The Philosophy of Information Retrieval Evaluation

Ellen M. Voorhees

National Institute of Standards and Technology, Gaithersburg MD 20899 USA
ellen.voorhees@nist.gov

Abstract. Evaluation conferences such as TREC, CLEF, and NTCIR are modern examples of the Cranfield evaluation paradigm. In the Cranfield paradigm, researchers perform experiments on test collections to compare the relative effectiveness of different retrieval approaches. The test collections allow the researchers to control the effects of different system parameters, increasing the power and decreasing the cost of retrieval experiments as compared to user-based evaluations. This paper reviews the fundamental assumptions and appropriate uses of the Cranfield paradigm, especially as they apply in the context of the evaluation conferences.

The evaluation of information retrieval (IR) systems is the process of assessing how well a system meets the information needs of its users. There are two broad classes of evaluation, system evaluation and user-based evaluation. User-based evaluation measures the user's satisfaction with the system, while system evaluation focuses on how well the system can rank documents. Since the goal is to determine how well a retrieval system meets the information needs of users, user-based evaluation would seem to be much preferable over system evaluation: it is a much more direct measure of the the overall goal. However, user-based evaluation is extremely expensive and difficult to do correctly. A properly designed user-based evaluation must use a sufficiently large, representative sample of actual users of the retrieval system (whose daily routine will be interrupted by the evaluation); each of the systems to be compared must be equally well developed and complete with an appropriate user interface; each subject must be equally well trained on all systems and care must be taken to control for the learning effect [18]. Such considerations lead IR researchers to use the less expensive system evaluation for some purposes.

System evaluation is, by design, an abstraction of the retrieval process that equates good performance with good document rankings. The abstraction allows experimenters to control some of the variables that affect retrieval performance thus increasing the power of comparative experiments. These laboratory tests are much less expensive than user-based evaluations while providing more diagnostic information regarding system behavior.

Laboratory testing of retrieval systems was first done in the Cranfield 2 experiment [3]. The experiment introduced a paradigm for system evaluation that has been the dominant experimental IR model for four decades, and is the model used in evaluation efforts such as the Text REtrieval Conference (TREC), the Cross-Language Evaluation Forum (CLEF), and the NII-NACSIS Test Collection for IR Systems (NTCIR). This paper examines the assumptions inherent in the Cranfield paradigm, thereby prescribing when such system testing is appropriate. The first section reviews the history of the Cranfield tradition. Section 2 describes how test collections used in current evaluation conferences are built using pooling, and examines the effect of pooling on the quality of the test collection. Section 3 summarizes a series of experiments run on TREC collections that demonstrates that comparative evaluations are stable despite changes in the relevance judgments. The results of the experiments validate the utility of test collections as laboratory tools. The evaluation of cross-language retrieval systems presents special challenges that are considered in section 4. The paper concludes with some general reminders of the limits of laboratory tests.

1 The Cranfield Paradigm

The Cranfield experiments were an investigation into which of several alternative indexing languages was best [3]. A design goal for the Cranfield 2 experiment was to create "a laboratory type situation where, freed as far as possible from the contamination of operational variables, the performance of index languages could be considered in isolation" [4]. The experimental design called for the same set of documents and same set of information needs to be used for each language, and for the use of both precision and recall to evaluate the effectiveness of the search. (Recall is the proportion of relevant documents that are retrieved while precision is the proportion of retrieved documents that are relevant.) Relevance was based on topical similarity where the judgments were made by domain experts (i.e., aeronautics experts since the document collection was an aeronautics collection).

While the Cranfield 2 experiment had its detractors [16], many other researchers adopted the concept of retrieval test collections as a mechanism for comparing system performance [18, 13, 17]. A test collection consists of three distinct components: the documents, the statements of information need (called "topics" in this paper), and a set of relevance judgments. The relevance judgments are a list of which documents should be retrieved for each topic.

The Cranfield experiments made three major simplifying assumptions. The first assumption was that relevance can be approximated by topical similarity. This assumption has several implications: that all relevant documents are equally desirable, that the relevance of one document is independent of the relevance of any other document, and that the user information need is static. The second assumption was that a single set of judgments for a topic is representative of the user population. The final assumption was that the lists of relevant documents for each topic is complete (all relevant documents are known). The vast majority of test collection experiments since then have also assumed that relevance is a binary choice, though the original Cranfield experiments used a five-point relevance scale.

