Overview of the TREC 2005 Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system's least effective topics.

The 2005 edition of the track used 50 topics that had been demonstrated to be difficult on one document collection, and ran those topics on a different document collection. Relevance information from the first collection could be exploited in producing a query for the second collection, if desired. The main measure for evaluating system effectiveness is "gmap", a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results. As in previous years, the most effective retrieval strategy was to expand queries using terms derived from additional corpora. The relative difficulty of topics differed across the two document sets.

Systems were also required to rank the topics by predicted difficulty. This task is motivated by the hope that systems will eventually be able to use such predictions to do topic-specific processing. This remains a challenging task. Since difficulty depends on more then the topic set alone, prediction methods that train on data from other test collections do not generalize well.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her request. The previous two editions of the robust track have demonstrated that average effectiveness masks individual topic effectiveness, and that optimizing standard average effectiveness measures usually harms the already ineffective topics.

This year's track used 50 topics that had been demonstrated to be difficult for the TREC Disks 4&5 document set (CD45) and ran those topics against the AQUAINT document set. Relevance information from the CD45 collection could be exploited in producing a query for the AQUAINT collection, if desired.

A focus of the robust track since its inception has been developing the evaluation methodology for measuring how well systems avoid abysmal results for individual topics. Two measures introduced in the initial track were subsequently shown to be relatively unstable even for as many as 100 topics in the test set [3]. Those measures have been dropped from this year's results and have been replaced by the geometric MAP, or "gmap", measure. Gmap is computed as a geometric mean of the average precision scores of the test set of topics, as opposed to the arithmetic mean used to compute the standard MAP measure. Experiments using the TREC 2004 robust track results suggest that the measure gives appropriate emphasis to poorly performing topics while being stable with as few as 50 topics.

In addition to producing a ranked list of documents for each topic, systems were also required to rank the topics by predicted difficulty. The motivation for this task is the hope that systems will eventually be able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the systems' retrieval results. Section 3 examines the differences in the test collections built with the different document sets. Despite the diversity of runs that contributed to the pools for the AQUAINT collection, analysis of the resulting relevance judgments suggests the pool depth was insufficient with respect to the document set size. Section 4 then examines the difficulty prediction task. The final section summarizes the results of the three-year run of the track: this is the concluding year of a separate robust track, though the gmap measure with its emphasis on poorly performing topics will be incorporated into ad hoc tasks in other tracks.

1 The Robust Retrieval Task

The task within the robust retrieval track is a traditional ad hoc task. The document set used in this year's track was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). This collection consists of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection.

The topic set consisted of 50 topics that had been used in ad hoc and robust tracks in previous years where they were run against the document set comprised of the documents on TREC disks 4&5 (minus the *Congressional Record*). These topics each had low median average precision scores in both the initial TREC in which they were used and in previous robust tracks, and were chosen for the track precisely because they are assumed to be difficult topics.

The 50 test topics were selected from a somewhat larger set based on having at least three relevant documents in the AQUAINT collection. NIST assessors were given a set of of topic statements and asked to search the AQUAINT collection looking for at least three relevant documents. Assessors were given the general guideline that they should spend no more than about 30 minutes searching for any one topic. The assessor stopped searching for relevant documents as soon as he or she found three relevant documents or when they felt they had exhausted the collection without finding three relevant documents. The topics for which fewer than three relevant documents were retrieved were discarded. The entire process stopped as soon as 50 topics with a minimum of three relevant documents were found.

The assessor who judged a topic on the AQUAINT data set was in general different from the assessor who originally judged the topic on the CD45 collection. Thus, both the document set and the assessor differed between original runs using the topics and the robust 2005 runs. Nonetheless, systems were allowed to exploit the existing judgments in creating their queries for the track if they chose to do so. (Such runs were labeled as manual or "human-assisted" runs since the previous judgments were manually created. Runs that used other types of manual processing are also labeled as human-assisted.) Using the existing judgments in this manner is equivalent to the routing task performed in early TRECs.

