What we have learned, and not learned, from TREC

Donna Harman
National Institute of Standards and Technology

Abstract

The Text REtrieval Conference (TREC), started in 1992, is a workshop series designed to encourage research on text retrieval for realistic applications by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing results. The number of participating systems has grown from 25 in TREC-1 to 66 in TREC-8, including participants from 16 different countries. This paper examines what has collectively been learned from this massive evaluation effort, and where knowledge gaps still exist.

1 Introduction

In early 1992 the twenty-five adventurous research groups in TREC-1 undertook to scale their prototype retrieval systems from searching 2 megabytes of text to searching 2 gigabytes of text. Large disk drives were scarce in 1992, typical research computers were much slower then, and most groups made herculean efforts to finish the task. The conference itself was enlivened by people telling all the funny stories that had happened along the way. But a truly momentous event had occurred: it had been shown that the statistical methods used by these various groups were capable of handling operational amounts of text, and that research on these large test collections could lead to new insights in text retrieval.

Since then there have been seven more TREC conferences, co-sponsored by NIST and DARPA, with the latest one (TREC-8) taking place in November of 1999. The number of participating systems has grown from 25 in TREC-1 to 66 in TREC-8, including participants from 16 different countries. The diversity of the participating groups has ensured that TREC represents many different approaches to retrieval, while the emphasis on individual experiments evaluated in a common setting has proven to be a major strength. The TREC effort represents literally thousands of experiments and many person-hours. So it is very valid to ask what has been learned from all this effort. It is equally valid, and maybe more important to future research, to ask where surprisingly little progress been made, or where more investigation is clearly needed.

This paper starts with some general background on the TREC main task, the *ad hoc* task, to explain the evaluation model on which TREC is based. The next section discusses what has been learned in this main task, and where there are gaps. The final section examines the *tracks*, or additional tasks that have been performed in TREC.

2 The Ad Hoc Task

The ad hoc task investigates the performance of systems that search a static set of documents using new questions (called *topics* in TREC). This task is similar to how a researcher might use a library—the collection is known but the questions likely to be asked are not known. NIST provides the participants approximately 2 gigabytes worth of documents and a set of 50 natural language topic statements. The participants produce a set of *queries* from the topic statements using either automatic or manual query construction, and run those queries against the documents. The output from this run, consisting of the top 1000 documents retrieved for each topic, is the official test result returned to NIST for the ad hoc task.

```
<num> Number:  409
<title> legal, Pan Am, 103

<desc> Description:
What legal actions have resulted from the destruction
of Pan Am Flight 103 over Lockerbie, Scotland, on
December 21, 1988?
<narr> Narrative:
Documents describing any charges, claims, or fines
presented to or imposed by any court or tribunal are
relevant, but documents that discuss charges made in
diplomatic jousting are not relevant.
```

Figure 1: A sample TREC-8 topic.

3 The Ad Hoc Test Collections

The creation of a set of large, unbiased test collections has been critical to the success of TREC. Like most traditional retrieval collections, there are three distinct parts to these collections: the documents, the topics, and the relevance judgments or "right answers".

There are currently five CD-ROM's of documents in the collections, with approximately 1 gigabyte of text per disk. Usually only two disks (2 gigabytes of data) are used for each TREC. These documents consist primarily of news articles (including the *Wall Street Journal*, the AP newswire, the *Financial Times*, the *San Jose Mercury News* , and the *Los Angeles Times*) and government documents (the *Federal Register*, the *Congressional Record*, patent applications, and abstracts from the US Department of Energy publications). The document selection criteria has been based on availability and also on having a wide variety of document characteristics such as a broad range of document lengths, a varied writing style and vocabulary, and different levels of editing.

The topics used in TREC have consistently been the most difficult part of the test collection to control. In designing the TREC task, there was a conscious decision made to provide "user need" statements rather than the more traditional queries. Starting in TREC-3, different lengths (and component parts) of topics have been used in each TREC to explore the effects of topic length, such as the use of short titles vs sentence length descriptions vs full user narratives (which include all parts of the topic). A sample TREC-8 topic is shown in Figure 3.

Starting in TREC-3, topics have generally been created by the same person (or *assessor*) who performed the relevance assessments for that topic. Each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the document collection (looking at approximately 100 documents per topic) to estimate the likely number of relevant documents per candidate topic. NIST personnel select the final 50 topics from among the candidates based on having a range of estimated number of relevant documents and balancing the load across assessors.

The relevance judgments are also of critical importance to a test collection. For each topic it is necessary to compile a list of relevant documents; hopefully as comprehensive a list as possible. TREC uses a sampling method known as pooling (Sparck Jones & van Rijsbergen, 1975) that takes the top 100 documents retrieved by each system for a given topic and merges them into a pool for relevance assessment. This is a valid sampling method since all the systems use ranked retrieval methods, with those documents most likely to be relevant returned first. The merged list of results is then shown to the human assessors, with each topic judged by a single assessor to insure the best consistency of judgment. For TREC-8 there was an average of 1736 documents judged per topic, with about 5% or 94 of these found relevant.