Of course, in general these assumptions are not true, which makes laboratory evaluation of retrieval systems a noisy process. Researchers have evolved a standard experimental design to decrease the noise, and this design has become an intrinsic part of the Cranfield paradigm. In the design, each retrieval strategy to be compared produces a ranked list of documents for each topic in a test collection, where the list is ordered by decreasing likelihood that the document should be retrieved for that topic. The effectiveness of a strategy for a single topic is computed as a function of the ranks of the relevant documents. The effectiveness of the strategy on the whole is then computed as the average score across the set of topics in the test collection.

This design contains three interrelated components—the number of topics used, the evaluation measures used, and the difference in scores required to consider one method better than the other—that can be manipulated to increase the reliability of experimental findings [2]. Since retrieval system effectiveness is known to vary widely across topics, the greater the number of topics used in an experiment the more confident the experimenter can be in its conclusions. TREC uses 25 topics as a minimum and 50 topics as the norm. A wide variety of different evaluation measures have been developed (see van Rijsbergen [22] for a summary), and some are inherently less stable than others. For example, measures based on very little data such as precision at one document retrieved (i.e., is the first retrieved document relevant?) are very noisy, and the mean average precision measure, which measures the area underneath the entire recall-precision curve, is much more stable. Requiring a larger difference between scores before considering the respective retrieval methods to be truly different increases reliability at the cost of not being able to discriminate between as many methods.

The remainder of this paper reviews a series of experiments that examine the reliability of test collections for comparing retrieval systems. The experiments show that the basic assumptions of the Cranfield paradigm need not be strictly true for test collections to be viable laboratory tools. The focus on comparative results is deliberate and important. A consequence of the abstraction used in the paradigm is that the absolute score of an evaluation measure for some retrieval run is not meaningful in isolation. The only valid use of such a score is to compare it to the score of a different retrieval run that used the exact same test collection. Note that this means that comparing the score a retrieval system obtained in TREC or CLEF one year to to the score the system obtained the following year is invalid since the test collection is different in the two years. It is also invalid to compare the score obtained over a subset of topics in a collection to the score obtained over the whole set of topics (or a different subset of topics).

2 Completeness of Relevance Judgments

A major departure from the original Cranfield 2 experiments and modern test collections such as the TREC collections is the relative emphasis given to collection size and the completeness of relevance judgments. In his account of the Cranfield experiments [4], Cleverdon remarks

> Experience had shown that a large collection was not essential, but it was vital that there should be a complete set of relevance decisions for every question against every document, and, for this to be practical, the collection had to be limited in size.

However, further experience has shown that collection size (i.e., the number of documents in the collection) *does* matter. IR is challenging because of the large number of different ways the same concept can be expressed in natural language, and larger collections are generally more diverse. Unfortunately, even a moderately large collection cannot possibly have complete relevance judgments. Assuming a judgment rate of one document per 30 seconds, and judging round-the-clock with no breaks, it would still take more than nine months to judge one topic for a collection of 800,000 documents (the average size of a TREC collection). Instead, modern collections use a technique called pooling [15] to create a subset of the documents (the "pool") to judge for a topic. Each document in the pool for a topic is judged for relevance by the topic author, and documents not in the pool are assumed to be irrelevant to that topic.

Before describing the effects of incompleteness on a test collection, we will describe the process by which the collections are built in more detail. The building process will be described in terms of the TREC workshops here, but the process is similar for other evaluation conferences as well.

2.1 Building Large Test Collections

NIST provides a document set and a set of topics to the TREC participants. Each participant runs the topics against the documents using their retrieval system, and returns to NIST a ranked list of the top 1000 documents per topic. NIST forms pools from the participants' submissions, which are judged by the relevance assessors. Each submission is then evaluated using the resulting relevance judgments, and the evaluation results are returned to the participant.

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. For example, if the operational setting is a medical library it makes little sense to use a collection of movie reviews as the document set. The TREC ad hoc collections contain mostly newspaper or newswire articles, though some government documents (the *Federal Register*, a small collection of patent applications) are also included in some of the collections to add variety. These collections contain about about 2 gigabytes of text (between 500,000 and 1,000,000 documents). The document sets used in various tracks have been smaller and larger depending on the needs of the track and the availability of data.

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. The format of a topic statement has evolved since the beginning of TREC, but it has been stable for the past several years. A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative.

TREC topic statements are created by the same person who performs the relevance assessments for that topic. Usually, each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the document collection using NIST's PRISE system to estimate the likely number of relevant documents per candidate topic. The NIST TREC team selects the final set of topics from among these candidate topics based on the estimated number of relevant documents and balancing the load across assessors. The TREC ad hoc collections contain 50 topics.

