June 2004

New Performance Metrics Based on Multigrade Relevance: Their Application to Question Answering

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

Abstract

This paper proposes two new Information Retrieval performance metrics based on multigrade relevance, called Q-measure and R-measure, which are akin to Cumulative Gain and Average Weighted Precision but are arguably more reliable. We then show how Q-measure can be applied to Question Answering involving ranked lists of exact answers, and discuss its advantages over Reciprocal Rank through an experiment using the QAC1 test collection. The appendices of this paper contain theorem proofs concerning Q-measure and R-measure, as well as a study of Q-measure and R-measure as Information Retrieval evaluation metrics using the runs submitted to the NTCIR-3 CLIR task. We plan to conduct similar experiments for the NTCIR QAC tasks using Q-measure as a Question Answering evaluation metric, if the QAC submission files become available for research purposes.
Keywords: *Q-measure, R-measure, Evaluation.*

1 Introduction

This paper proposes two new Information Retrieval (IR) performance metrics based on multigrade relevance, called Q-measure and R-measure, which are akin to Cumulative Gain [6] and Average Weighted Precision (originally called Weighted Average Precision [8]; See Section 2.3) but are arguably more reliable. We also show how Q-measure can be applied to Question Answering (QA) evaluation involving ranked lists of exact answers, and discuss its advantages over Reciprocal Rank through an experiment using the QAC1 test collection [4]. By providing full details of Q-measure and R-measure, this paper serves as the backbone of our two NTCIR-4 site reports: Q-measure and R-measure are used as *IR* metrics with the NTCIR-4 CLIR test collections in [14], while Q-measure is used as a QA metric with the NTCIR-4 QAC2 test collection in [15].

In the early TREC English QA tracks (TREC-8 through TREC 2001) [18, 19], systems returned up to five candidate answers in decreasing order of confidence for the Main Task, i.e., "single-answer" task. Thus, if we let L and L' denote the system output size and the maximum output size allowed, respectively, then $L \leq L' = 5$ for all "single-answer" questions. *Reciprocal Rank* (RR) was used as the evaluation metric. TREC 2001 also introduced the List Task, in which systems were required to return an *unranked* list of answers. The answers were evaluated using *Accuracy*. The TREC List Task was *explicit* (up to TREC 2002) in that L' was clearly specified within each List question. However, these early TREC QA tracks dealt with fixed-length *text snippets* rather than exact answers.

The first Japanese Question Answering Challenge (QAC1) took place at NTCIR-3 [4]. QAC1 dealt with exact answers instead of text snippets, but basically followed the TREC QA evaluation methodology in that the Main Task (Subtask 1) used Reciprocal Rank with $L' = 5$. On the other hand, the QAC1 List Task (Subtask 2) used *F-measure* rather than accuracy for dealing with unranked answer lists, as the QAC1 List questions were in general *implicit*. Thus, in principle, the system had to determine the system output size L for each List question. (In fact, the QAC1 List question set was identical to the QAC1 Main question set, and the top performer in the List Task simply let $L = 1$ for all questions.) The task settings for NTCIR-4 QAC2 are similar to those for QAC1.

Existing problems in QA evaluation include:

1. Different evaluation metrics need to be used for "different" QA tasks, as each of the metrics has its weaknesses: Reciprocal Rank can only look at the first correct response, while Accuracy and F-measure ignore answer priorities. However, the distinction between the above two tasks is not always clear, as there are more than one correct answer for many seemingly "single-answer" questions. Consider: Q: "What is the official language in Switzerland?" A: "Italian, German and French". It is also impossible to tell whether Q: "Who in Japan received the Nobel Prize in Physics?" is a List question or not unless you

know the answer (or answers).

2. There is no QA evaluation metric that takes the *correctness level* of the answer into account. For example, for Q: "When did French revolutionaries storm the Bastille?" [18], A: "July 14, 1789" is probably more informative than A: "July 14" or A: "1789". For Q: "Where is Tokyo Disneyland?" A:"Chiba prefecture" is probably more useful than A: "Japan". However, currently there is no way to reflect these differences.

Our new metrics, which are applicable to QA evaluation with ranked lists of exact answers, are designed to solve the above two problems. That is, we aim at integrating "single-answer" and List tasks to some extent *and* incorporating answer correctness levels.

We are aware that Reciprocal Rank was abandoned at TREC 2002 with the requirement that the system must return *exactly one answer* (i.e. $L = L' = 1$) for the Main Task, and that CLEF 2004 is also following this move. However, we believe that evaluating ranked lists for QA is still important for the following reasons:

1. Returning a single exact answer is not the only possibility in practical QA systems. That is, a small ranked list of possible answers may be perfectly acceptable for some applications, e.g., when *answer recall* is considered to be important.

2. From a statistical viewpoint, evaluation based on single answers may not be reliable, as this is like measuring document retrieval performance by examining the document at Rank 1 only. Thus, a very good system that unluckily returned a correct answer at Rank 2 for *all* questions would be judged as "complete rubbish". To circumvent this danger, a large question set is often used, which can be burdensome for test collection constructors.

3. There appears to be some room for improvement in QA evaluation with $L' = 1$. The aim of introducing *Confidence Weighted Score* (CWS) at TREC 2002 was to measure a system's ability to recognise when it has found a correct answer to a given question [20]. However, it is clear from its definition that CWS only measures the system's ability to determine whether it is more confident about one question than another in a given question set: that is, it only measures *relative* confidence. Moreover, the idea of ranking questions in CWS may be counter-intuitive in some cases: for example, two TREC 2002 systems had nearly identical CWS values even though one system answered 28 more questions than the other one [20].

The remainder of this paper is organised as follows. Section 2 describes some existing IR metrics that are akin to our proposed metrics, and Section 3 proposes these metrics, namely, Q-measure and R-measure, as well as how to apply them to QA evaluation. Section 4 describes an experiment using the QAC1 QA test collection to discuss the advantages of Q-measure over Reciprocal Rank. Section 5 discusses extensions and limitations of the present work, and Section 6 concludes this paper. **Appendix A** contains some mathematical proofs concerning Q-measure and R-measure. In addition, **Appendix B** describes a set of experiments that demonstrate the practicality and reliability of Q-measure and R-measure as *IR* metrics, by actually ranking the systems submitted to the NTCIR-3 CLIR tasks [17]. As future work, we would like to conduct similar system ranking experiments for the NTCIR *QAC* tasks as well, if the QAC submission files become available for research purposes.

2 IR Metrics akin to Q-measure

2.1 Average Precision

Average Precision (e.g. [2]) is one of the most widely-used IR metric, although it cannot handle multigrade relevance. Let R denote the total number of known relevant documents for a particular search request (or a *topic*), and let $count(r)$ denote the number of relevant documents within the top r documents of the ranked output. Clearly, the Precision at Rank r is $count(r)/r$. Let $isrel(r)$ denote a binary flag, such that $isrel(r) = 1$ if the document at Rank r is relevant and $isrel(r) = 0$ otherwise. Then, Average Precision (AveP) is defined as:

$$AveP = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{count(r)}{r} \qquad (1)$$

where L is the ranked output size.

Another useful measure often used along with AveP is *R-Precision*, although this is not as sensitive as AveP to the changes in the top ranks:

$$R\text{-}Precision = \frac{count(R)}{R} \qquad (2)$$

These measures are known to "average well" across a set of topics, in contrast to metrics that are based on *fixed* document ranks (See Section 2.2).

2.2 Cumulative Gain

Järvelin and Kekäläinen proposed (Discounted) Cumulative Gain for evaluation based on multigrade relevance [6]. Their basic idea is that a system output, scanned from the top, receives a score for each retrieved relevant document. The score for retrieving a highly relevant document is high, and that for retrieving a partially relevant one is low.

Formally, let X denote a relevance level, and let $gain(X)$ denote the *gain value* for successfully retrieving an X-relevant document. For the NTCIR CLIR test collections, $X \in \{S, A, B\}$ [8], and a typical gain value assignment would be $gain(S) = 3, gain(A) = 2, gain(B) = 1$. Hereafter, we use the above NTCIR relevance levels and the gain value assignment by default without loss of generality. 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 nonrelevant. The *cumulative gain at Rank* r is given by $cg(r) = g(r) + cg(r-1)$ for $r > 1$ and $cg(1) = g(1)$. (In fact, a concept that is equivalent to cumulative gain already existed in the 1960s, when Pollack proposed the *sliding ratio* measure [10].)

In [6], Järvelin and Kekäläinen used the Cumulative Gain by averaging $cg(r)$ across a given topic set for each r, from the viewpoint of *how many documents the user has to go through*. However, as Kando *et al.* [8] and Sakai [12] have pointed out, this is not desirable from a *statistical* viewpoint, as the number of relevant documents (R) differs across the search request set, and therefore the upperbound performance at a fixed rank differs across the set as well. This also applies to Precision at a fixed document rank. For example, consider a ranked output with three nonrelevant documents and two B-relevant documents at the very top, such that its *gain sequence* is $(g(1), g(2), \ldots) = (0, 0, 0, 1, 1)$, so that its *cumulative gain sequence* is $(cg(1), cg(2), \ldots) = (0, 0, 0, \mathbf{1}, \mathbf{2})$. (Here, $cg(r)$ values are shown in **boldface** whenever $g(r) > 0$.) Let $R(X)$ denote the number of known X-relevant documents so that $\sum_X R(X) = R$, and suppose that such a ranked output was returned for both Topic One with $R = R(B) = 2$, and for Topic Two with $R = R(B) = 100$. Then, for *both* of these topics, the Precision at Rank 5 is 2/5=0.4 and the Cumulative Gain at Rank 5 is $cg(5) = 2$. However, these values clearly represent the *best possible* performance at Rank 5 for Topic One, while they are far from it for Topic Two.