4 What have we learned in the Ad Hoc testing

The basic TREC ad hoc paradigm has presented three major challenges to search engine technology from the beginning. The first is the vast scale-up in terms of number of documents to be searched, from several megabytes of documents to 2 gigabytes of documents. This system engineering problem occupied most systems in TREC-1, and has continued to be the initial work for most new groups entering TREC. The second challenge is that these documents are mostly full-text and therefore much longer than most algorithms in TREC-1 were designed to handle. The document length issue has resulted in major changes to the basic term weighting algorithms, starting in TREC-2. The third challenge has been the idea that a test question or topic contains multiple fields, each representing either facets of a user's question or the various lengths of text that question could be represented in. The particular fields, and the lengths of these fields, have changed across the various TRECs, resulting in different research issues as the basic environment has changed.

Table 1 summarizes the ad hoc task results from TREC-2 to TREC-7. It illustrates some of the common issues that have affected all groups, and also shows the initial use and subsequent spread of some of the now-standard techniques that have emerged from TREC.

Five different research areas are shown in the table, with research in many of these areas triggered by changes in the TREC evaluation environment. For example, the use of subdocuments or passages was caused by the initial difficulties in handling full text documents, particularly excessively long ones. The use of better term weighting, including correct length normalization procedures, made this technique less used in TREC's 4 and 5, but it resurfaced in TREC-6 to facilitate better input to relevance feedback.

The first research area shown in the table is that of term weighting. Most of the initial participants in TREC used term weighting that had been developed and tested on very small test collections with short documents (abstracts). Many of these algorithms were modified to handle longer documents in simple ways, however some algorithms were not amenable to this approach, resulting in some new fundamental research. The group from the Okapi system, City University, London (Robertson, Walker, Hancock-Beaulieu, & Gatford, 1994) decided to experiment with a completely new term weighting algorithm that was both theoretically and practically based on term distribution within longer documents. By TREC-3 this algorithm had been "perfected" into the BM25 algorithm. Continuing along this same row in table 1, three other systems (the SMART system from Cornell (Singhal, Buckley, & Mitra, 1996), the PIRCS system from CUNY (Kwok, 1996) and the INQUERY system from the University of Massachusetts (Allan, Ballesteros, Callan, Croft, & Lu, 1996) changed their weighting algorithms in TREC-4 based on analysis comparing their old algorithms to the new BM25 algorithm. By TREC-5 and TREC-6, many of the groups had adopted these new weighting algorithms, with the early adopters being those systems with similar structural models.

It could be expected that 6 years of term weighting experiments would lead to a convergence of the algorithms. However, a snapshot of the top 8 systems in TREC-7 shows that these systems are derived from many models and use differing term weighting algorithms and similarity measures. Of particular note here is that new models and term weighting algorithms are still being developed (Hiemstra & Kraaij, 1999; Miller, Leek, & Schwartz, 1999), and that these are competitive with the more established methods.

The second new technique started back in TREC-2 (the second line of table 1) was the use of smaller sections of documents, called subdocuments, by the PIRCS system at City University of New York (Kwok & Grunfeld, 1994). This issue was forced by the difficulty of using the PIRCS spreading activation model for documents having a wide variety of lengths. By TREC-3 many of the groups were also using subdocuments, or passages, to help with retrieval. But, as mentioned before, TREC's 4 and 5 saw far less use of this technique as many groups dropped the use of passages due to minimal added improvements in performance. There was a revival in TREC-6 of the use of passages, but generally only for specific uses, such as topic expansion. This diverse use of passages has continued in TREC, with passages clearly becoming one of the standard tools for experimentation.

The query expansion/modification techniques shown in the third and fourth lines of the table 1 were started when the topics were substantially shortened in TREC-3. In the search for some technique that would automatically expand the topic, several groups revived an old one of assuming that the top retrieved documents

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7
Term weighting	baseline for most systems beginning of Okapi weighting experiments	Okapi perfects BM25 algorithm	new weighting algorithms in SMART, INQUERY, and PIRCS systems	use of Okapi / SMART weighting algorithms by other groups	adaptations of Okapi / SMART algorithms in most systems	new retrieval models by TNO and BBN
Passages	use of subdocuments by PIRCS system	heavy use of passages / subdocuments	decline in use of passages		use of passages in relevance feedback	multiple uses of passages
Automatic query expansion		beginning of expansion using top X documents	heavy use of expansion using top X documents	beginning of more complex expansion schemes	more sophisticated expansion experiments by many groups	
Manual query modification		beginning of manual expansion using other sources	major experiments in manual editing, user-in-the-loop	extensive user-in-the-loop experiments	simpler user-specific strategies tested	
Other new areas		initial use of "data fusion"		start of more concentration on initial topic	more complex use of data fusion continued focus on initial topic, especially the title	

Table 1: Use of new techniques in the ad hoc task

are relevant, and then using them in relevance feedback. This technique, which had not worked on smaller collections, turned out to work very well in the TREC environment.

By TREC-6 almost all groups were using variations on expanding queries using information from the top retrieved documents (often called *pseudo-relevance feedback*). There are many parameters needed for success here, such as how many top documents to use for mining terms, how many terms to select, and how to weight those terms. There has been general convergence on some of these parameters in that the parameters used in TREC-7 are more similar than those used in earlier TRECs. But there continues to be further investigations by new systems adopting these techniques as there can be subtle differences between systems that strongly influence parameter selection.