The relevance judgments are what turns a set of documents and topics into a test collection. TREC has almost always used binary relevance judgments—either a document is relevant to the document or it is not. To define relevance for the assessors, the assessors are told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

The judgment pools are created as follows. NIST selects the maximum number of runs that can be contributed to the pools by a single participating group; each group contributes this many runs to the pools unless they submitted fewer runs to NIST (in which case all their runs contribute to the pools). When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST merges the runs into the pools respecting this preferred ordering. For each selected run, the top X documents (usually, $X = 100$) per topic are added to the topics' pools. Since the retrieval results are ranked by decreasing similarity to the query, the top documents are the documents most likely to be relevant to the topic. Many documents are retrieved in the top X for more than one run, so the pools are generally much smaller the theoretical maximum of $X \times$ *the-number-of-selected-runs* documents (usually about 1/3 the maximum size). Each pool is sorted by document identifier so assessors cannot tell if a document was highly ranked by some system or how many systems (or which systems) retrieved the document.

2.2 Effects of Incompleteness

As mentioned above, pooling violates one of the original tenets of the Cranfield paradigm since it does not produce complete judgments. The concern is that evaluation scores for methods that did not contribute to the pools will be deflated relative to methods that did contribute because the non-contributors will have highly ranked unjudged documents that are assumed to be not relevant.

Figure 1 shows that there are relevant documents that are contributed to the pool by exactly one group (call these the unique relevant documents). The figure contains a histogram of the total number of unique relevant documents found by a group over the 50 test topics for four of the TREC ad hoc collections. The totals are subdivided into categories where "Automatic" designates documents that were retrieved only by completely automatic runs, "Manual" designates documents that were retrieved only by manual runs, "Mixture" designates documents that were retrieved by runs of different types, and "Others" designates documents that were retrieved by other tracks that contributed to the ad hoc pools. Each of the histograms in the figure uses the same scale. A dot underneath the x-axis indicates a group is plotted there, and all groups that retrieved at least one unique relevant document are plotted. For each year, the majority of unique documents was retrieved by manual runs. The distribution of unique relevant documents found was roughly the same over the four years.

Figure 1 suggests that the TREC collections contain relevant documents that have not been judged. If the particular group that contributed a unique relevant document had not participated in TREC that year, that document would not have been judged and would have been assumed to be not relevant. Presumably, there are other documents that didn't make it into the pools that also would have been judged relevant. Indeed, a test of the TREC-2 and TREC-3 collections demonstrated the presence of unjudged relevant documents [8]. In this test, relevance assessors judged the documents in new pools formed from the second 100 documents in the ranked results submitted by participants. On average, the assessors found approximately one new relevant document per run (i.e., one relevant document that was not in the pool created from the top 100 documents of each ranking). The distribution of the new relevant documents was roughly uniform across runs, but was skewed across topics—topics that had many relevant documents initially also had many more new relevant documents.

Zobel found the same pattern of unjudged documents in his analysis of the effect of pooling [26]. He also demonstrated that the TREC collections could still be used to compare different retrieval methods since the collections were not biased against the unjudged runs. In this test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

Most of the subsequent TREC collections have been examined using a similar test. In these tests, the entire set of uniquely retrieved relevant documents for a group across all of the runs that group submitted are removed from the relevance judgment set when evaluating a run from that group. This is a more stringent variation of the test used by Zobel in that it completely removes any effect of that group's participation. For the TREC-8 ad hoc collection, the mean