More recently, Järvelin and Kekäläinen have proposed *normalised* versions of their cumulative gain metrics [7]. They will be discussed in Section 5.2.

2.3 Average Weighted Precision

Average Weighted Precision (AWP) proposed by Kando *et al.* [8] is based on Cumulative Gain, but is arguably more statistically reliable as it performs comparison with an *ideal ranked output* [6] before averaging across topics. (An ideal ranked output for NTCIR can be obtained by listing up all S-relevant documents, then all A-relevant documents, then all B-relevant documents.) Let $cig(r)$ represent the cumulative gain at Rank r for an ideal ranked output. AWP is given by:

$$AWP = \frac{1}{R} \sum_{1 \leq r \leq L, g(r) > 0} \frac{cg(r)}{cig(r)}$$

$$= \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cg(r)}{cig(r)} \quad (3)$$

Kando *et al.* [8] have also proposed *R-Weighted Precision*:

$$R\text{-}WP = \frac{cg(R)}{cig(R)} \quad (4)$$

If the relevance assessments are *binary* so that each relevant document gives a gain value of 1, then, by definition, both

$$cg(r) = count(r) \quad (5)$$

and

$$cig(r) = r \quad (6)$$

hold for $r \leq R$. Thus, with binary relevance,

$$cg(r)/cig(r) = count(r)/r \quad (7)$$

holds for $r \leq R$. (Thus, $cg(r)/cig(r)$ is a kind of "weighted precision", and from Equation (3), we prefer the name Average Weighted Precision to Weighted Average Precision.) Therefore, From Equations (1) and (3), if the relevance assessments are binary *and* if the system output does not have any relevant documents below Rank R, then

$$AveP = AWP \quad (8)$$

holds. Similarly, from Equations (2) and (4), with binary relevance,

$$R\text{-}Precision = R\text{-}WP \quad (9)$$

holds.

Although AWP appears to be a natural extension of AveP, it suffers from a serious problem. Since there are no more than R relevant documents,

$$cig(r) = cig(R) \quad (10)$$

holds for $r > R$. That is, after Rank R, $cig(r)$ becomes a *constant*, which implies, from Equation (3), that AWP cannot distinguish between System A that has a relevant document at Rank R and System B that has a relevant document at Rank L (i.e. at the very bottom of the ranked list). For example, suppose that $R = R(B) = 5$ for a topic. Given that $gain(B) = 1$, the sequence of $cig(r)$ is clearly $(1, 2, 3, 4, 5, 5, , \ldots)$. Now, suppose that both System A and System B retrieved only one relevant document, but that System A has it at Rank 5 and that System B has it at Rank 1000. Then, for System A, the sequence of $cg(r)$ is $(0, 0, 0, 0, \mathbf{1}, 1, \ldots)$ and $AWP =$

$(cg(5)/cig(5))/5 = (\mathbf{1}/5)/5 = 0.04$. For System B, the sequence of $cg(r)$ is $(0, 0, 0, 0, 0, \ldots, \mathbf{1})$ and $AWP = (cg(1000)/cig(1000))/5 = (\mathbf{1}/5)/5 = 0.04$. Thus the two systems would be considered to be identical in performance.

In short, AWP is not a reliable metric because its denominator $cig(r)$ "freezes" after Rank r. In contrast, AveP is free from this problem because its denominator r is guaranteed to increase steadily. R-Precision and R-WP are also free from the problem because they only look at the top R documents.

3 Proposed Metrics

3.1 Q-measure and R-measure

We now propose Q-measure and R-measure to solve the above problem of AWP.

First, we introduce the notion of *bonused gain at Rank* r, simply given by $bg(r) = g(r) + 1$ if $g(r) > 0$ and $bg(r) = 0$ if $g(r) = 0$. Then, the *cumulative bonused gain at Rank* r is given by $cbg(r) = bg(r) + cbg(r-1)$ for $r > 1$ and $cbg(1) = bg(1)$. That is, the system receives an *extra reward* for each retrieved relevant document. Q-measure is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \le r \le L, g(r) > 0} \frac{cbg(r)}{cig(r) + r}$$

$$= \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{cbg(r)}{cig(r) + r} \quad (11)$$

Note that the denominator in the above equation ($cig(r) + r$) is guaranteed not to "freeze", so that relevant documents found below Rank R can be handled properly. For the example given in Section 2.3, for System A, the sequence of $cbg(r)$ is $(0, 0, 0, 0, \mathbf{2}, 2, \ldots)$ and $Q\text{-}measure = (cbg(5)/(cig(5) + 5))/5 = (\mathbf{2}/(5 + 5))/5 = 0.04$. But for System B, the sequence of $cbg(r)$ is $(0, 0, 0, 0, 0, \ldots, \mathbf{2})$ and $Q\text{-}measure = (cbg(1000)/(cig(1000) + 1000))/5 = (\mathbf{2}/(5 + 1000))/5 = 0.0004$. (Here, $cbg(r)$ values are also shown in **boldface** whenever $g(r) > 0$.)

As the proofs in **Appendix A** show, Q-measure is equal to one iff a system output (s.t. $L \ge R$) is an ideal one. In contrast, both R-measure and R-WP are equal to one iff all the top R documents are (at least partially) relevant. Thus, for example, B-relevant documents may be ranked above the A-relevant ones. In this respect, Q-measure is clearly superior to R-measure.

By definition of the cumulative bonused gain,

$$cbg(r) = cg(r) + count(r) \quad (12)$$

holds for $r \ge 1$. Therefore, Q-measure and R-measure can alternatively be expressed as:

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{cg(r) + count(r)}{cig(r) + r} \quad (13)$$

$$R\text{-}measure = \frac{cg(R) + count(R)}{cig(R) + R} \quad (14)$$

By comparing Equation (13) with Equations (1) and (3), and Equation (14) with Equations (2) and (4), it can be observed that Q-measure and R-measure are "blended" metrics: Q-measure inherits the properties of both AWP and AveP, and R-measure inherits the properties of both R-WP and R-Precision. Moreover, it is clear that using large gain values would emphasise the AWP aspect of Q-measure, while using small gain values would emphasise its AveP aspect. Similarly, using large gain values would emphasize the R-WP aspect of R-measure, while using small gain values would emphasise its R-Precision aspect. For example, letting $gain(S) = 30$, $gain(A) = 20$, and $gain(B) = 10$ (or conversely $gain(S) = 0.3$, $gain(A) = 0.2$, and $gain(B) = 0.1$) instead of $gain(S) = 3$, $gain(A) = 2$, and $gain(B) = 1$ is equivalent to using the following generalised equations and letting $\beta = 10$ (or conversely $\beta = 0.1$):

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \le r \le L} isrel(r) \frac{\beta cg(r) + count(r)}{\beta cig(r) + r} \quad (15)$$

$$R\text{-}measure = \frac{\beta cg(R) + count(R)}{\beta cig(R) + R} \quad (16)$$

If the relevance assessments are binary, then, from Equations (5) and (6),

$$\frac{cg(r) + count(r)}{cig(r) + r} = \frac{2count(r)}{2r} = \frac{count(r)}{r} \quad (17)$$

holds for $r \le R$. Therefore, From Equations (1) and (13), if the relevance assessments are binary *and* if the system output does not have any relevant documents below Rank R, then Equation (8) can be generalised as:

$$AveP = AWP = Q\text{-}measure \quad (18)$$

Similarly, from Equations (2) and (14), with binary relevance, Equation (9) can be generalised as:

$$R\text{-}Precision = R\text{-}WP = R\text{-}measure \quad (19)$$

Appendix B describes a set of experiments for demonstrating the reliability of Q-measure and R-measure as *IR* metrics [17]. However, we now focus on the application of Q-measure to *QA* evaluation.

3.2 Application to Question Answering

This section describes how to apply Q-measure (and R-measure) to QA evaluation involving ranked lists of exact answers. The difficuly of QA evaluation lies in the fact that arbitrary answer strings need to be evaluated, in contrast to an IR situation in which only a closed-class, unique document IDs need to be evaluated. To overcome this problem (at least partially; See Section 5.3), we propose to provide equivalence classes of answers, or *answer synsets*, at the time of QA test collection construction. Using answer synsets, we can handle both "single-answer" and List questions in an "answer ranking" task, and can avoid rewarding systems that return duplicate answers that mean the same thing. *In fact, the QAC1 and QAC2 answer files already include equivalence class data for computing F-measure based on answer instances, so this is not a problem*. Our second proposal is to assign a *correctness level* to each answer string within each answer synset, so that we can distinguish between "good" answers and the "bad" (but somewhat correct) ones.

Let $AS(i)$ $(1 \leq i \leq R)$ denote an answer synset, and let $a(i,j)$ denote the j-th answer string in $AS(i)$. Let $x(i,j)$ denote the correctness level of $a(i,j)$, and let $xmax(i) = max_j x(i,j)$. That is, $xmax(i)$ is the *highest correctness level* within $AS(i)$. Then we define $R(X)$ as the number of answer synsets such that $xmax(i) = X$. Thus, if we extend the NTCIR document relevance levels to answer correctness levels, $R(S) + R(A) + R(B) = R$.

Below, we show some examples of how to prepare QA test collections in this way. (More examples, based on the actual QAC2 answer file, are provided in [15].)