Groups that build their queries manually also looked into better query expansion techniques starting in TREC-3 (see fourth line of table 1). At first these expansions involved manual editing or using other sources to manually expand the initial query. However the rules governing manual query building changed in TREC-5 to allow unrestricted interactions with the systems. This change caused a major evolution in the manual query expansion, with most participating groups not only manually expanding the initial queries, but then looking at retrieved documents in order to further expand the queries, much in the manner that users of these systems could operate. Several of the groups (Milic-Frayling, Evans, Tong, & Zhai, 1997; Strzalkowski et al., 1997) ran experiments with complex user interaction scenarios. But by TREC-6 the manual experiments moved back to the simpler scenario of having users edit the automatically-generated query, or having users select documents to be used in automatic relevance feedback.

The final line in table 1 shows some of the other areas that have seen concentrated research in the ad hoc task. Data fusion has been used in TREC by many groups in various ways, but has increased in complexity over the years. The INQUERY system from the University of Massachusetts has worked in all TREC's to automatically build more structure into their queries, based on information they have "mined" from the topics (Brown, 1995). Starting in TREC-5, there have been experiments by other groups to use more information from the initial topic, including the use of term co-occurrance and proximity as alternative methods for ranking.

The creation of two formal topic lengths in TREC-5 inspired many experiments comparing results using those different topic lengths, and the addition of a formal "title" in TREC-6 increased these investigations. It should be noted that whereas most of the best runs use the full topic, there is now a smaller performance difference between runs that use the full topic and runs that use only the title and description sections than was seen in earlier TRECs. The improvement going to the full topic was only 1% in TREC-7 for several groups. This is most likely due to improved query expansion methods, but could be due to variations across topic sets.

The graph in Figure 2 shows that retrieval effectiveness has approximately doubled since the beginning of TREC. This means, for example, that retrieval engines that could retrieve three good documents within the top ten documents retrieved in 1992 are now likely to retrieve six good documents in the top ten documents retrieved for the same search. The figure plots retrieval effectiveness for one well-known retrieval engine, the SMART system of Cornell University. The SMART system has consistently been one of the more effective systems in TREC, but other systems are comparable with it, so the graph is representative of the increase in effectiveness for the field as a whole. Researchers at Cornell ran the version of SMART used in each of the seven TREC conferences against each of the seven ad hoc test sets (Buckley, Mitra, Walz, & Cardie, 1999). Each line in the graph connects the mean average precision scores produced by each version of the system for a single test. For each test, the TREC-7 system has a markedly higher mean average precision than the TREC-1 system.

5 What we have not learned in the Ad Hoc testing

Figure 2 also shows a flattening of the improvements by TREC-7. There are many potential reasons for this, including a dilution of effort from the ad hoc task to all the new (and more exciting) tracks. However, I personally think that we may also be seeing the limits to how far we can push the current technology in the difficult test environment that systems face at TREC.

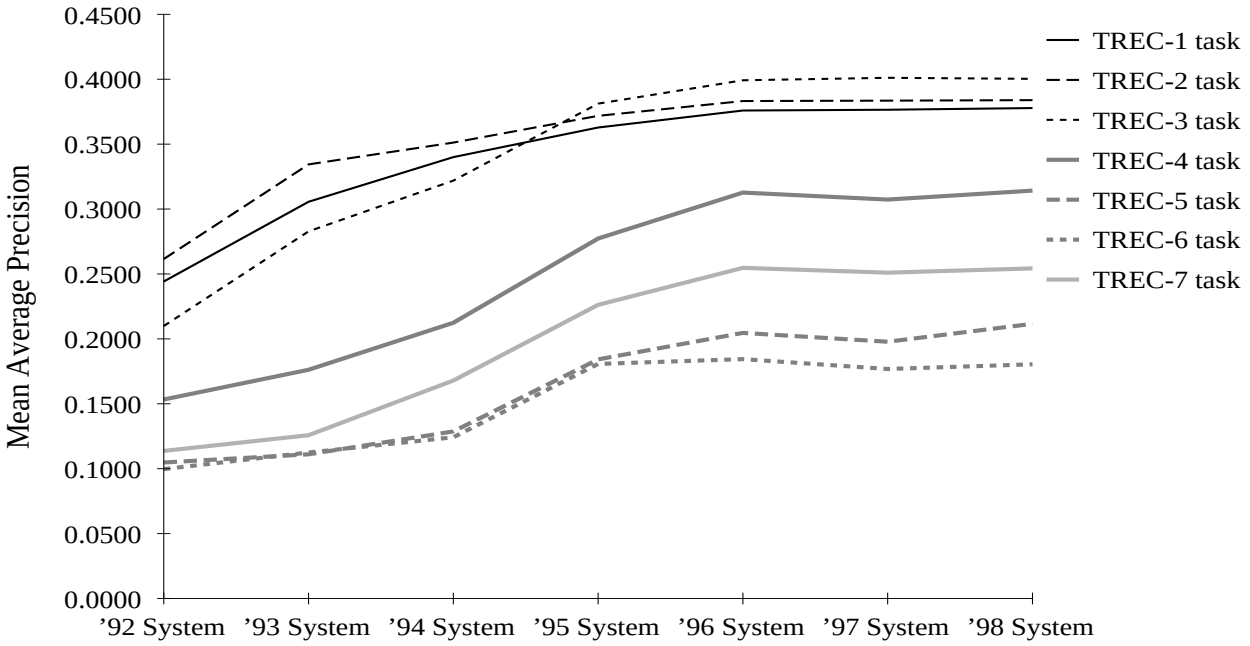

Figure 2: Retrieval effectiveness improvement for Cornell's SMART system, TREC-1 – TREC-7.