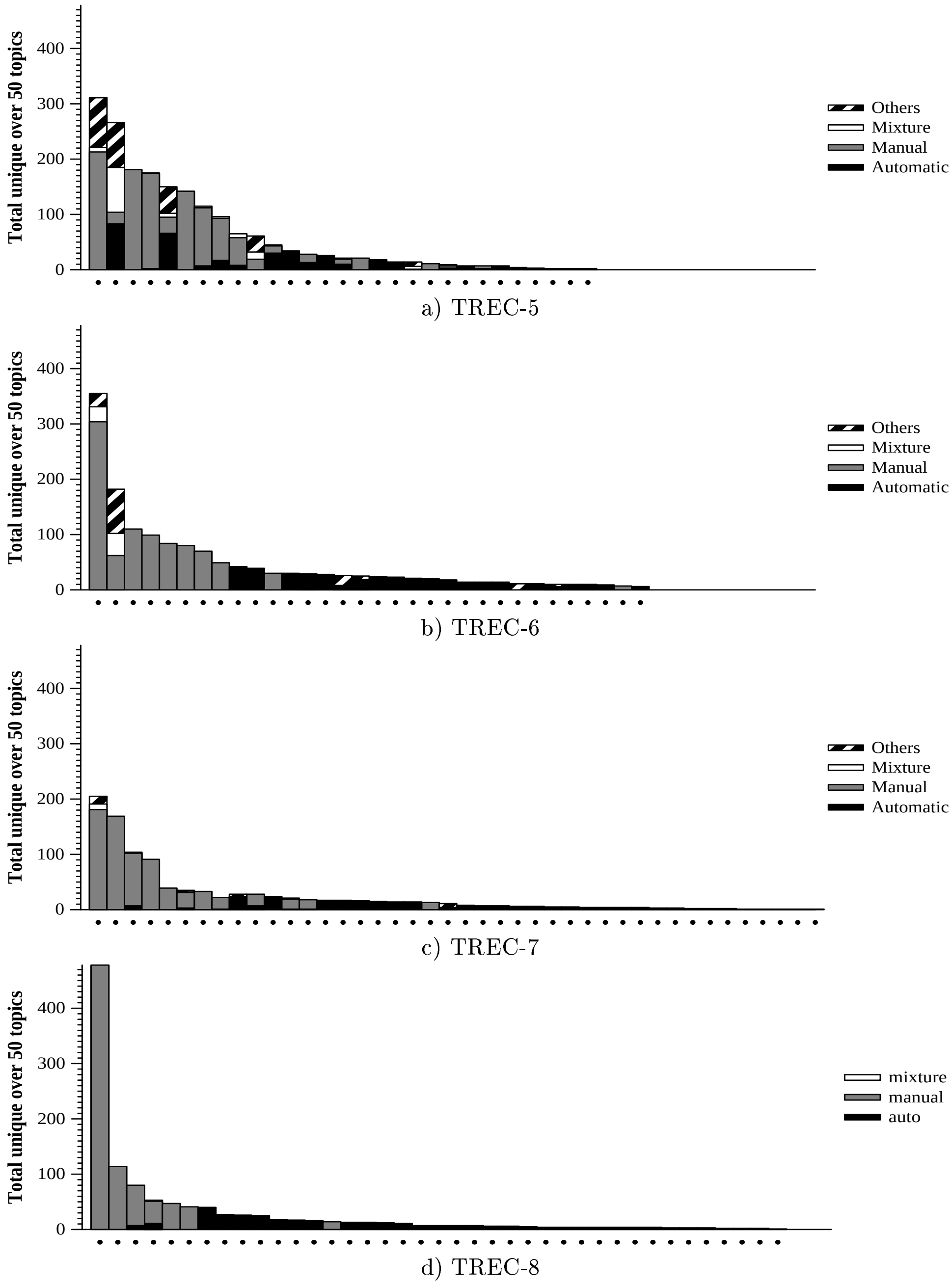

Fig. 1. Total number of unique relevant documents retrieved per TREC. Each total gives the percentages of the total that were retrieved by Automatic, Manual, Mixed, or Other runs. Groups are indicated by a dot beneath the x-axis. All groups that retrieved at least one unique relevant document are plotted.

percentage difference in mean average precision over the 71 runs that contributed to pools was 0.78 %, with a maximum difference of 9.9 %. Not surprisingly, the manual groups that had the largest number of unique relevant documents (see figure 1) also had the largest percentage differences in mean average precision. But given that the manual runs' contributions are in the pool, the difference in evaluation results for automatic runs is negligible. For automatic runs, the largest percentage difference in mean average precision scores was 3.85 %, which corresponded to an absolute difference of only .0001. Every automatic run that had a mean average precision score of at least .1 had a percentage difference of less than 1 %.

Figure 2 shows the absolute difference in mean average precision scores plotted against the number of unique relevant documents contributed by that run's group for each automatic run for TREC-8. The runs are sorted by increasing difference and then by number of unique relevant documents. The two obvious outliers in number of unique relevant documents (for runs `GE8ATDN1` and `iit99au1`) reflect organizations that submitted manual runs in addition to automatic runs; the vast majority of their unique relevant documents were contributed by their manual run.