Example 1:
Q: "Who played in the Beatles?" ($R = R(S) = 4$)
$AS(1) = \{$<"Sir Paul McCartney", $S >$, <"Paul McCartney", $S >$, <"McCartney", $A >$, <"Paul", $B >\}$
$AS(2) = \{$<"John Lennon", $S >$, <"Lennon", $A >$, <"John", $B >\}$
$AS(3) = \{$<"George Harrison", $S >$, <"Harrison", $A >$, <"George", B>$\}$
$AS(4) = \{$<"Ringo Starr", $S >$, <"Starr", $A >$, <"Ringo", $B >\}$
Some test collection constructors may prefer to add more answer synsets with relatively low correctness levels, representing early/temporary members of the Beatles, such as:
$AS(5) = \{$<"Stuart Sutcliffe", $B >$, <"Sutcliffe", $B >$, <"Stuart", $B >\}$
If the fifth answer synset is added, then $R(B) = 1$ and therefore $R = R(S) + R(B) = 5$.

Example 2:
Q: "What does DVD stand for?" ($R = R(S) = 1$)
$AS(1) = \{$<"Digital Versatile Disk", $S >$,<"Digital Video Disk", $A >\}$
If a system that returns *both* of the above answer strings is preferrable, then the above data should be broken into two separate answer synsets.

Example 3:
Q: "What is love?" ($R = R(A) = 1$)
$AS(1) = \{$<"NIL", $A >\}$
The answer data for NIL questions should be prepared as above. The correctness level of the NIL answer does not affect the QA performance, as we shall see later.

For some questions, answer correctness judgement may be more difficult than traditional document relevance judgement. However, for such questions, assigning "flat" correctness levels would suffice (e.g. treat all answer strings as A-relevant). That is, *distinction among answer strings by answer correctness levels is not mandatory*.

Figures 1 and 2 show an example of how to implement Q-measure and R-measure calculation for QA. Firstly, the algorithm in Figure 1 reads a ranked list of answers and marks the correct ones with S, A or B, but avoids marking duplicate answers from the same answer synset. Then, the algorithm in Figure 2 reads the above *marked* answers to calculate Q-measure and R-measure. Moreover, Figure 1 includes a special treatment of NIL answers: only a NIL answer at Rank 1 is marked as correct, in contrast to the TREC 2001 evaluation in which systems could be rewarded for including "NIL" somewhere in the ranked list [19].

Let us return to *Example 1* (without the fifth answer synset), and suppose that the system output was ("McCartney", "Lennon", "Paul", "George Harrison", "Starr"). Then, $(g(1), g(2), \ldots) = (2, 2, 0, 3, 2)$, and $(bg(1), bg(2), \ldots) = (3, 3, 0, 4, 3)$. Hence $(cbg(1), cbg(2), \ldots) = (\mathbf{3}, \mathbf{6}, 6, \mathbf{10}, \mathbf{13})$. Whereas, an example ideal ranked output for this question is ("Paul McCartney", "John Lennon", "George Harrison", "Ringo Starr"), so that $(cig(1), cig(2), \ldots) = (3, 6, 9, 12, 12, \ldots)$. Therefore, $Q\text{-}measure = (\mathbf{3}/(3+1) + \mathbf{6}/(6+2) + \mathbf{10}/(12+4) + \mathbf{13}/(12+5))/4 = 0.722$, and $R\text{-}measure = \mathbf{10}/(12+4) = 0.625$.

For *Example 3* (where "NIL" is regarded as A-correct), if the system correctly returns "NIL" at Rank 1, then $(g(1), g(2), \ldots) = (2, 0, \ldots)$, $(bg(1), bg(2), \ldots) = (3, 0, \ldots)$, and $(cbg(1), cbg(2), \ldots) = (\mathbf{3}, 3, \ldots)$. Whereas, $(cig(1), cig(2), \ldots) = (2, 2, \ldots)$. Thus, $Q\text{-}measure = R\text{-}measure = \mathbf{3}/(2+1) = 1$. In general, if the answer at Rank 1 is correct, then both $cbg(1) = cg(1) + 1$ and $cig(1) = cg(1)$ hold. Hence $cbg(1)/(cig(1) + 1) = 1$. Therefore, the NIL answer at Rank 1 would receive a Q/R-measure of 1.0 regardless of whether it is treated as S-, A- or B-correct.

```
/* initialize flag for each answer synset.
The flags avoid marking multiple answers from
the same answer synset. */
for( i=1; i<=R; i++ ) flag[i]=0;

r=1; /* system output rank */
while read o(r){ /* system's r-th answer */
  if( there exists a(i,j) s.t. o(r)==a(i,j) ){
  /* o(r) matches with a correct answer */
    if( o(r)=="NIL" ){
    /* special treatment of NIL */
      if( r==1 ){ /* i.e. NIL at Rank 1 */
        print o(r), x(i,j);
        /* marked as correct */
      }
      else{
        print o(r);
        /* NOT marked as correct */
      }
    }
    else{ /* not NIL */
      if( flag[i]==0 ){
      /* AS(i) is a NEW answer synset */
        print o(r), x(i,j);
        /* marked as correct */
        flag[i]=1;
      }
      else{ /* i.e. flag[i]==1 */
        print o(r);
        /* duplicate answer from AS(i)
        NOT marked as correct */
      }
    }
  }
  else{ /* no match with a correct answer */
    print o(r);
    /* NOT marked as correct */
  }
  r++; /* examine next rank */
}
```

Figure 1. Algorithm for marking a system output.

```
rmax=max(L,R); /* L: system output size */
               /* R: #answer synsets */

/* obtain cumulative gains for the
IDEAL ranked output */
r=0; cig[0]=0;
for each X in (S,A,B) { /* X: correctness level */
  for( k=1; k<=R(X); k++ ){
  /* R(X): #answer synsets in which the
     highest correctness level is X. */
    r++;
    cig[r]=cig[r-1]+gain(X);
  }
}
for( r=R+1; r<=rmax; r++ ){ /* in case L>R */
  cig[r]=cig[R];
}

/* obtain cumulative bonused gains for
the system output */
r=0; cbg[0]=0;
for( r=1; r<=L; r++ ){
  if( o(r) is marked with X ){
    cbg[r]=cbg[r-1]+gain(X)+1;
  }
  else{
    cbg[r]=cbg[r-1];
  }
}
for( r=L+1; r<=rmax; r++ ){ /* in case L<R */
  cbg[r]=cbg[L];
}

/* calculation */
sum=0;
for( r=1; r<=L; r++ ){
  if( cbg[r]>cbg[r-1] ){
  /* i.e. correct answer at Rank r */
    sum+=cbg[r]/(cig[r]+r);
  }
}
Q-measure=sum/R;
R-measure=cbg[R]/(cig[R]+R);
```

Figure 2. Algorithm for calculating Q-measure/R-measure.

4 Experiments

This section discusses the advantages of Q-measure over Reciprocal Rank through an experiment using the QAC1 test collection.

4.1 Extended QAC1 Collection

To use Q-measure and R-measure with the QAC1 test colleciton, the author manually converted the "flat" QAC1 answer data into answer synsets (based on the equivalence class information already provided in the answer file), and assigned a correctness level to each answer string. Although we could not hire a second judge for enhancing the reliability of the new answer data, here we assume that inter-judge differences do not affect comparative evaluation [18]. Strictly speaking, however, whether inter-judge differences in defining answer synsets and multigrade relevance affect evaluation is an open question. More importantly, the *reusability* of the QAC1 test collection has never been guaranteed: it is known that QA test collections are inherently less reusable than IR test collections [18]. We would like to study the impact of the inter-judge differences on system ranking if the actual QAC submission files (or *runs*) become available to us for research purposes.

We were able to add answer synsets and correctness levels to the original QAC1 answer data without any major problems. We have also extended the *QAC2* answer data in the same way in [15], and have constructed our own QA test collections with answer synsets and correctness levels in [16]. Based on our experience, we believe that constructing answer synsets and assigning correctness levels in QA test collection construction is feasible. Note that the number of answer strings used in short, exact-answer QA evaluation is generally much smaller than the number of relevant documents used in IR evaluation. We argue that, if assigning multigrade relevance for IR is feasible, so is our QA evaluation methodology. Recall also that "flat" answer correctness levels can be used whenever it is difficult to judge whether one answer string is better than another.

Table 1 (a) shows the distribution of the number of answer synsets for the Extended QAC1 data: it can be

observed that there is only one answer synset (i.e. $R = 1$) for 161 questions. Thus, R-measure is probably too demanding for this test collection as it only evaluates top R answers. Note also that Kando's AWP is clearly not suitable for QA evaluation: From Equation (10), $R = 1$ implies that $cig(r)$ remains constant for all r. Therefore, System A that returns the correct answer at Rank 1 and System B that returns the same answer at Rank 5 would receive the same score.

The outlier with $R = 18$ in Table 1 (a) is a very ambiguous List question: QAC1-1097 ("What are the Three Sacred Treasures?"). Although the phrase "Three Sacred Treasures" originally refer to specific historic items that symbolise the Imperial Throne, it is often used in newspaper contexts such as "Three Sacred Treasures of the Modern Era". Thus, consumer products such as "color TV" and "refrigerator" are included in the original answer set. Ideally, such outlier questions should be discarded from the evaluation set, because, if the system output size L is smaller than R, it is impossible to achieve a Q-measure of 1. However, here we follow the TREC/NTCIR traditions and let $L \leq L' = 5$.

Table 1 (b) shows the distribution of correctness levels of the QAC1 answer strings. As there are 282 answer *synsets* in total, each answer synset contains 616/282=2.18 answer strings on average.