The first limitation is the issue of realistic performance expectations for the systems in lieu of the known conflicts in relevance judgements. Whereas it has been shown that disagreements about relevance do not affect the *relative* performance when comparing systems (Voorhees, 2000), they certainly affect *absolute* performance. Consistency checking has shown that the overlap of relevant documents between any two assessors is on the order of 40% on average, and this low overlap statistic is consistent with earlier studies. Voorhees converts this to a "cross-assessor" recall and precision of about 65% precision at 65% recall. This implies a practical upper bound on retrieval system performance of 65% precision at 65% recall on average, with a wide variation across topics. Note that this limitation on performance should not be considered as an evaluation problem in TREC. The ability of the systems to operate within such noisy environment is critical to operational systems and therefore an important part of TREC environment.

The second limitation is the obvious lack of user interaction. TREC was originally conceived as a test-collection-based evaluation in the Cranfield tradition. What is basically being measured are the initial results a user would see after they input a query, but before any interaction. Whereas this point of measurement is definitely important, and many users will be satisfied with these initial results, the average precision measure shown in Figure 2 has a strong recall component. The recall performance will only be further improved by user interaction and appropriate new tools. These tools are generally not being tested in TREC except for the groups operating in a manual mode or participating in the interactive track.

To get some idea of how much further improvement could be obtained from user interaction, we can compare the results from manual and automatic systems. CLARITECH Corporation has consistently done extensive user-in-the-loop experiments since TREC-5, using both complex and simple interaction models. Table 2 shows both the precision at 30 and the average precision for the Cornell system shown in Figure 2 and the best CLARITECH manual run.

The lack of user involvement is a major gap in the TREC experiments and many have urged a movement towards more interactive testing. Whereas this is clearly desirable, it is much less clear how to implement. The experience of the interactive track (section 6.5) in the design of cross-site comparisons shows the difficulties in translating the ad hoc experimental framework to include users. This would force a shift towards a less controlled experimental environment in TREC, without cross-site comparisons. It is not personally clear to me that this is desirable, since a major contribution of TREC is both the cross-site comparison

	TREC-5	TREC-6	TREC-7
Cornell (automatic)	0.29/0.21	0.32/0.21	0.39/0.27
CLARITECH (manual)	0.36/0.25	0.46/0.37	0.57/0.37

Table 2: Comparison of performance using precision at 30 and average precision for automatic and manual systems.

	Long	Desc	Title
Okapi	28	13	9
CUNY	27	10	13
Cornell	22	17	11

Table 3: Number of TREC-7 topics performing best by topic length.

and the building of re-usable test collections. This does not say that TREC evaluation should not pursue this admirable goal, but just that we need to be careful not to lose more than we gain in moving towards interactive testing.

I would like to concentrate the rest of this section on the third issue: that of the extremely wide variation in performance across topics. This variation is reflected in many ways, such as in the following examples.

1. a wide variation in which system, or even which run within a given system does best on a given topic.
2. a wide variation in the absolute performance (average precision) of the best performing system on each topic.
3. a wide variation in performance across topics of the effectiveness of particular devices such as relevance feedback.
4. a wide variation between two system variants with respect to the rank of the same retrieved document.

More analysis of these variations will be presented at the talk, but Table 3 shows an typical example of topic variation. This table shows the number of topics that had the best performance from among a group's three runs using different topic input lengths (full, short and title only). Not only is there a wide variation across topics, there is also a wide variation across systems in that topics that work best at a particular length for one group did not necessarily work best at that length for the other groups.

6 The Tracks

One of the goals of TREC is to provide a common task evaluation that allows cross-system comparisons, and this has proven to be a key strength in TREC. A second major strength is the loose definition of the ad hoc task, which allows a wide range of experiments. The addition of secondary tasks (called tracks) in TREC-4 combined these strengths by creating a common evaluation for retrieval subproblems. The tracks invigorate TREC by focusing research on new areas or particular aspects of text retrieval. To the extent that the same retrieval techniques are used for the different tasks, the tracks also validate the findings of the ad hoc task. Figure 3 shows the various tracks that have been run in TREC over the 8 year period, in addition to the number of experiments (or groups) that have participated.

Each track has a set of guidelines developed under the direction of the track coordinator and participants are free to choose which, if any, of the tracks they will join. The set of tracks, their primary goals and their major results are listed below. See the track reports in the various TREC proceedings for a more complete description of each track and its results.