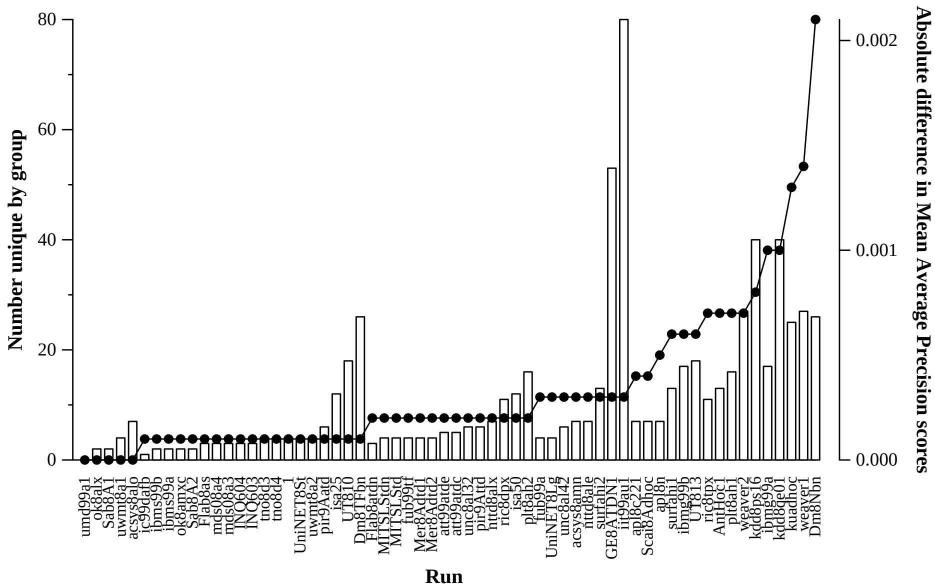

Fig. 2. Absolute difference in mean average precision scores when a run is evaluated using relevance pools with and without that group's unique relevant document for TREC-8 automatic, ad hoc runs. Also plotted is the number of unique relevant documents contributed to the pools by that group. Runs are ordered by increasing absolute difference and by increasing number of unique relevant documents.

While the lack of any appreciable difference in the scores of the automatic runs is not a guarantee that all relevant documents have been found, it is very strong evidence that the test collection is reliable for comparative evaluations of retrieval runs. Pool depth and diversity are important factors in building a test collection through pooling, but with adequate controls, the resulting test collection is not biased against runs that did not contribute to the pools. The quality of the pools is significantly enhanced by the presence of recall-oriented manual runs, an effect exploited by the organizers of the NTCIR workshops who perform their own manual runs to supplement the pools [11].

In the end, the concern regarding completeness is something of a red herring. Since test collections support only comparative evaluations, the important factor is whether the relevance judgments are unbiased. Having complete judgments ensures that there is no bias in the judgments; pooling with sufficiently diverse pools is a good approximation. The importance of an unbiased judgment set argues against the proposals for different pooling strategies that find more relevant documents in fewer total documents judged. Zobel suggests judging more documents for topics that have had many relevant documents found so far and fewer documents for topics with fewer relevant documents found so far [26]. However, assessors would know that documents added later in the pools came from lower in the systems' rankings and that may affect their judgments. Cormack et al. suggest judging more documents from runs that have returned more relevant documents recently and fewer documents from runs that have returned fewer relevant document recently [5]. But that would bias the pools toward systems that retrieve relevant documents early in their rankings. For test collections, a smaller, fair judgment set is always preferable to a larger biased set.

3 Differences in Relevance Judgments

Incompleteness is a relatively recent criticism of the Cranfield paradigm since the original test collections were complete. Inconsistency—the fact that different relevance assessors produce different relevance sets for the same topics—has been the main perceived problem with test collections since the initial Cranfield experiments [20, 7, 9]. The main gist of the critics' complaint is that relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [14]. Critics question how valid conclusions can be drawn when the evaluation process is based on something as volatile as relevance.

To study the effect of different relevance judgments on the stability of comparative evaluation results, NIST obtained three independent sets of assessments for each of the 49 topics used in the TREC-4 evaluation (Topics 202–250). The TREC-4 relevance assessors were asked to judge additional topics once they had finished with the main TREC-4 assessing. Call the author of a topic its primary assessor. After the primary assessor was finished with a topic, a new document pool was created for it. This new pool consisted of all of the relevant documents as judged by the primary assessor up to a maximum of 200 relevant documents (a random sample of 200 relevant documents was used if there were more than 200 relevant documents) plus 200 randomly selected documents that the primary assessor judged not relevant. The new pool was sorted by document identifier and given to two additional assessors (the secondary assessors) who each independently judged the new pool for relevance. Because of the logistics involved, a topic was given to whatever secondary assessor was available at the time, so some individual assessors judged many more topics than others. However, each topic was judged by three individuals.

3.1 Assessor Agreement

The overlap of the relevant document sets can be used to quantify the level of agreement among different sets of relevance assessments [12]. Overlap is defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets. Table 1 gives the mean overlap for each pair of assessors and the set of three assessors. Documents that the primary assessor judged relevant but that were not included in the secondary pool (because of the 200 document limit) were added as relevant documents to the secondary assessors' judgments for the analysis. The overlap is less than 50%, indicating that the assessors substantially disagreed on the judgments. Across all topics, 30% of the documents that the primary assessor marked relevant were judged nonrelevant by both secondary assessors.