As *supporting documents* [18, 19, 20] were not evaluated at QAC1, our Extended QAC1 data are based on answer strings rather than answer-document pairs. Thus our evaluation is *lenient* in TREC parlance.

Table 1. Distribution of R and correctness levels for the 195 QAC1 questions.

(a)	
R	#questions
1	161
2	14
3	12
4	4
5	1
9	2
18	1
total	195

(b)	
correctness level	#answer strings
S	401
A	118
B	97
total	616

Table 2. Performance of ASKMi for the 195 QAC1 questions.

	RR	Q-measure	R-measure
AF	0.682	0.684	0.543
	4↓26↑	5↓35↑	4↓20↑
noAF	0.637	0.639	0.469

4.2 ASKMi Japanese QA System

ASKMi, the Japanese QA system used in the present experiments, is described fully in [13, 15]. It suffices to treat it as a "black box" for the purpose of this study. To illustrate the advantages of Q-measure over Reciprocal Rank, this paper examines two ASKMi runs, namely, those with and without the *Answer Formulator* module. The primary function of the Answer Formulator is *answer string consolidation*: For example, if the original ranked list of answers contains "*Koizumi shushō* (prime minister Koizumi)" at Rank 1 and "*Koizumi*" at Rank 4, the answer formulator tries to erase the latter to minimise redundancy. Note that the above operation would move the answer at Rank 6 (i.e. out of the answer list) to Rank 5.

4.3 Results and Discussions

Table 2 summarises the performance of ASKMi for the 195 non-NIL questions from QAC1. (Currently, ASKMi cannot detect NIL questions.) The runs with and without the Answer Formulator are represented by **AF** and **noAF**, respectively. The table also shows question-by-question comparisons: for example, in terms of Reciprocal Rank, **AF** outperforms **noAF** for 26 questions while **noAF** outperforms **AF** for 4 questions. While these differences are statistically significant with the Sign Test for all three metrics, it can be observed that Q-measure is more sensitive than Reciprocal Rank: while Q-measure detected a performance difference for 40 questions (5 down and 35 up), Reciprocal Rank detected a performance difference for only 30 questions (4 down and 26 up). R-measure, on the other hand, appears to be less sensitive to the effect of Answer Formulator, as R is generally very small for the QAC1 questions.

Although the *Mean* Reciprocal Rank and the *Mean* Q-measure values are very similar for this test collection, individual values are in fact quite different. Among the 195 questions, there were 23 questions for which the **AF** performance was 1.0 in terms of Reciprocal Rank *and* less than one in terms of Q-measure. This happens when a system returns a (somewhat) correct answer at Rank 1 *and*: (a) the above answer is not the *best* answer; or (b) there is at least one more answer synset and the system did not handle it well. An example of (a) is QAC1-1012 "When did Yasunari Kawabata become the first Japanese to receive the Nobel Prize in Literature?". **AF**'s first response for this question was "1968", which was only B-correct. Thus, $cbg(1) = g(1) + 1 = \mathbf{2}$. There was only one answer synset for this question, which included "December 10, 1968" as an S-correct answer. Thus, $R = R(S) = 1$, and $cig(1) = 3$. Therefore, $Q\text{-}measure = (\mathbf{2}/(3+1))/1 = 0.5$. An example of (b) is QAC1-1058 "Japanese who received the Nobel Prize in Physics". The **AF** run returned "Hideki Yukawa" at Rank 1 and "Shinichiro Tomonaga" at Rank 5, both

of which are S-correct. Thus, the bonused gain sequence is $(4, 0, 0, 0, 4)$ and the cumulative bonused gain sequence is $(\mathbf{4}, 4, 4, 4, \mathbf{8})$. There are three answer synsets (i.e., three researchers) and $R = R(S) = 3$ for this question. Thus, $(cig(1), cig(2), cig(3), \ldots) = (3, 6, 9, 9, 9, \ldots)$. Therefore, $Q\text{-}measure = (\mathbf{4}/(3 + 1) + \mathbf{8}/(9 + 5))/3 = 0.524$. Note that Reciprocal Rank *ignores* the correct answer at Rank 5, and would have fully accepted "incomplete" answers such as "Yukawa".

Figure 3 visualises the performance differences between **AF** and **noAF** in terms of each evaluation metric for the first one-third of the QAC1 questions. Thus, dots above and below zero represent the positive and negative effects of the Answer Formulator, respectively, and they correspond to the "arrows" in Table 2. Although the Answer Formulator can occasionally hurt performance, some of the seemingly negative effects are because of the *reusability* problem mentioned in Section 4.1. For example, the only "negative dot" in Figure 3 represents QAC1-1021 "How was Prime Minister Obuchi criticized just after inauguration?": The **noAF** run returned "ordinary man" at Rank 1 and "cold pizza" at Rank 2, both of which were S-correct. However, the Answer Formulator replaced "cold pizza" with "Obuchi is as uninspiring as cold pizza", as it judged the longer answer to be more informative. Unfortunately, the longer answer string was beyond the scope of QAC1 and was *not* listed as a correct answer. Hence **AF** received a lower score.

Let us go back to the discussion of the *sensitiveness* of QA metrics in terms of comparison between **AF** and **noAF**. For QAC1-1013, 1021, 1037, 1056 and 1058 in Figure 3, the difference in terms of Reciprocal Rank is zero while that in terms of Q-measure is not. That is, for these questions, Q-measure detected the effect of the Answer Formulator which Reciprocal Rank overlooked. For QAC1-1058 mentioned earlier in this section, the **noAF** run failed to return the second correct answer "Shinichiro Tomonaga", as its answer list contained *duplicates*, namely, "Hideki Yukawa" at Rank 1 and "Doctor Hideki Yukawa" at Rank 4. Thus, after answer string consolidation, "Shinichiro Tomonaga" rose to Rank 5 and received credit in terms of Q-measure. Also, we have examined QAC1-1021 already. These examples show that Q-measure not only handles both "single-answer" and List questions properly but also evaluates the system's power to minimise redundancy in the answer list.

R-measure is also more sensitive than Reciprocal Rank for QAC1-1021, 1037 and 1056 in Figure 3: for these questions, the R-measure values were actually equal to the Q-measure ones. However, as mentioned earlier, R-measure can be insensitive to changes in the ranked list for questions with small R. For example, for QAC1-1006 "When will NTT Communications take over NTT International Network?" included in Figure 3, the difference in R-measure is zero while those in Reciprocal Rank and Q-measure are 0.083 and 0.096, respectively. Although the Answer Formulator managed to move the correct answer "October 1" from Rank 4 to Rank 3 by erasing "October" (treated as incorrect in the QAC1 data) which was originally at Rank 3, R-measure did not detect this improvement because, for this question, $R = R(S) = 1$. Probably, R-measure is more suitable for IR than for QA.

5 Extensions and Limitations

Sections 5.1 and 5.2 discuss possible extensions of Q-measure and R-measure as IR metrics. Section 5.3 discusses some unsolved issues in QA evaluation.

5.1 Average Gain Ratio for IR

Recently, Sakai [12] has proposed *Average Gain Ratio* (AGR) and *R-Gain Ratio* for IR evaluation based on multigrade relevance. These metrics are the same as Kando's AWP and R-WP, respectively, except that they use *topic adjusted* gain values instead of *fixed* gain values such as $gain(S) = 3, gain(A) = 2, gain(B) = 1$. Thus, Sakai proposes to perform the following transformation for each topic:

$$gain'(X) = gain(X) - \frac{R(X)}{R}(gain(X) - gain(X')) \quad (20)$$

where X' is the relevance level that is one level lower than X. (If X is the lowest relevance level, then $gain(X')$ is taken to be zero. Moreover, the above transformation is not applied if $R(X) = R$.) The above transformation was proposed based on the observation that the ratio $R(S) : R(A) : R(B)$ differs considerably across topics for the NTCIR CLIR test collections. For example, $R(B) >> R(S)$ holds for many questions, but not for *all* questions.

Although AGR itself inherits the problem of AWP discussed in Section 3.1, Equation (20) can easily be applied to Q-measure and R-measure as well.

5.2 Discounted Gains for IR

As discussed in Section 3.1, Q-measure penalises "late arrival" of relevant documents by incorporating the AveP aspect into AWP. In contrast, Järvelin and Kekäläinen [6] have used *discounted* cumulative gains for the penalisation. While R-measure and R-WP can be equal to one for a suboptimal system output (See **Appendix A**), using *discounted* cumulative gains instead of the raw ones would alleviate this problem. However, discounting requires a parameter that must be given from outside, namely the logarithm base.