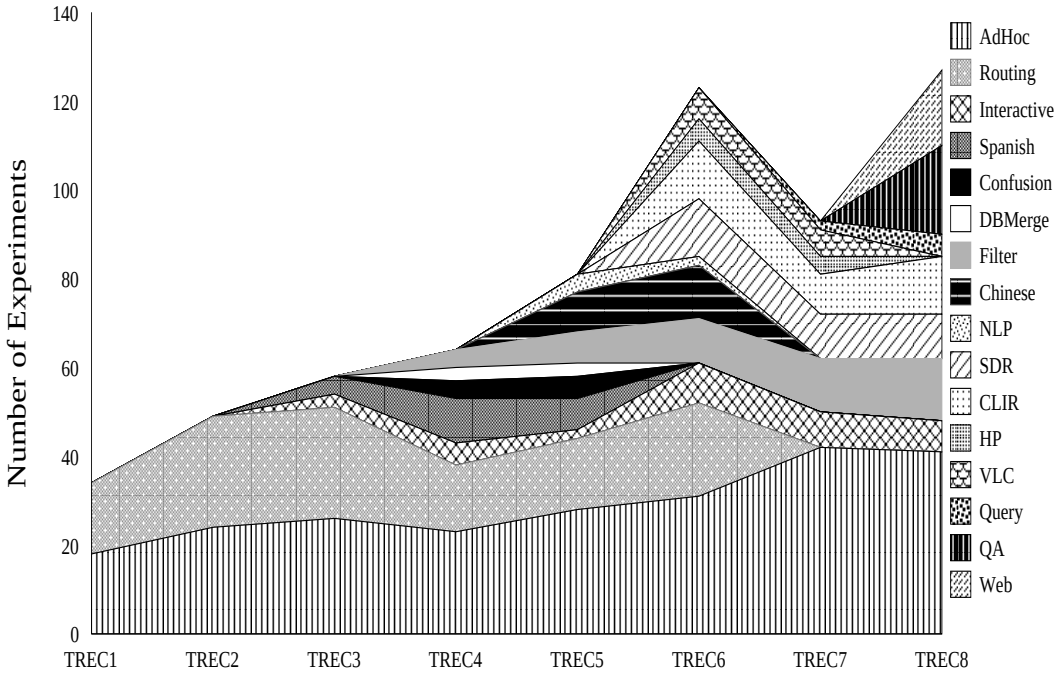

Figure 3: Number of TREC experiments by TREC task.

6.1 The Spanish and Chinese Tracks

Track reports– (Smeaton & Wilkinson, 1997; Wilkinson, 1998)

The first non-English track was started in TREC-3. Four groups worked with 25 topics in Spanish, using a document collection consisting of about 200 megabytes (58,000 documents) of a Mexican newspaper from Monterey (*El Norte*). Since there was no training data for testing (similar to the startup problems for TREC-1), the groups used simple techniques. The major result from this very preliminary experiment in a second language was the ease of porting the retrieval techniques across languages. Cornell (Buckley, Salton, Allan, & Singhal, 1995) reported that only 5 to 6 hours of system changes were necessary (beyond creation of any stemmers or stopword lists).

In TREC-4 10 groups took part, using the same document collection and 25 new topics. The final round of Spanish retrieval took place in TREC-5, again with 25 new topics and also with additional text (1994 newswire from *Agence France Presse*, including 308 megabytes or 173,950 documents). Seven groups took part in Spanish, with several of them building more elaborate procedures for testing, such as Spanish POS taggers. But in the main these did not improve performance and the major outcome of the Spanish track was that most of the techniques used in English retrieval, including the advanced ones used in the ad hoc task, can be successfully applied to Spanish.

The purpose of the Chinese track was to investigate retrieval performance for a language whose orthographics are not word-oriented. Participants performed an ad hoc search in which both the topics and the documents were in Chinese. The document set was a collection of articles selected from the *Peoples Daily* newspaper and the *Xinhua* newswire, a total of 168,811 documents in 170 megabytes. Twenty-eight topics were created for the track in TREC-5 and an additional 26 topics for TREC-6.

Nine groups submitted Chinese runs in TREC-5, and since it was the first year for Chinese in TREC, most groups concentrated on segmentation issues. In TREC-6 there were 12 participating groups, and again the majority of the experiments compared different methods of segmenting the text into retrieval features. In general, approaches that used single characters or bi-grams as features were competitive with word-based approaches and had the advantage of not requiring complicated segmentation schemes.

In terms of what was not learned, we need to examine the characteristics of the test collections. Whereas the Spanish test collection in TREC-5 was sufficiently large and diverse, the Chinese test collection had problems with similar/duplicate documents and with rather simplistic topics. The retrieval effectiveness was quite high (the median mean average precision was greater than 0.5), and was similar across systems. It was therefore difficult to distinguish more effective techniques when all techniques appear to work equally well. Without further testing, it was not possible to determine whether the TREC-6 Chinese test collection was simply "easy" or if there is something inherent in Chinese that facilitates retrieval.

This leads to an obvious general question about retrieval in various languages: are there specific characteristics of a language that need special attention, and if so, what parts of a retrieval system are language dependent or independent. Associated with this is the question of whether the improvements in performance that can be obtained using language dependent techniques are critical, particularly in regions where that language is dominant.