The TREC 2005 HARD track used the same test collections as the robust track. Pools for document judging were created from one baseline and one final run for each HARD track participant, and one run per robust track participant. Because there were limited assessing resources, relatively shallow pools were created. The top 55 documents per topic for each pool run were added to the pools, producing pools that had a mean size of 756 documents (minimum 350, maximum 1390). While these pools are shallow, the expectation was that the diversity of the runs used to make the pools would result in sufficiently comprehensive relevance judgments. This hypothesis is explored later in section 3. Documents in the pools were judged not relevant, relevant, or highly relevant, with both highly relevant and relevant judgments used as the relevant set for evaluation.

Runs were evaluated using `trec_eval`, and the standard measures are included in the evaluation report for robust runs. The primary measure for the track is the geometric MAP (gmap) score computed over the 50 test topics. Gmap was introduced in the TREC 2004 robust track [3] as a measure that emphasizes poorly performing topics while remaining stable with as few as 50 topics. Gmap takes a geometric mean of the individual topics' average precision scores, which has the effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between larger scores. The geometric mean is equivalent to taking the log of the the individual topics' average precision scores, computing the arithmetic mean of the logs, and exponentiating back for the final gmap score. The gmap value reported for robust track runs was computed using the current version of `trec_eval` (invoked with the -a option). In this implementation, all individual topic average precision scores that are less than 0.00001 are set to 0.00001 to avoid taking logs of 0.0.

2 Retrieval Results

The robust track received a total of 74 runs from the 17 groups listed in Table 1. Participants were allowed to submit up to five runs. To have comparable runs across participating sites, if the participant submitted any automatic runs, one run was required to use just the description field of the topic statements, and one run was required to use just the title field of the topic statements. Four of the runs submitted to the track were human-assisted runs; the remaining seventy were completely automatic runs. Of the automatic runs, 24 runs were description-only runs, 34 were title-only runs,

Table 1: Groups participating in the robust track.

Arizona State University (Roussinov)	Chinese Academy of Sciences (ICT)
Ecole des Mines de Saint-Etienne	The Hong Kong Polytechnic University
Hummingbird	IBM Research, Haifa
Indiana University	IRIT/SIG
Johns Hopkins University/APL	Meiji University
Queens College, CUNY	Queensland University of Technology
RMIT University	Sabir Research, Inc.
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Massachusetts	

Table 2: Evaluation results for the best title-only and description-only runs for the top eight groups ordered by gmap.

Title-only Runs				Description-only Runs			
Run	gmap	MAP	P10	Run	gmap	MAP	P10
uic0501	0.233	0.310	0.592	ASUDE	0.178	0.289	0.536
indri05RdmmT	0.206	0.332	0.524	indri05RdmeD	0.161	0.282	0.498
pircRB05t2	0.196	0.280	0.542	ICT05qerfD	0.155	0.259	0.446
ICT05qerfTg	0.189	0.271	0.444	JuruDWE	0.129	0.230	0.472
UIUCrAt1	0.189	0.268	0.498	pircRB05d1	0.125	0.230	0.466
JuruTiWE	0.157	0.239	0.496	sab05rod1	0.114	0.184	0.404
humR05txle	0.150	0.242	0.490	humR05dle	0.114	0.201	0.432
wdf1t3qs0	0.149	0.235	0.456	wdf1t3qd	0.110	0.187	0.376

and 12 used various combinations of the topic statement.

Table 2 gives the evaluation scores for the best run for the top eight groups who submitted either a title-only run or a description-only run. The table gives the gmap, MAP, and average P(10) scores over the 50 topics. The run shown in the table is the run with the highest gmap; the table is sorted by this same value.