Järvelin and Kekäläinen [7] have recently proposed Average Normalised Cumulative Gain (ANCG),

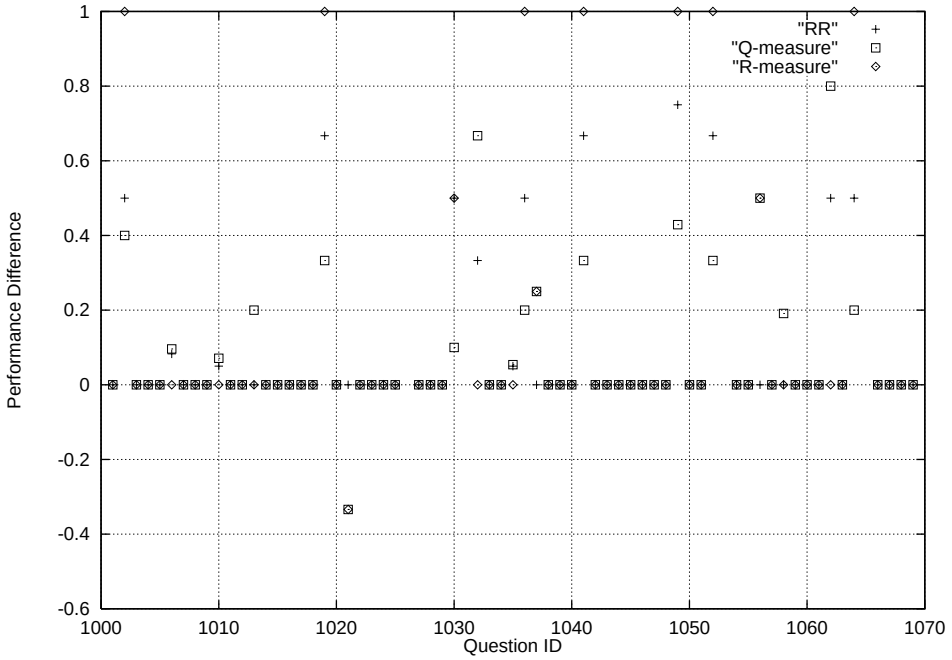

Figure 3. Performance difference (AF-noAF) for QAC1 (1001-1069).

which can be expressed as follows:

$$ANCG = \frac{1}{l} \sum_{1 \leq r \leq l} \frac{cg(r)}{cig(r)} \qquad (21)$$

where l is the number of documents to be examined, a parameter that must be given from outside. They have proposed Average Normalised *Discounted* Cumulative Gain (ANDCG) as well.

ANCG *appears* to resemble AWP (See Equation (3)), but is in fact a very different metric: ANCG is a *rank-based* (or *DCV-based* [5]) metric, while AWP and Q-measure are *recall-based* metrics just like Average Precision. It has been argued that *rank-based* metrics are *user-oriented*, and that *recall-based* metrics are *system-oriented* [5, 7]. As future work, we plan to compare Q-measure with ANCG and ANDCG by actually ranking the NTCIR systems as we have done in **Appendix B**.

5.3 Definition/Why/How Questions in QA

Clearly, our QA evaluation methodology cannot fully handle definition/why/how type questions as it is almost impossible to prepare *exhaustive* lists of such answers in advance. QAC1-1021 mentioned in Section 4.3 is an example of a similar problem. Although some automatic evaluation methods based on comparison with gold-standard texts have been proposed for Machine Translation, Summarisation and QA [1, 9], problems remain for QA: Suppose that the user asks "What is exothermic reaction?" and the system responds with "a chemical reaction accompanied by the absorption of heat". The correct answer is, however, "a chemical reaction accompanied by the *evolution* of heat". Using existing automatic evaluation metrics, the system would receive a high score despite the fact that it is telling a complete lie, as the two answer strings do share word N-grams and are identical in length [9]. These problems are beyond the scope of Q-measure and R-measure (and traditional QA measures such as Reciprocal Rank).

6 Conclusions

We have proposed Q-measure and R-measure, which are statistically reliable IR metrics for multi-grade relevance. Through an experiment using the QAC1 test collection, we also showed that Q-measure can handle both "single-answer" and List questions, as well as answer correctness levels, in QA evaluation with ranked lists of exact answers. As mentioned earlier, we would like to demonstrate the usefulness of Q-measure as a QA metric by actually ranking the systems submitted to the NTCIR QAC tasks. We therefore hope that the QAC submission files will soon become available to us.

Appendix A: Theorem Proofs.

Theorem 1 *Q-measure is equal to one iff the system output (s.t. $L \geq R$) is an ideal one.*

Proof: Given that the system output (s.t. $L \geq R$) is an ideal one, then both

$$cg(r) = cig(r) \tag{22}$$

and

$$count(r) = r \tag{23}$$

hold for $r \geq 1$. Therefore,

$$\begin{aligned} Q\text{-}measure &= \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{cig(r) + r}{cig(r) + r} \\ &= \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \\ &= \frac{1}{R} (\sum_{1 \leq r \leq R} isrel(r) + \sum_{R < r \leq L} isrel(r)) \end{aligned} \tag{24}$$

Now, since the system output is an ideal one, the top R documents are relevant and those below Rank R are nonrelevant. In other words, $isrel(r) = 1$ for $r \leq R$ and $isrel(r) = 0$ for $r > R$. Therefore,

$$Q\text{-}measure = \frac{1}{R} (\sum_{1 \leq r \leq R} +0) = \frac{1}{R} * R = 1. \tag{25}$$

Conversely, given that Q-measure is one, then

$$R = \sum_{1 \leq r \leq L} isrel(r) \frac{cg(r) + count(r)}{cig(r) + r} \tag{26}$$

holds. Now, since both

$$cg(r) \leq cig(r) \tag{27}$$

and

$$count(r) \leq r \tag{28}$$

hold for $r \geq 1$,

$$\frac{cg(r) + count(r)}{cig(r) + r} \leq \frac{cg(r) + r}{cig(r) + r} \leq 1 \tag{29}$$

holds for $r \geq 1$.

From Equations (26) and (29),

$$R \leq \sum_{1 \leq r \leq L} isrel(r) \frac{cg(r) + r}{cig(r) + r} \leq \sum_{1 \leq r \leq L} isrel(r) \tag{30}$$

holds. However, as there are no more than R relevant documents,

$$R \geq \sum_{1 \leq r \leq L} isrel(r) \tag{31}$$

should hold. Therefore, from Equations (30) and (31), both

$$R = \sum_{1 \leq r \leq L} isrel(r) \tag{32}$$

and

$$\sum_{1 \leq r \leq L} isrel(r) \frac{cg(r) + r}{cig(r) + r} = \sum_{1 \leq r \leq L} isrel(r) \tag{33}$$

hold. Equation (32) implies that the system output includes *all* relevant documents. Whereas, From Equations (29) and (33), it is necessary that $cg(r) = cig(r)$ for every r s.t. $isrel(r) = 1$. Therefore, the system output must be an ideal one.

Theorem 2 *R-measure is equal to one iff all the top R documents are (at least partially) relevant.*

Proof: Given that the top R documents are (at least partially) relevant, that is, *all* the relevant documents are listed at the top of the ranked list, then both

$$cg(R) = cig(R) = \sum_{X} R(X) gain(X) \tag{34}$$

and

$$count(R) = R \tag{35}$$

hold. (Recall that $R = \sum_X R(X)$, and that $X \in S, A, B$ for NTCIR.) Therefore,

$$R\text{-}measure = \frac{cig(R) + R}{cig(R) + R} = 1. \tag{36}$$

Convsersely, given that R-measure is one, then

$$cg(R) + count(R) = cig(R) + R \tag{37}$$

holds. From Equations (27), (28) and (37), both

$$cg(R) = cig(R) \tag{38}$$

and

$$count(R) = R \tag{39}$$

hold. Therefore, the top R documents are all relevant.

Appendix B: Ranking the NTCIR-3 CLIR Systems based on Q-measure.

This appendix shows the reliability of Q-measure and R-measure using the actual submitted runs from the NTCIR-3 CLIR task [17]. The following files, provided by National Institute of Informatics, Japan, were used for the analyses reported in this paper.

- `ntc3clir-allCruns.20040511.zip` (45 Runs for retrieving Chinese documents)
- `ntc3clir-allJruns.20040511.zip` (33 Runs for retrieving Japanese documents)
- `ntc3clir-allEruns.20040511.zip` (24 Runs for retrieving English documents)
- `ntc3clir-allKruns.20040511.zip` (14 Runs for retrieving Korean documents)

The above files contain runs submitted by 14 different participants, and include both monolingual and cross-language runs, as well as runs using different topic fields, e.g. TITLE, DESCRIPTION etc. (There were 23 participants at the NTCIR-3 CLIR Task, but not all of them have agreed to the release of their submission files.)

Tables 3-6 show the Spearman and Kendall Rank Correlations for Q-measure and its related metrics based on the NTCIR-4 CLIR C-runs, J-runs, E-runs, and K-runs, respectively. The correlation coefficients are equal to 1 when two rankings are identical, and are equal to -1 when two rankings are completely reversed. (It is known that the Spearman's coefficient is usually higher than the Kendall's.) Values higher than 0.99 (i.e. extremely high correlations) are indicated in **boldface**. "Relaxed" represents Relaxed Average Precision, "Rigid" represents Rigid Average Precision, and "Q-measure" and "AWP" use the *default* gain values: $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$. Moreover, the columns in Part (b) of each table represent Q-measure with different gain values: For example, "Q30:20:10" means Q-measure using $gain(S) = 30$, $gain(A) = 20$ and $gain(B) = 10$ (Recall Equation (15)). Thus, "Q1:1:1" implies binary relevance, and "Q10:5:1" implies stronger emphasis on highly relevant documents.

Figures 4-7 visualise the above tables, respectively, by sorting systems in decreasing order of *Relaxed Average Precision* and then renaming each system as System No. 1, System No. 2, and so on. Thus, the Relaxed Average Precision curves are guaranteed to decrease monotonically, and the other curves (representing system rankings based on other metrics) would also decrease monotonically only if their rankings agree perfectly with that of Relaxed Average Precision. That is, an increase in a curve represents a *swop*.

The above tables and figures are shown in order of decreasing reliability: Table 3/Figure 4 are based on 45 systems, while Table 6/Figure 7 are based on only 14 systems. Furthermore, Table 7 condenses Tables 3-6 into one by taking averages over the four sets of data.