6.2 The Cross Language (CLIR) Track

Track reports– (Schäuble & Sheridan, 1998; Braschler, Krause, Peters, & Schäuble, 1999; Braschler, Schäuble, & Peters, 2000)

The CLIR task focused on retrieving documents that are written in different languages (English, French, German and Italian) using topics that are in one language only. This track was run in cooperation with four European institutions: University of Zurich, Switzerland (working on the French portion); Social Science Information Centre, Bonn and the University of Koblenz (working on the German portion); and CNR, Pisa, Italy (doing the Italian portion). The track was first held in TREC-6 (minus the Italian) using 25 topics created at NIST. Italian was added in TREC-7, along with 25 new topics; an additional 28 topics were built for TREC-8.

Note that for TRECs 7 and 8 these topics were created in each of the cooperating institutions in their native language. Each institution developed candidate topics such that a third of the candidates targeted international events, a third targeted items of interest in Europe generally, and a third targeted local items of interest. The intention was to create topics that had different distributions of relevant documents across languages. Twenty-eight topics were selected (7 from each native language), and relevance judgments for these topics were made separately for each language.

The track has used a document set composed of 250MB of French documents from the Swiss news agency *Schweizerische Depeschen Agentur* (SDA); 330MB of German documents from SDA plus 200MB from the newspaper *New Zurich Newspaper* (NZZ); 90MB of Italian documents from SDA; and 750MB of English documents from the AP newswire. All of the document sets contain news stories from approximately the same time period, but are not aligned or specially coordinated with one another.

The task in TREC-6 was to retrieve in the various language pairs both in a monolingual and a cross-lingual manner. Three major approaches to cross-language retrieval were represented: machine translation, where either the topics or the documents were translated into the target language; the use of machine-readable bilingual dictionaries or other existing linguistic resources; and the use of corpus resources to train or otherwise enable the cross-language retrieval mechanism. The approaches all behaved similarly in that some group obtained good cross-language performance for each method. In general, the best cross-language performance was between 50%–75% as effective as a quality monolingual run.

In TRECs 7 and 8 participants were provided with sets of topics that had translations available in English, French, German, and Italian, but had to pick one topic language to search the combined document set. This was thought to be a more realistic task than the paired language approach because it required groups to merge documents from different languages. The task of merging results across different languages turned out to be particularly difficult (like all results merging tasks). There are many unresolved issues from this track. More realistic (and larger) document collections are needed, in particular independent newspapers in each language in addition to the Swiss newswire. There needs to be better separation of the effects of

cross-language retrieval from those of merging. Both of these issues will be further examined in the new European CLEF workshop (`http://www.iei.pi.cnr.it/DELOS/CLEF`).

6.3 Routing and the Filtering Track

Track reports– (Lewis, 1997; Hull, 1998, 1999; Hull & Robertson, 2000)

The routing or filtering problem can be viewed as the inverse of the ad hoc retrieval task in that the question is assumed to be known and the document stream changes. These searches are similar to those required by news clipping services and library profiling systems. As the routing task was defined in TREC, participants used old topics with existing relevance judgments to form routing queries. These queries are then run against a previously unseen document collection to produce a ranked document list. However, real routing applications generally require a system to make a binary decision whether or not to retrieve the current document, not to form a ranking of a document set. The filtering track was started in TREC-4 to address this more difficult version of the routing task.

The question of how to evaluate filtering runs has been a major focus of the filtering track since its inception. Since filtering results are an unordered set of documents, the rank-based measures used in the ad hoc and routing tasks are not appropriate. The main approach has been to try utility functions as measures of the quality of the retrieved set—the quality is computed as a function of the benefit of retrieving a relevant document and the cost of retrieving an irrelevant document. Each TREC since TREC-4 has tried different types of utility functions in search of the elusive "ideal" measure.

There has been a slow evolution in the filtering tasks, with TREC's 7 and 8 containing three tasks of increasing difficulty (and realism). The first task was the traditional routing task. The second task was a *batch* filtering task in which systems were given topics and relevance judgments as in the routing task, and must then decide whether or not to retrieve each document in the test portion of the collection. The third task was an *adaptive* filtering task. In this task, a filtering system starts with just the query derived from the topic statement, and processes documents one at a time in date order. If the system decides to retrieve a document, it obtains the relevance judgment for it, and can modify its query as desired.

The routing task since TREC-2 has served both as an "introductory" task in TREC and as a training ground to learn new techniques for later trial in the ad hoc testing. But the issue of metrics for performance in the two filtering tasks continues to plague this track. In addition to the metrics issues, it is obvious that the filtering task (and particularly the adaptive filtering task) is a challenging problem for current systems. Indeed, when using the F1 utility measure to evaluate performance, the "baseline" system which retrieves no documents was the most effective system overall. Comparison with batch filtering results show that setting an appropriate threshhold for when to retrieve a document is a critical, and difficult, task in adaptive filtering.

6.4 The High Precision Track

Track reports– (Buckley, 1998, 1999a)

In TRECs 6 and 7 a high precision track was run. The task in the track was to retrieve fifteen (ten in TREC-6 relevant documents for a topic within five minutes (wall clock time). Users could not collaborate on a single topic, nor could the system (or user) have previous knowledge of the topic. Otherwise, the user was free to use any available resources as long as the five minute time limit was observed. The task is an abstraction of a common retrieval problem: quickly find a few good documents to get a feel for the topic area.