As in previous robust tracks, the best performing runs used some sort of external corpus to perform query expansion. Usually the external corpus was the web as viewed from the results of web search engines, though other large data sets such as a collection of TREC news documents (University of Massachusetts) or the .GOV collection (Chinese Academy of Sciences) were used as well. The behavior of the topics on the CD45 and AQUAINT document sets (examined in more detail below) is sufficiently different that expanding queries using a large external corpus was more effective on average than exploiting relevance information from the CD45 collection. For example, IBM Hafia found that using web expansion was more effective than no expansion, but expanding based on the CD45 relevance information was less effective than no expansion [4]. Sabir Research used the CD45 relevance information to produce "optimal" queries in its `sab05ror1` run [1]; these queries produced the best average precision scores for nine topics on the AQUAINT collection, but the average effectiveness across all topics was less than that of the best performing runs.

The top title-only runs, `uic0501` from the University of Illinois at Chicago and `indri05RdmmT` from the University of Massachusetts, illustrate the difference between the gmap and MAP measures. The `uic0501` run obtained a higher gmap score than the `indri05RdmmT` run, while the reverse is true for MAP. Figure 1 shows the per-topic average precision scores for the two runs. In the figure the topics are plotted on the x-axis and are sorted by decreasing average precision score obtained by the `indri05RdmmT` run. The horizontal line in the graph is plotted at an average precision of 0.05. The `indri05RdmmT` run has a better average precision score for more topics, but has seven topics for which the average precision score is less than 0.05. In contrast, the `uic0501` run has only two topics with an average precision score less than 0.05.

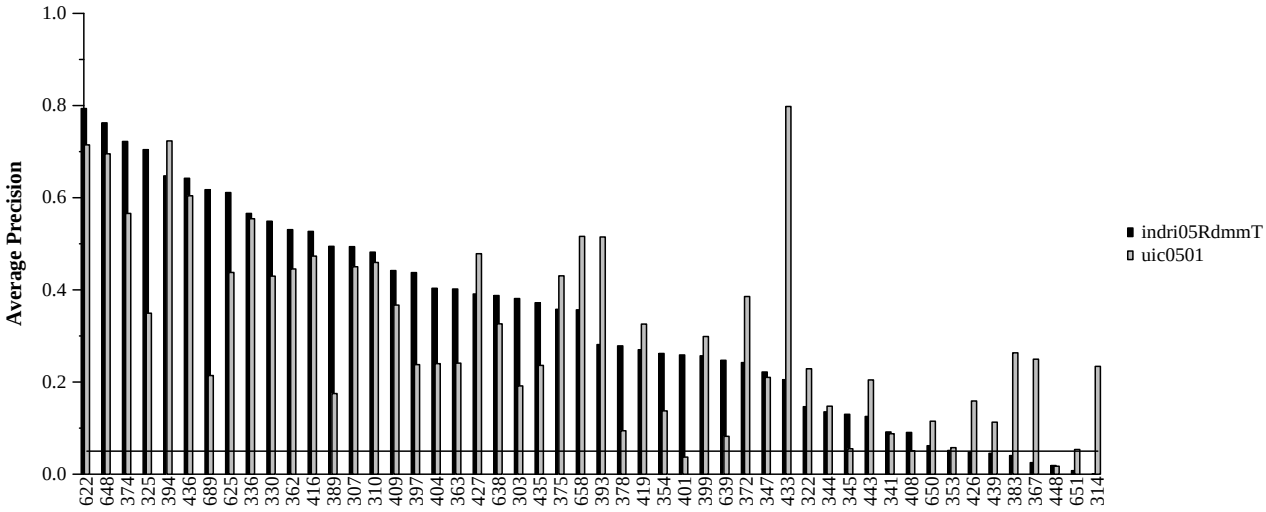

Figure 1: Per-topic average precision scores for top title-only runs. The `uic0501` run has a higher gmap score since it has fewer topics with a score less than 0.05, while the `indri05RdmmT` run has a higher average precision score for more topics and a greater MAP score.