From the above results regarding Q-measure, we can observe the following:

1. While it is theoretically clear that AWP is unreliable when relevant documents are retrieved below Rank R, our experimental results confirm this fact. The AWP curves include many swops, and some of them are represented by a very "steep" increase. This is because AWP overestimates a system's performance which rank many relevant documents below Rank R. For example, in Figure 4, System No. 4 outperforms System No. 3 according to AWP, even though all other metrics suggest the contrary.

2. Compared to AWP, the Q-measure curves are clearly more stable. Moreover, from Part (a) of each table, Q-measure is more highly correlated with Relaxed AveP than AWP is, *and* is more highly correlated with Rigid AveP than AWP is.

3. From Part (a) of each table, it can be observed that Q-measure is more highly correlated with *Relaxed* AveP than with *Rigid* AveP. (The same is true for AWP as well.) This is natural, as Rigid AveP ignores the B-relevant documents completely.

4. It can be observed that the behaviour of Q-measure is relatively stable with respect to the choice of gain values. Moreover, by comparing "Q30:20:10", "Q-measure" (i.e. Q3:2:1) and "Q0.3:0.2:0.1" in terms of correlations with "Relaxed", it can be observed that using smaller gain values implies more resemblance with Relaxed AveP (Recall Equation (15)). For example, in Table 3, the Spearman's correlation with "Relaxed" is 0.9909 for "Q30:20:10", 0.9982 for "Q-measure", and 0.9997 for "Q0.3:0.2:0.1". This property is also visible in the graphs: while each "Q30:20:10" curve resembles the corresponding AWP curve, each "Q0.3:0.2:0.1" curve is almost indistisguishable from the "Relaxed" curve.

5. From Part (b) of each table, it can observed that "Q1:1:1" (i.e. Q-measure with binary relevance) is very highly correlated with Relaxed AveP (Recall Equation (18)).

Tables 8-11 show the Spearman and Kendall Rank Correlations for R-measure and its related metrics based on the NTCIR-4 CLIR C-runs, J-runs, E-runs, and K-runs, respectively. Table 12 condenses Tables 8-11 into one by taking averages over the four sets of data. Again, "Q-measure", "R-measure" and "R-WP" use the default gain values, "R30:20:10" represents R-measure using $gain(S) = 30$, $gain(A) = 20$ and $gain(B) = 10$, and so on. As "R1:1:1" (R-measure with binary relevance) is identical to R-Precision (and R-WP), it is not included in the tables.

From the above results regarding R-measure, we can observe the following:

1. From Part (a) of each table, it can be observed that R-measure, R-WP and R-Precision are very highly correlated with one another. Moreover, R-measure is slightly more highly correlated with R-Precision than R-WP is.

2. From the tables, it can be observed that R-measure is relatively stable with respect to the choice of gain values. By comparing "R30:20:10", "R-measure" (i.e. R3:2:1) and "R0.3:0.2:0.1" in terms of correlations with R-Precision, it can be observed that using smaller

gain values implies more resemblance with R-Precision (Recall Equation (16)). For example, in Table 8, the Spearman's correlation with R-Precision is 0.9939 for "R30:20:10", 0.9960 for "R-measure", and 0.9982 for "R0.3:0.2:0.1".

Thus, our experiments show that Q-measure and R-measure are reliable IR performance metrics for evaluations based on multigrade relevance. However, recall that, while Q-measure is one iff the system output is an ideal one, R-measure (and R-WP) can be one for a suboptimal system output (See **Appendix A**). We therefore recommend the use of Q-measure as the primary IR metric based on multigrade relevance.

At the NTCIR-4 Open Submission Session, Q-measure and *Average Distance Measure* (ADM) [3] were separately proposed as official IR evaluation metrics for NTCIR. ADM requires *continuous* User Relevance Scores (URSs) and System Relevance Scores (SRSs) *that can be compared directly with the URSs*. However, the goal of most IR systems is to *rank* the documents (usually by means of calculating some kind of document scores) and *not* to estimate the relevance score of each document. That is, ADM defines a *different task altogether* (which itself is interesting, as such a task would make SRSs comparable across systems). Finding an appropriate function for transforming the document scores of a particular IR system into estimated relevance scores is a task that is probably more akin to document *filtering* than to IR [11]. Moreover, as ADM is simply based on the absolute difference between the pair of SRS and URS for each document, (a) Using *discrete* (i.e. multigrade) URSs as in NTCIR implies that an optimal IR system is also supposed to output discrete SRSs; and (b) Documents at low ranks are considered to be of equal importance as those at high ranks, which is a clear disadvantage compared to Average Precision and Q-measure for the purpose of traditional IR evaluation.

Acknowledgement

The author is indebted to the NTCIR-3 Organisers, most of all Noriko Kando, for making the NTCIR-3 CLIR data available to us for research purposes. I would also like to thank the NTCIR-3 CLIR participants who have agreed to the release of their submission files.

References

[1] Breck, E. J. *et al.*: How to Evaluate Your Question Answering System Every Day... and Still Get Real Work Done, *LREC 2000 Proceedings*, 2000.

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

[3] Della Mea, V. and Mizzaro, S.: Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation, *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 6, pp. 530-543, 2004.

[4] Fukumoto, J., Kato, T. and Masui, F.: Question Answering Challenge (QAC-1): An Evalution of Question Answering Tasks at the NTCIR Workshop 3, *AAAI Spring Symposium: New Directions in Question Answering*, pp. 122-133, 2003.

[5] Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *ACM SIGIR '93 Proceedings*, pp. 329-338, 1993.

[6] Järvelin, K. and Kekäläinen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents, *ACM SIGIR 2000 Proceedings*, pp. 41-48, 2000.

[7] 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.

[8] Kando, N., Kuriyama, K. and Yoshioka, M.: Information Retrieval System Evaluation using Multi-Grade Relevance Judgments - Discussion on Averageable Single-Numbered Measures (in Japanese), *IPSJ SIG Notes*, FI–63–12, pp. 105-112, 2001.

[9] Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics, *HLT-NAACL 2003 Proceedings*, 2003.

[10] Pollack, S. M.: Measures for the Comparison of Information Retrieval Systems, *American Documentation* Vol. 19, No. 4, pp. 387-397, 1968.

[11] Robertson, S. and Soboroff, I.: The TREC 2002 Filtering Track Report, *TREC 2002 Proceedings*, 2002.

[12] Sakai, T.: Average Gain Ratio: A Simple Retrieval Performance Measure for Evaluation with Multiple Relevance Levels, *ACM SIGIR 2003 Proceedings*, pp. 417–418, 2003.

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

[14] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/Bilingual IR and Flexible Feedback, *NTCIR-4 CLIR Proceedings*, to appear, 2004.

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

[16] Sakai, T. *et al.*: The Effect of Back-Formulating Questions in Question Answering Evaluation, *ACM SIGIR 2004 Proceedings*, pp. 474-475, 2004.

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

[18] Voorhees, E. M.: Building A Question Answering Test Collection, *ACM SIGIR 2000 Proceedings*, pp. 200–207, 2000.

[19] Voorhees, E. M.: Overview of the TREC 2001 Question Answering Track, *TREC 2001 Proceedings*, 2001.

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

Table 3. Spearman/Kendall Rank Correlations for the 45 C-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9874/.9273	**.9982**/.9798	.9802/.8990
Rigid	-	.9858/.9192	.9648/.8667
Q-measure	-	-	.9851/.9152
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9909**/.9374	**.9997**/**.9960**	**.9989**/.9879	**.9947**/.9556
Rigid	.9788/.8970	.9874/.9273	.9851/.9192	.9829/.9111
Q-measure	**.9901**/.9333	**.9978**/.9798	**.9984**/.9798	**.9955**/.9636