Table 1. Mean overlap for each assessor pair and the set of three assessors.

Assessor Group	Overlap
Primary & A	.421
Primary & B	.494
A & B	.426
All three	.301

3.2 Effect of Inconsistency

To test how evaluation results change with these differences in assessments, we compare system rankings produced using different relevance judgments sets. A system ranking is a list of the systems under consideration sorted by the value each system obtained for some evaluation measure.

Each system is evaluated over a set of topics, and each topic has a different set of judgments produced by each assessor. Call the concatenation of one judgment set per topic a *qrels* (for query-relevance set). With three independent judgments for each of 49 topics, we can theoretically create 3^{49} different qrels by using different combinations of assessor's judgments for the topics, and evaluate the systems using each qrels. Note that each of these qrels might have been the qrels produced after the TREC conference if that set of assessors had been assigned those topics. To simplify the analysis that follows, we discarded Topic 214 since Secondary Assessor A judged no documents relevant for it. That leaves 48 topics and 3^{48} possible qrels.

Three of the 3^{48} possible qrels are special cases. The original qrels set consists of the primary assessments for each topic—this is the qrels released after TREC-4 except that it lacks Topic 214. The set of judgments produced by the Secondary A judge for each topic, and the set of judgments produced by the Secondary B judge for each topic constitute the Secondary A qrels and the Secondary B qrels respectively. We created a sample of size 100,000 of the remaining qrels by randomly selecting one of the primary or secondary assessors for each of the 48 topics and combining the selected judgments into a qrels. Adding the three distinguished qrels to the sample gives a total of 100,003 qrels that were used to evaluate retrieval systems. Finally, two additional qrels, the union and intersection qrels, were created from the relevance judgments. In the union qrels a document is considered to be relevant to a topic if any assessor judged it relevant to that topic; in the intersection qrels a document is considered to be relevant to a topic if all three assessors judged it relevant to that topic. Because two topics had no documents that all assessors agreed were relevant (219 and 232), the intersection qrels contains only 46 topics.

There were 33 category A ad hoc retrieval systems used in TREC-4. We evaluated each of these systems against each of the qrels in the sample of 100,003 qrels and computed the sample mean of the mean average precision for each system. The means are plotted in Figure 3 where the systems are sorted by decreasing mean. The error bars in Figure 3 indicate the minimum and the maximum mean average precision obtained for that system over the sample. Also plotted in the figure are the mean average precision scores computed using the original, union, and intersection qrels. These points demonstrate how the system ranking changes for an individual qrels versus the ranking by the mean: a system with a symbol higher than the corresponding symbol of a system to its left would be ranked differently in the individual qrels ranking. For example, the `pircs2` and `uwgcl1` systems (shown in position 2 and 3 in Figure 3) would switch positions when evaluated by the Original qrels.

The plot in Figure 3 demonstrates that the mean average precision score *does* change depending on the qrels used in the evaluation. The difference between the minimum and maximum mean average precision values is greater than .05 for most systems. However, the changes are very highly correlated across systems. That is, if a particular system gets a relatively high score with a particular qrels, then it is very likely that the other systems will also get a relatively high score with that qrels. The union qrels (the triangle in Figure 3) is close to the top of the range for each system, for example.

The correlation can be quantified by using a measure of association between the different system rankings. We used a correlation based on Kendall's tau [19] as the measure of association between two rankings. Kendall's tau computes the distance between two rankings as the minimum number of pairwise adjacent