3 The AQUAINT Test Collection

Retrieval effectiveness is in general better on the AQUAINT collection than the CD45 collection as illustrated in figure 2. The figure shows box-and-whisker plots of the average precision scores for each of the topics across the set of description-only runs submitted to TREC 2005 (top plot) and TREC 2004 (bottom plot). The line in the middle of a box indicates the median average precision score for that topic. The plots are computed over different numbers of runs (24 description-only runs in TREC 2005 vs. 30 description-only runs in TREC 2004) and in general involve different systems, but aggregate scores should be valid to compare. The majority of topics have higher medians for TREC 2005 than for TREC 2004. It is extremely unlikely that the entire set of systems that submitted description-only runs to TREC 2005 are significantly improved over TREC 2004 systems. Instead, these results remind us that topics are not inherently easy or difficult in isolation—the difficulty depends on the interaction between the information need and information source.

There are a number of differences between the ACQUAINT and CD45 test collections. The AQUAINT document set is much larger than the disks 4&5 document set: AQUAINT has more than one million documents and 3 gigabytes of text while the CD45 collection has 528,000 documents and 1904 MB of text. The AQUAINT collection contains newswire data only while the CD45 collection contains the 1994 *Federal Register* and FBIS documents. The AQUAINT collection covers a later time period. Different people assessed a given topic for the two collections. Any or all of these differences could affect retrieval effectiveness.

Earlier work in the TREC VLC track demonstrated that P(10) scores generally increase when the size of the document set increases [2]. The near doubling of the number of documents between the CD45 and AQUAINT document sets is likely a major reason for the increase in absolute scores. Aggregate statistics regarding the number of relevant documents for the two collections are not starkly different—for the AQUAINT test set there is a mean of 131.2 relevant documents per topic with a minimum number of relevant of 9 and a maximum number of relevant of 376, while the corresponding statistics for the CD45 test set are are a mean of 86.4, minimum 5, and maximum 361. But as figure 3 shows, the AQUAINT collection has many fewer topics with very small numbers of relevant documents. The figure contains a histogram of the number of relevant documents per topic for the two collections. The AQUAINT collection has only 2 topics with fewer than 20 relevant documents while the CD45 collection has 9 such topics. Good early precision scores are clearly easier to obtain when there are more relevant documents.

As figure 2 suggests, however, it is not the case that effectiveness scores simply increased by some common amount for all topics. The relative difficulty of the topics differs between the two collections. Figure 4 shows the topics sorted