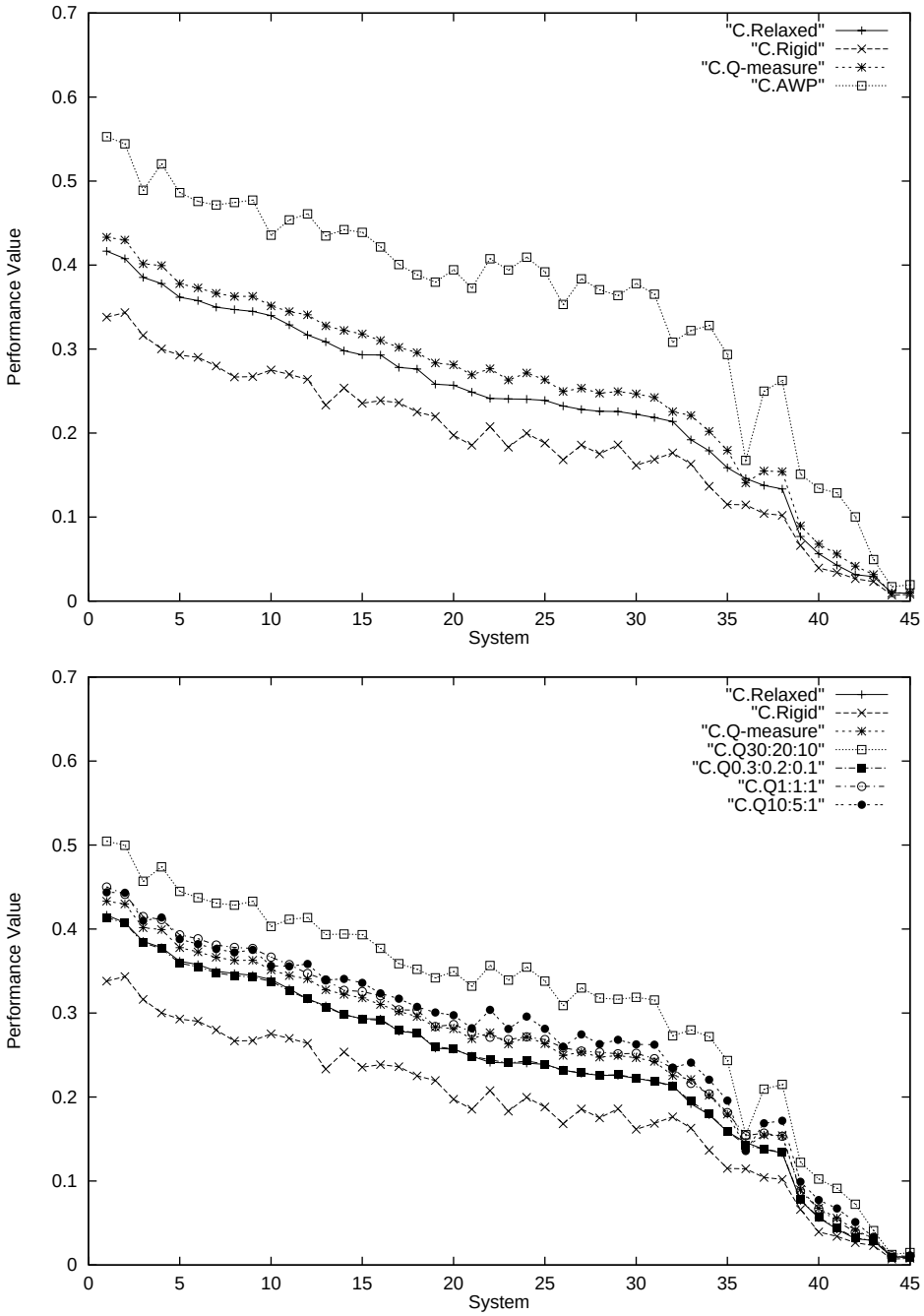

Figure 4. System ranking comparisons with Relaxed Average Precision (C-runs).

Table 4. Spearman/Kendall Rank Correlations for the 33 J-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9619/.8561	**.9947**/.9583	.9833/.9242
Rigid	-	.9616/.8447	.9505/.8182
Q-measure	-	-	.9813/.9129
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	.9769/.9015	**.9980**/.9811	**.9990**/.9886	.9759/.8977
Rigid	.9395/.7879	.9592/.8447	.9616/.8523	.9519/.8144
Q-measure	.9729/.8826	**.9943**/.9545	**.9943**/.9545	.9706/.8864

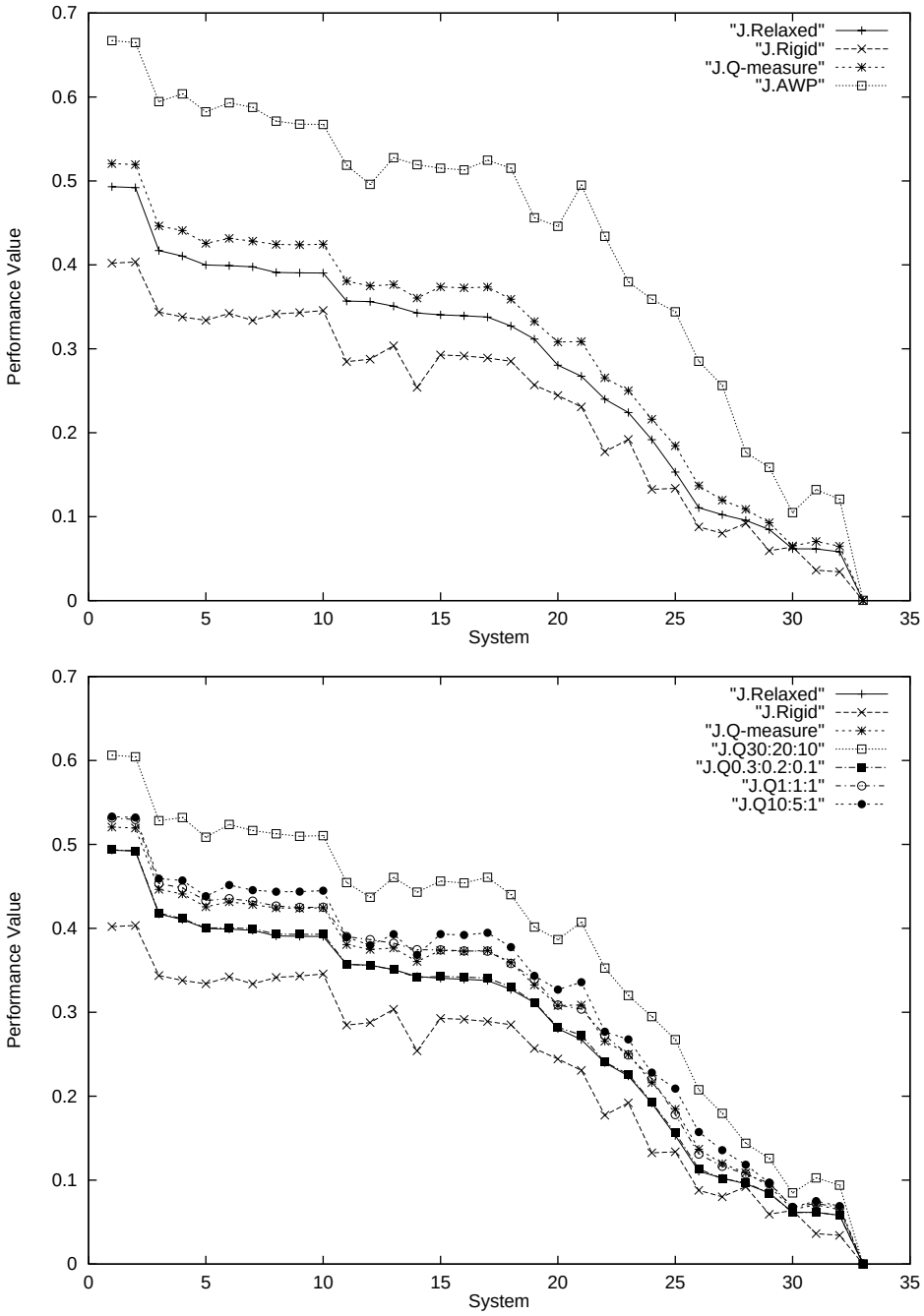

Figure 5. System ranking comparisons with Relaxed Average Precision (J-runs).

Table 5. Spearman/Kendall Rank Correlations for the 24 E-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	**.9922**/.9565	**.9974**/.9783	.9835/.9058
Rigid	-	**.9948**/.9638	.9748/.8913
Q-measure	-	-	.9843/.9130
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9922**/.9565	**1.000**/**1.000**	**.9965**/.9783	.9887/.9348
Rigid	.9852/.9275	**.9922**/.9565	**.9904**/.9493	.9887/.9348
Q-measure	**.9904**/.9493	**.9974**/.9783	**.9957**/.9710	.9887/.9420

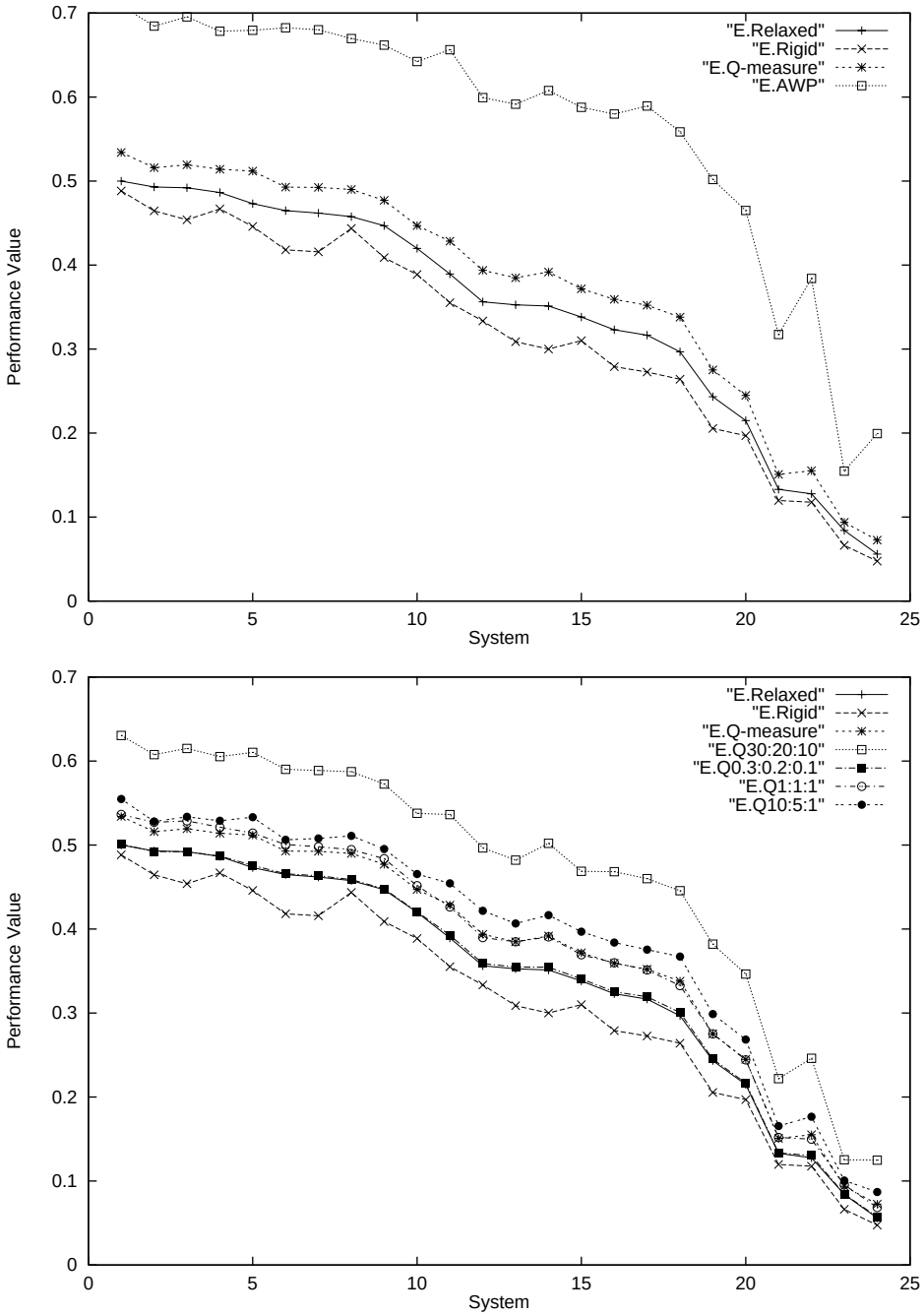

Figure 6. System ranking comparisons with Relaxed Average Precision (E-runs).