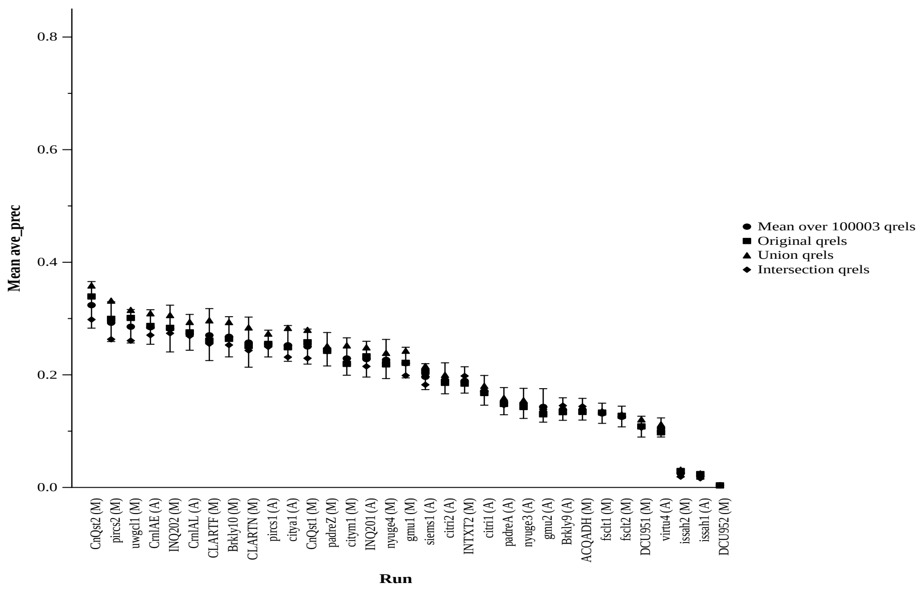

Fig. 3. Sample mean, min, and max of the mean average precision computed for each of 33 TREC-4 systems over a sample of 100,003 qrels. Also plotted are the mean average precision for the original, union, and intersection qrels. Systems are labeled as either manual (M) or automatic (A).

swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is -1.0, and the expected correlation of two rankings chosen at random is 0.0.

We computed the mean of the Kendall correlations in the sample of 100,003 qrels in two ways. In the first case, we took the mean of the correlations between the ranking produced by the original qrels and the rankings produced by each of the other 100,002 qrels. In the second case, we took a random subsample of 1000 qrels and computed the mean correlation across all pairs in the subsample. The mean, minimum, and maximum Kendall correlations for the two methods are given in Table 2. The numbers in parentheses show the number of pairwise adjacent swaps a correlation represents given that there are 33 systems in the rankings. (Note that the way in which the qrels were constructed means that any two qrels are likely to contain the same judgments for 1/3 of the topics. Since the qrels are not independent of one another, the Kendall correlation is probably slightly higher than the correlation that would result from completely independent qrels.)

On average, it takes only 16 pairwise, adjacent swaps to turn one ranking into another ranking. The vast majority of the swaps that do take place are between systems whose mean average precisions are very close (a difference of less than .01).

Table 2. Kendall correlation of system rankings and corresponding number of pairwise adjacent swaps produced by different qrels. With 33 systems, there is a maximum of 528 possible pairwise adjacent swaps.

	Mean	Min	Max
with Original	.9380 (16)	.8712 (34)	.9962 (1)
in subsample	.9382 (16)	.8409 (42)	.9962 (1)

The plots of the union and intersection qrels evaluations in Figure 3 show that the evaluation using those qrels sets is not different from the evaluation using the other qrels sets. The intersection and union qrels differ from the other sets in that each topics' judgments are a combination of judges' opinions in the union and intersection qrels. The intersection qrels set represents a particularly stringent definition of relevance, and the union qrels a very weak definition of relevance. Nonetheless, in all but two cases (intersection qrels for systems `pircs2` and `uwgcl1`), the mean average precision as computed by the union and intersection qrels falls within the range of values observed in the sample of 100,003 qrels. The corresponding rankings are also similar: the Kendall correlation between the original qrels ranking and the union qrels ranking is .9508, and between the original and intersection rankings is .9015.

This entire experiment was repeated several times using different evaluation measures, different topic sets, different systems, and different groups of assessors [23]. The correlation between system rankings was very high in each experiment, thus confirming that the comparative evaluation of ranked retrieval results is stable despite the idiosyncratic nature of relevance judgments.

4 Cross-Language Test Collections

The Cranfield paradigm is appropriate for cross-language retrieval experiments, though building a good cross-language test collection is more difficult than building a monolingual collection. In monolingual collections, the judgments for a topic are produced by one assessor. While this assessor's judgments may differ from another assessor's judgments, the judgment set represents an internally consistent sample of judgments. CLEF cross-language collections are produced using a separate set of assessors for each language represented in the document set. Thus multiple assessors judge a single topic across the entire collection. This necessitates close coordination among assessors so that "gray areas" can be judged consistently across languages.

Pooling is also much more difficult to coordinate for cross-language collections. As the results in section 2 demonstrate, the quality of the test collection depends on having diverse pools. Yet it is very difficult to get equally large, diverse pools for all languages contained within a multilingual collection. Both the number of runs submitted by participants and the documents retrieved within a run are usually skewed in favor of some languages at the expense of others. As a result, the pools for the minority languages are smaller and less diverse than the pools for the majority languages, which introduces an unknown bias into the judgments.