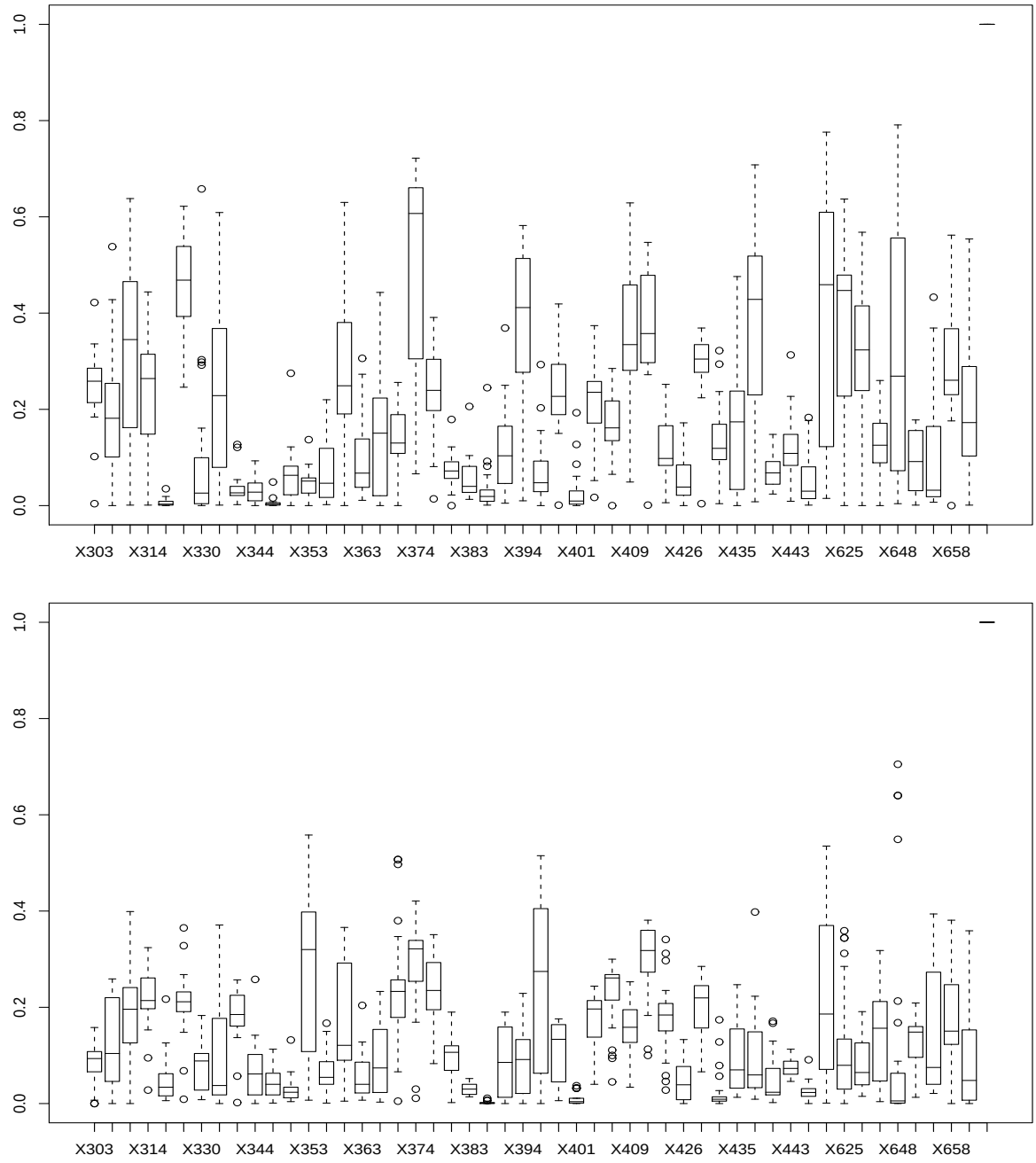

Figure 2: Box-and-whiskers plot of average precision scores for each of the 50 TREC 2005 test topics across description-only runs submitted to TREC 2005 (top) and TREC 2004 (bottom).