Table 6. Spearman/Kendall Rank Correlations for the 14 K-runs (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9560/.8462	**.9912**/.9560	**.9912**/.9560
Rigid	-	.9385/.8022	.9385/.8022
Q-measure	-	-	**1.000**/**1.000**
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	**.9912**/.9560	**.9956**/.9780	**1.000**/**1.000**	**.9912**/.9560
Rigid	.9385/.8022	.9516/.8242	.9560/.8462	.9385/.8022
Q-measure	**1.000**/**1.000**	**.9956**/.9780	**.9912**/.9560	**1.000**/**1.000**

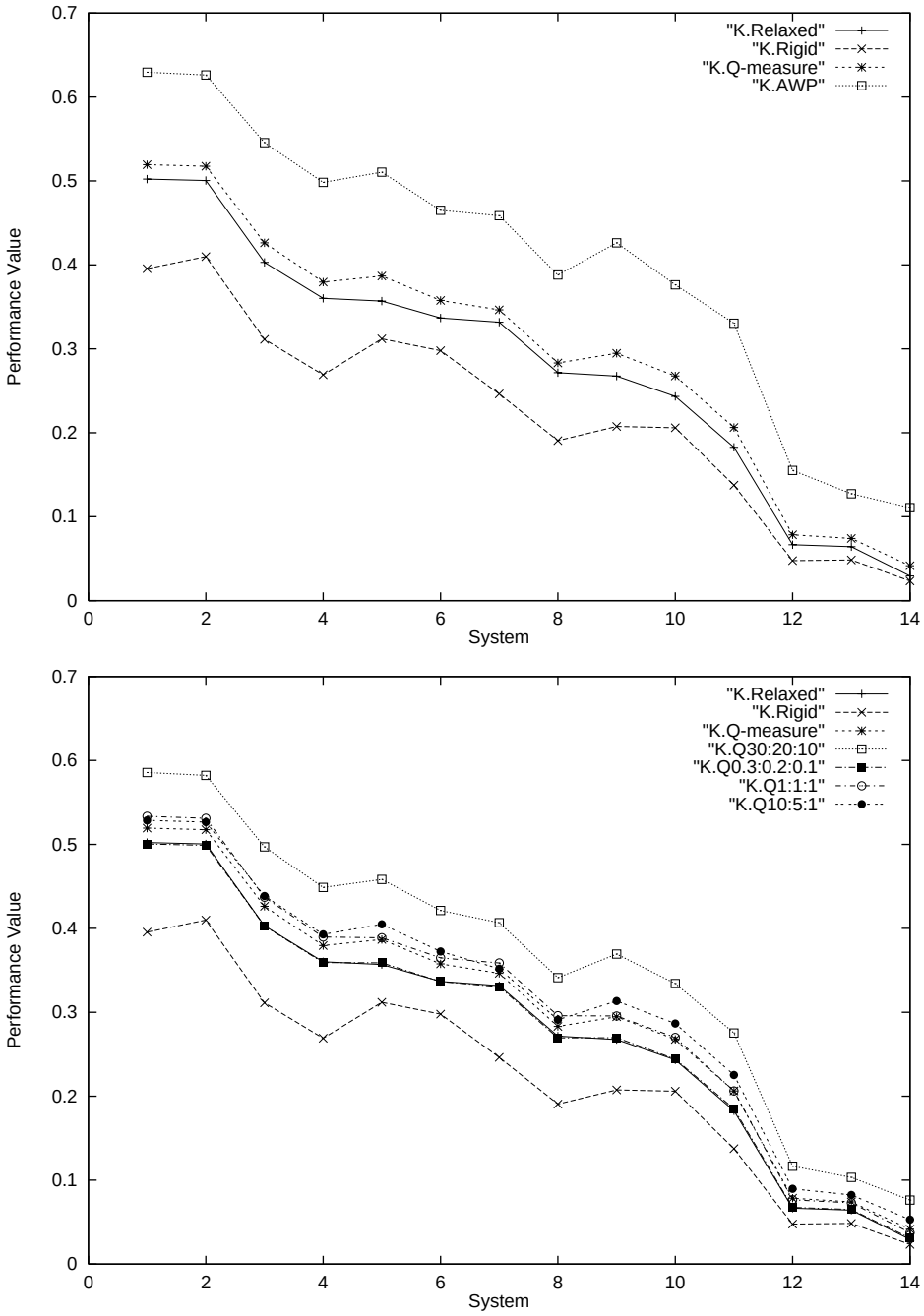

Figure 7. System ranking comparisons with Relaxed Average Precision (K-runs).

Table 7. Spearman/Kendall Rank Correlations: Averages over C, J, E and K (Q-measure etc.).

(a)	Rigid	Q-measure	AWP
Relaxed	.9744/.8965	**.9954**/.9681	.9846/.9213
Rigid	-	.9702/.8825	.9571/.8446
Q-measure	-	-	.9877/.9353
AWP	-	-	-

(b)	Q30:20:10	Q0.3:0.2:0.1	Q1:1:1	Q10:5:1
Relaxed	.9878/.9378	**.9983**/.9888	**.9986**/.9887	.9876/.9360
Rigid	.9605/.8537	.9726/.8882	.9733/.8918	.9655/.8656
Q-measure	.9884/.9413	**.9963**/.9727	**.9949**/.9653	.9887/.9480

Table 8. Spearman/Kendall Rank Correlations for the 45 C runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9864/.9313	.9867/.9293	.9863/.9293
Q-measure	.9867/.9232	.9871/.9253	.9883/.9333
R-Precision	-	**.9960**/.9616	**.9938**/.9495
R-measure	-	-	**.9971**/.9758
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9862/.9273	.9870/.9333	.9838/.9232
R-Precision	**.9939**/.9515	**.9982**/.9818	.9845/.9152
R-measure	**.9972**/.9778	**.9976**/.9758	.9893/.9333

Table 9. Spearman/Kendall Rank Correlations for the 33 J runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9886/.9356	.9866/.9318	.9843/.9242
Q-measure	**.9913**/.9318	**.9903**/.9356	.9880/.9280
R-Precision	-	**.9923**/.9583	**.9900**/.9356
R-measure	-	-	**.9910**/.9470
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9850/.9280	.9883/.9356	.9830/.9205
R-Precision	**.9920**/.9470	**.9957**/.9697	.9873/.9242
R-measure	**.9930**/.9583	**.9910**/.9583	.9883/.9356

Table 10. Spearman/Kendall Rank Correlations for the 24 E runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9852/.9275	.9870/.9348	.9870/.9348
Q-measure	.9843/.9203	.9835/.9130	.9835/.9130
R-Precision	-	**.9948**/.9638	**.9948**/.9638
R-measure	-	-	**1.000**/**1.000**
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9870/.9348	.9852/.9275	.9713/.8913
R-Precision	**.9948**/.9638	**.9983**/.9855	.9626/.8478
R-measure	**1.000**/**1.000**	**.9965**/.9783	.9591/.8551

Table 11. Spearman/Kendall Rank Correlations for the 14 K runs (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9868/.9560	.9868/.9560	.9824/.9341
Q-measure	.9780/.9121	.9780/.9121	.9824/.9341
R-Precision	-	**1.000**/**1.000**	**.9956**/.9780
R-measure	-	-	**.9956**/.9780
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9824/.9341	.9868/.9560	.9824/.9341
R-Precision	**.9956**/.9780	**1.000**/**1.000**	**.9956**/.9780
R-measure	**.9956**/.9780	**1.000**/**1.000**	**.9956**/.9780

Table 12. Spearman/Kendall Rank Correlations: Averages over C, J, E and K (R-measure etc.).

(a)	R-Precision	R-measure	R-WP
Relaxed	.9868/.9376	.9868/.9380	.9850/.9306
Q-measure	.9851/.9219	.9847/.9215	.9856/.9271
R-Precision	-	**.9958**/.9709	**.9936**/.9567
R-measure	-	-	**.9959**/.9752
R-WP	-	-	-

(b)	R30:20:10	R0.3:0.2:0.1	R10:5:1
Relaxed	.9852/.9311	.9868/.9381	.9801/.9173
R-Precision	**.9941**/.9601	**.9980**/.9843	.9825/.9163
R-measure	**.9964**/.9785	**.9963**/.9781	.9831/.9255