The major finding of the track was that retrieving 15 good documents is a simple enough task for current retrieval systems that disagreements between the searcher and the assessor regarding what constitutes a relevant document bounds performance. Note that this result correlates with the previously discussed limitations on the performance in the ad hoc track.

6.5 The Interactive Track

Track reports– (Over, 1997, 1998, 1999, 2000)

The interactive track, one of the first tracks to be started in TREC, has studied text retrieval systems in interaction with users and is interested in the process as well as the results. Two particular issues have dominated the track from the beginning. The first is overcoming the difficulty in comparing results for interactive experiments across participating sites, and the second involves the problem of selecting appropriate tasks for testing of interactive searching.

The major track result for TREC-4 was that there was no way to compare across sites in a manner similar to other TREC tracks. TRECs 5 and 6 concentrated on a new experimental design that involved comparing the particular retrieval system used at a site (an *experimental system*) to a common *control system* that was also run at each site. The direct comparison between the experimental and control systems was used to derive a measure of how much better the experimental system was than the control, independent of topic, searcher, and any other site-specific effects. Different experimental systems could then be indirectly compared across sites relative to the common control.

As a first step in analyzing the cross-site data, the best model for each site's results in terms of which factors and interactions to include was determined. Then a cross-site analysis of variance (ANOVA) was performed, which indicated that there was a significant difference between some systems. However, a multiple comparisons test (Tukey's), run to determine which systems differed, found no significant pair-wise differences. Additional experiments before and after TREC-6 addressed the effectiveness of the control (i.e., the equivalence of the direct and indirect comparison of systems) but neither confirmed nor refuted its effectiveness (Lagergren & Over, 1998; Swan & Allan, 1998). As a practical matter, it is difficult to justify the cost of adding a control system to an experimental design in the absence of clear positive evidence for its effectiveness.

TRECs 7 and 8 used a similar experimental framework, but without the requirement to use the single control system. The framework both defined a common task for participants to perform and prescribed an experimental matrix for running experiments with a minimum of 8 searchers. The search task used the title and description sections plus a special "Instances" section of eight ad hoc topics; the documents searched were the *Financial Times* collection from Disk 4. The topics each described a need for information of a particular type such that multiple distinct examples or instances of that information were contained in the document collection. The searchers job was to save documents covering as many distinct answers to the question as possible in a 15-minute time limit.

Eight groups participated in the TREC-7 track, performing a total of ten experiments. For TREC-8 there were 7 groups, and the task was basically the same, with only six topics, and a 20-minute time limit. Participants were also required to collect demographic and psychometric data from the searchers, and to report extensive data on each searcher's interactions with the search systems.

Whereas the track has abandoned efforts on cross-site comparisons, all groups ran individual experiments using the same task, and this provides a common focus for experimentation. The results of the interactive track need to be understood in the context of the particular research goals of the individual research groups. In general, however, many of the results have been inconclusive, illustrating the extreme difficulty in user testing of systems (as opposed to usability testing). Part of the problem here has been finding a suitable task or scenario for evaluation. For example, if users spend a majority of their limited time reading documents and not using a browsing system, then it is hard to evaluate the differences between two browsing systems.

6.6 The Natural Language Processing (NLP) Track

(Track report– (Strzalkowski & Jones, 1997))

The NLP track was started in TREC-5 to explore whether the natural language processing techniques available today are mature enough to have an impact on IR, and specifically whether they can offer an

advantage over more conventional methods. Four groups participated in the initial running of the natural language processing track; only 2 groups participated in TREC-6.

To date, specific NLP processing has not proved essential to obtaining effective retrieval in TREC. The most useful NLP techniques for text retrieval generally have been methods that recognize and normalize names and other multi-word terms. However, the TREC topics do not require processing at this level of detail. Other information seeking tasks such as fact extraction or story summarization may be a more appropriate test of current NLP technology.

6.7 The Question Answering track

(Evaluation report– (Voorhees & Tice, 2000))

TREC-8 was the first time the Question Answering track was run. The purpose of the track was to encourage research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question.

The track used the TREC-8 ad hoc document collection and 198 fact-based, short-answer questions such as "How many calories are there in a Big Mac?". Each question was guaranteed to have at least one document in the collection that answered the question. Participants were to return a ranked list of five strings per question such that each string was believed to contain an answer to the question. Depending on the run type, answer strings were limited to either 50 or 250 bytes. Human assessors read each string and made a binary decision as to whether or not the string actually did contain an answer to the question. Individual questions received a score equal to the reciprocal of the rank at which the first correct response was returned (or 0 if none of the five responses contained a correct answer). The score for a run was the mean of the individual questions' reciprocal ranks.

6.8 The Query Track

(Track report– (Buckley, 1999b, 2000))

The query track was a new track in TREC-7 whose goal was to create a large query collection. The variability in topic performance makes it impossible to reach meaningful conclusions regarding query-dependent processing strategies unless there is a very large query set—much larger than the sets of 50 topics used in the TREC collections. The query track was designed as a means for creating a large set of different queries for an existing TREC topic set.

Participants in the track created different types of queries from the topic statements and/or relevance judgments. A query of a given type was created for each of the 50 topics, forming one query set. Five different query types were used:

Very short: two or three words extracted from the topic statement.

Sentence: an English sentence based on the topic statement and the relevant documents.