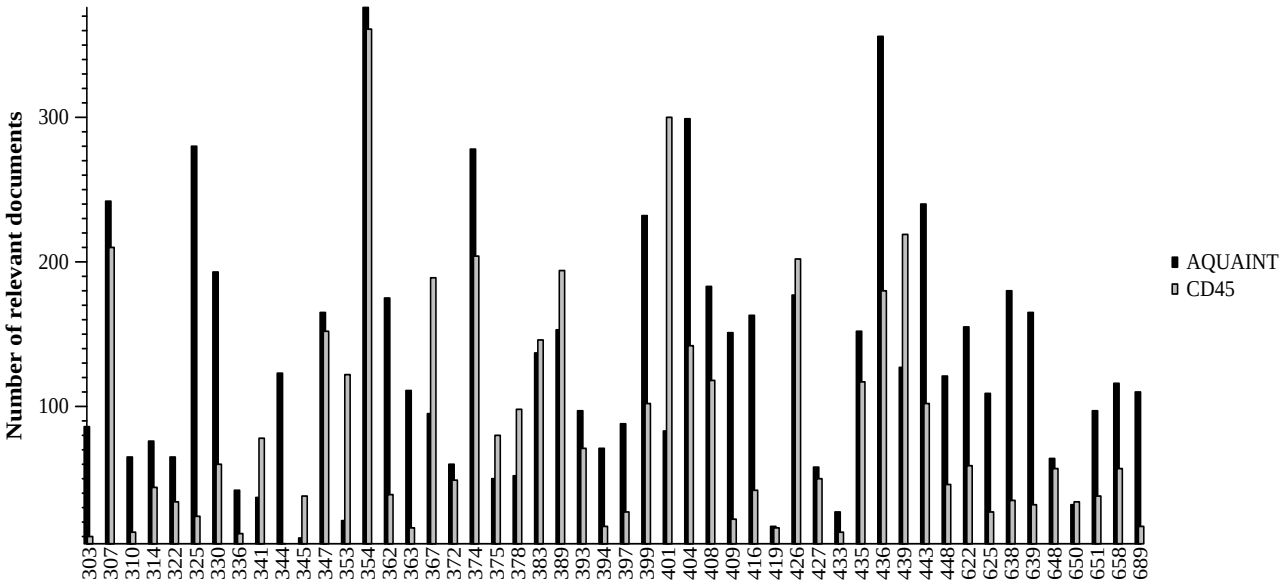

Figure 3: Number of relevant documents per topic in the TREC 2005 test set for the AQUAINT and CD45 document sets.

TREC 2005	374	325	622	625	436	394	416	310	409	638
	427	648	314	658	362	303	375	336	404	399
	435	307	689	367	408	372	639	433	443	393
	650	419	363	378	439	347	353	397	354	426
	383	651	448	344	341	330	389	401	345	322
TREC 2004	374	353	416	397	408	372	375	427	314	325
	310	404	622	341	419	639	658	409	650	399
	362	378	307	394	625	303	367	330	393	435
	651	443	638	344	354	436	689	345	336	363
	426	322	383	347	439	448	433	648	401	389
Kendall τ between rankings:		0.326								

Figure 4: Ranking of TREC 2005 test topics by decreasing median average precision score across description-only runs.

from easiest to hardest for the two collections. The difficulty of a topic is defined here as the median average precision score as computed over description-only runs submitted to either TREC 2004 and TREC 2005. The Kendall τ score between the two topic rankings is only 0.326, demonstrating that the topics have different relative difficulty on the two document sets.

The pools from which the AQUAINT test collection was created were more shallow than previous pools. Topics first used in the ad hoc tasks for TRECs 6–8 (topics 301–450) in particular had pools that were deeper and were comprised from more groups' runs than this year's pools. The expectation when the pools were formed was that the pools would be of sufficient quality because the runs contributing to the pools included both routing-type runs from the robust track and runs created after clarification form interaction from the HARD track. Unfortunately, the track results suggest that the resulting relevance judgments are dominated by a certain kind of relevant document—specifically, relevant documents that contain topic title words—and thus the AQUAINT test collection will be less reliable for future experiments where runs retrieve documents without a title-word emphasis. Note that the results of this year's HARD and robust tracks remain valid since runs from those tracks were judged.

There were two initial indications that the AQUAINT collection might be flawed. First, title-only runs are more effective than description-only runs for the AQUAINT collection, while the opposite is true for the CD45 collection. While hardly conclusive evidence of a problem, title-only queries would be expected to be better if the AQUAINT collection's shallow pools contain only easy-to-retrieve relevant documents. Second, the "optimal query" run produced by Sabir Research, a run that explicitly did not rely only on topic title words, contributed 405 unique relevant documents to the pools across the 50 topics (out of a total of $55 \times 50 = 2750$ documents contributed to the pools).

A unique relevant document is a document that was judged relevant and was contributed to the pool by exactly one group. Such documents would not have been in the pool, and therefore would be assumed irrelevant, if the one group that retrieved it had not participated in the collection building process. The difference in evaluation scores when a run is evaluated with and without the unique relevant documents from its group is used as an indication of how reusable a test collection is, since future users of the collection will not have the opportunity for their runs to be judged. The Sabir run's MAP score suffered a degradation of 23% when evaluated without its unique relevant documents, a definite warning sign.