Another concern with the TREC cross-language collections is that the cross-language tasks have tended not to receive the recall-oriented manual runs that are beneficial for collection building. Analysis of the effect of incompleteness on the cross-language collections (as described in section 2) has sometimes showed mean differences in mean average precision scores computed with and without a group's unique relevant documents as large as 6–8 %, compared to mean percentage differences of less than 2 % for automatic runs for the monolingual TREC collections [24, 25]. (The same analysis on the CLEF 2000 multilingual collection showed an average difference of less than 1 % [1].) The somewhat larger average differences do not invalidate the collections since comparative results are generally remain quite stable. However, experiments who find many unjudged documents in the top-ranked list of only one of a pair of runs to be contrasted should proceed with care.

5 Conclusion

Test collections are research tools that provide a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. As such they are abstractions of an operational retrieval environment. Test collections are useful because they allow researchers to control some of the variables that affect retrieval performance, increasing the power of comparative experiments while drastically decreasing the cost as compared to user-based evaluations.

This paper has summarized a set of experiments that demonstrate the validity of using test collections as laboratory tools. In response to the criticism that test collections use of a static set of binary, topical relevance judgments to represent correct retrieval behavior, we showed that comparative retrieval results are stable despite changes in the relevance judgments. We also showed that an adequately controlled pooling process can produce an unbiased judgment set, which is all that is needed for comparative evaluation.

The final test of the Cranfield paradigm is whether the conclusions reached from the laboratory experiments transfer to operational settings. Hersh and his colleagues suggest that the results may not transfer since they were unable to verify the conclusions from a laboratory experiment in either of two user studies [10, 21]. However, the first user study involved only 24 searchers and six topics, and the second user study involved 25 searchers and eight topics. The results of the user studies did not show that the conclusions from the laboratory test were wrong, simply that the user studies could not detect any differences. As such, the studies are a good illustration of the difficulties of performing user studies to evaluate retrieval systems. Any measure of the retrieval technology actually in use today demonstrates the results do transfer. Basic components of current web search engines and other commercial retrieval systems—including full text indexing, term weighting, and relevance feedback—were first developed on test collections.

Because the assumptions upon which the Cranfield paradigm is based are not strictly true, the evaluation of retrieval systems is a noisy process. The primary consequence of the noise is the fact that evaluation scores computed from a test collection are valid *only* in comparison to scores computed for other runs using the exact same collection. A second consequence of the noise is that there is an (unknown) amount of error when comparing two systems on the same collection. This error can be reduced by using many topics, by repeating the whole experiment (different sets of topics/judgments) multiple times, and by accepting two systems as different only if the delta between their respective scores is larger.

Acknowledgements

My thanks to Chris Buckley for many discussions regarding retrieval system evaluation, and for his comments on a draft of this paper.

References

1. Martin Braschler. CLEF 200 – Overview of results. In Carol Peters, editor, *Cross-Language Information Retrieval and Evaluation; Lecture Notes in Computer Science 2069*, pages 89–101. Springer, 2001.
2. Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In N. Belkin, P. Ingwersen, and M.K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.
3. C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967. (Reprinted in *Readings in Information Retrieval*, K. Sparck-Jones and P. Willett, editors, Morgan Kaufmann, 1997).
4. Cyril W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Reserach and Development in Information Retrieval*, pages 3–12, 1991.
5. Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In Croft et al. [6], pages 282–289.
6. W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press, New York.
7. C. A. Cuadra and R. V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4):291–303, 1967.
8. Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, October 1996. NIST Special Publication 500-236.
9. Stephen P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

10. William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In N. Belkin, P. Ingwersen, and M.K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 17–24, 2000.
11. Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, 1999.
12. M.E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359, 1969.
13. G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.
14. Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.
15. K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.
16. Karen Sparck Jones. The Cranfield tests. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, chapter 13, pages 256–284. Butterworths, London, 1981.
17. Karen Sparck Jones. *Information Retrieval Experiment.* Butterworths, London, 1981.
18. Karen Sparck Jones and Peter Willett. Evaluation. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, chapter 4, pages 167–174. Morgan Kaufmann, 1997.
19. Alan Stuart. Kendall's tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.
20. M. Taube. A note on the pseudomathematics of relevance. *American Documentation*, 16(2):69–72, April 1965.
21. Andrew H. Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–231, 2001.
22. C.J. van Rijsbergen. *Information Retrieval*, chapter 7. Butterworths, 2 edition, 1979.
23. Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.
24. Ellen M. Voorhees and Donna Harman. Overview of the eighth Text REtrieval Conference (TREC-8). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, 2000. NIST Special Publication 500-246. Electronic version available at http://trec.nist.gov/pubs.html.
25. Ellen M. Voorhees and Donna Harman. Overview of TREC 2001. In *Proceedings of TREC 2001 (Draft)*, 2001. To appear.
26. Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [6], pages 307–314.