As a result of these findings, Chris Buckley of Sabir Research and NIST examined the relevance judgments more closely. We defined a measure called titlestat as the percentage of a set of documents that a topic title word occurs in, computed as follows. For each word in the title of the current topic that is not a stop word, calculate the percentage of the set of documents, C, that contains that word, normalized by the maximum possible percentage. (The normalization is necessary because in rare cases a title word will have a collection frequency smaller than $|C|$.) Average over all title words for the topic, then average over all topics in the collection. A maximum value of 1.0 is obtained when all the documents in the set contain all topic title words; a minimum value of 0.0 means that all documents in the set contain no title words at all. Titlestat computed over the known relevant documents for the AQUAINT collection is 0.719, while the corresponding value for the CD45 collection is only 0.588. Further, the titlestat values computed over individual topics' relevant sets was greater for the AQUAINT collection than for the CD45 collection for 48 of the 50 topics.

None of the differences between the CD45 and AQUAINT document sets can plausibly explain such a change in the frequency of occurrence of topic title words in the relevant documents. If anything, title words would be expected to occur more frequently in the longer CD45 documents. Instead, the most likely explanation is that pools did not contain the documents with fewer topic title words that would have been judged relevant had they been in the pool. Topic title words are generally good descriptors of the information need stated in the topic, and retrieval systems naturally emphasize those words in their retrieval (especially when one of the mandated conditions of the track is to produce queries using only the title section!). In a collection with as many documents as the AQUAINT collection, there will be many documents containing topic title words, and these documents will fill up the pools before documents containing fewer title words will have a chance to be added.

The `sab05ror1` Sabir run further supports that contention that the majority of pool runs are dominated by documents containing topic title words while other relevant documents do exist. The titlestat computed over `sab05ror1`'s retrieved set is 0.388 while the average titlestat on the retrieved sets of the other robust track runs is 0.600. Using the unique relevant documents retrieved by the `sab05ror1` run as the set of documents the titlestat is computed over results in a value of 0.530, compared to a titlestat of 0.719 for all known relevants (including the unique relevants of the Sabir run).

Zobel demonstrated that the quality of a test collection built through pooling depends on both the diversity of the runs that contribute to the pools and the depth to which the runs are pooled [5]. In those experiments he down-sampled from existing TREC pools and saw problems only when the pools were very shallow in absolute terms. These results demonstrate how "too shallow" is relative to the document set size, a disappointing if not unexpected finding. As document collections continue to grow, traditional pooling will not be able to scale to create ever-larger reusable test collections. One of the goals of the TREC terabyte track is to examine how to build test collections for large document sets.

4 Predicting difficulty

Having a system predict whether it can effectively answer a topic is a necessary precursor to having that system modify its behavior to avoid poor performers. The difficulty prediction task was introduced into the robust track in

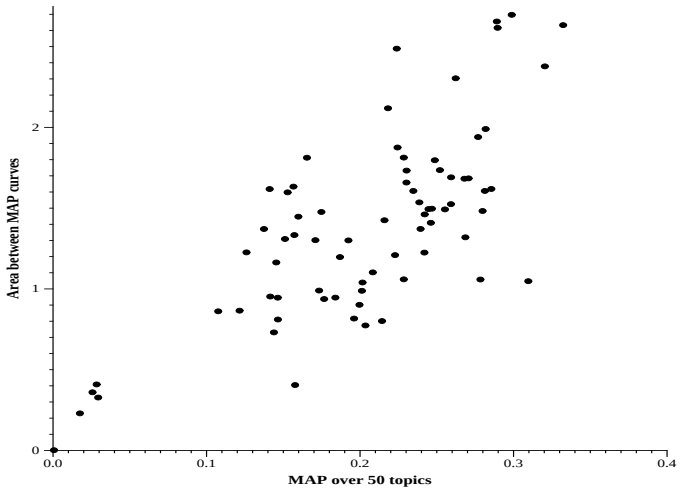

Figure 5: Scatter plot of area prediction measure vs. MAP for TREC 2005 robust track runs illustrating strong positive correlation of the scores.

TREC 2004. The task requires systems to rank the test set topics in strict order from 1 to 50 such that the topic at rank 1 is the topic the system predicted it had done best on, the topic at rank 2 is the topic the system predicted it had done next best on, etc.

Since relevance data from the CD45 collection was available for the test topics, some groups tried using that data to train difficulty predictors. These attempts were largely unsuccessful, though, since topic difficulty varied across the collections.

The difficulty-predicting task is also hampered by the lack of a suitable measure of how well a system can perform the task. Call the ranking submitted by a system its *predicted* ranking, and the topics ranked by the average precision scores obtained by the system the *actual* ranking. Clearly the quality of a system's prediction is a function of how different the predicted ranking is from the actual ranking, but this has been difficult to operationalize. The original measure used in 2004 for how the rankings differed was the Kendall τ measure between the two rankings, though it quickly became obvious that this is not a good measure for the intended goal of the predictions. The Kendall τ measure is sensitive to any change in the ranking across the entire set of topics, while the task is focused on the poor performers. A second way to measure the difference in the rankings is to look at how MAP scores change when successively greater numbers of topics are eliminated from the evaluation. In particular, compute the MAP score for a run over the best X topics where $X = 50 \ldots 25$ and the best topics are defined as the first X topics in either the predicted or actual ranking. The difference between the two curves produced using the actual ranking on the one hand and the predicted ranking on the other is the measure of how accurate the predictions are.

While the area between the two curves is a better match than Kendall τ as a quality measure of predictions for our task, it has its own faults. The biggest fault is that the area between the MAP curves is dependent on the quality of the run itself, making the area measure alone unreliable as a gauge of how good the prediction was. For example, poorly performing runs will have a small area (implying good prediction) simply because there is no room for the graphs to differ. Figure 4 shows a scatter plot of the area measure vs. the MAP score over all 50 topics for each of the runs submitted to the TREC 2005 robust track. A perfect submission would have a MAP of 1.0 and an area score of 0.0, making the lower right corner of the graph the target. Unfortunately, the strong bottom-left to top-right orientation of the plot illustrates the dependency between the two measures. Some form of normalization of the area score by the full-set MAP score may render the measure more usable.

5 Conclusion

The TREC 2005 edition of the robust retrieval track was the third, and final, running of the track in TREC. The results of the track in the various years demonstrated how optimizing average effectiveness for standard measures generally

degrades the effectiveness of poorly performing topics even further. While pseudo-relevance feedback within the target collection helps only the topics that have at least a moderate level of effectiveness to begin with, expanding queries using external corpora can be effective for poorly performing topics as well. The gmap measure introduced in the track is a stable measure that emphasizes a system's worst topics. Such an emphasis can help system builders tune their systems to avoid topics that fail completely. Gmap has been incorporated into the newest version of the trec_eval software, and will be reported for future ad hoc tasks in TREC.

References

[1] Chris Buckley. Looking at limits and tradeoffs: Sabir Research at TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2006. `http://trec.nist.gov/pubs/trec14/papers/sabir.tera.robust.qa.pdf`.

[2] David Hawking and Stephen E. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–105, 2003.

[3] Ellen M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, pages 70–79, 2005.

[4] Elad Yom-Tov, David Carmel, Adam Darlow, Dan Pelleg, Shai Errera-Yaakov, and Shai Fine. Juru at TREC 2005: Query prediction in the terabyte and the robust tracks. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2006. `http://trec.nist.gov/pubs/trec14/papers/ibm-haifa.tera.robust.pdf`.

[5] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In 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*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.