Manual feedback: an English sentence based on reading 5–10 relevant documents only (by someone who doesn't know the topic statement).

Manual structured query (only for TREC-7): a manually constructed query based on the topic statement and relevant documents. The use of operators supported by the participant's system was encouraged. The TIPSTER DN2 format was used to represent the query structure.

Automatic structured query (only for TREC-7): a query constructed automatically from the topic statement and relevance judgments. TIPSTER DN2 format used to represent the query structure.

Participants exchanged the query sets they created with all other participants in the track, and all participants ran all query sets their system could support. Since the track design included all groups running all query sets, a number of direct comparisons were possible. First, participants could see how effective their system was using their own queries. Second, they could see how effective their search component was when using other queries, and finally, participants could evaluate how effective their query construction strategies were by seeing how other groups fared with their queries.

Although only 2 groups ran this track in TREC-7, 5 groups participated in TREC-8, generating over 23 different sets of queries. This has created a rich research environment. One tentative conclusion from this data is the extreme importance of good initial user queries. Further analysis of the query sets and their interaction with differing systems is liable to move us closer towards a better understanding of how to deal with topic variability.

6.9 The Confusion Track

(Track report– (Kantor & Voorhees, 2000))

A confusion (or data corruption) track was run in TREC-4 and TREC-5 to investigate the problems with using "corrupted" data such as would come from OCR or speech input. The TREC-4 track used the ad hoc task, but with data that was randomly corrupted at NIST using character deletions, substitutions, and additions to create data with a 10% and 20% error rate (i.e., 10% or 20% of the characters were affected). Note that this process is neutral in that it does not model OCR or speech input. Four groups used the baseline and 10% corruption level; only two groups tried the 20% level. As was somewhat expected, the 10% error rate did not hurt performance in general and the track results were somewhat inconclusive.

In TREC-5, the test data was actual OCR output of scanned images of the 1994 *Federal Register*. This time a new task was tried: *known-item searching*, where the goal was to retrieve a single specific document, rather than a set of relevant documents. Three versions of the documents were used, including the original documents, the documents that resulted after the originals were subjected to an optical character recognition (OCR) process with a character error rate of approximately 5%, and the documents produced through OCR with a 20% error rate (caused by down-sampling the image before doing the OCR). Five groups tried very different methods, with the group from the Swiss Federal Institute of Technology (ETH) (Ballerini et al., 1997) performing the best, using a type of expansion of possible candidate words to improve the best match score.

It was decided to migrate the confusion track to the speech area in TREC-6, where it was called the Spoken Document Retrieval (SDR) track.

6.10 The Spoken Document Retrieval (SDR) Track

Track reports– (Garofolo, Voorhees, Stanford, & Jones, 1998; Garofolo, Voorhees, Auzanne, Stanford, & Lund, 1999; Garofolo, Voorhees, & Auzanne, 2000)

The SDR track fosters research on retrieval methodologies for spoken documents (i.e., recordings of speech). It was run in TRECs 6, 7, and 8, using different document sets and different tasks.

The TREC-6 document set was a set of transcripts from 50 hours of broadcast news originally collected by the Linguistic Data Consortium for DARPA Hub-4 speech recognition evaluations (Garofolo, Fiscus, & Fisher, 1997). Three versions of the transcripts were used: a "truth" transcript that was hand-produced; a transcript produced by an IBM baseline speech recognition system; and a transcript produced by the participant's own speech recognition system. Document boundaries were given in the hand-produced transcript, and the same boundaries were used in the other two versions. While recognizing fifty hours of news presented a serious challenge to the speech systems, the resulting document set was small by retrieval standards, consisting of only 1451 stories.

Like the earlier confusion track, the task in the TREC-6 SDR track was a known-item search. Thirteen groups submitted SDR track runs and the results suggested that speech recognition and IR technologies are sufficiently advanced to do a credible job of retrieving specific documents. The better systems were able to retrieve the target document at rank 1 over 70% of the time using their own recognizer transcripts, compared to the best performance on the truth transcripts of 78.7%. Search performance was a bigger factor in the overall results than recognition accuracy, although the best results were obtained by groups that included both speech and IR experts.

The TREC-7 track implemented a full ranked retrieval task. The document collection was doubled to approximately 100 hours, representing about 3000 news stories. NIST created a set of 23 topics Different versions of the transcripts (similar to TREC-6 but with two baselines having 35% and 50% error rates) allowed participants to observe the effect of recognizer errors on their retrieval strategy. Eleven groups participated, and the results of the track displayed a linear correlation between the error rate of the recognition and a decrease in retrieval effectiveness, Not surprisingly, the correlation was stronger when recognizer error rate is computed over content-based words (e.g., named entities) rather than *all* words.

The TREC-8 track made a major jump in collection size, with more than 550 hours of news broadcasts (21,500 stories) and 50 topics. It also investigated the effects of lack of story boundaries. The ten participating groups found that the scaling up of the collection size did not affect results and that spoken document retrieval is effective even in "very large" speech corpora. The lack of story boundaries did create some problems and these will be investigated in TREC-9.

6.11 The Very Large Corpus (VLC) Track

Track reports– (Hawking & Thistlewaite, 1998; Hawking, Craswell, & Thistlewaite, 1999)

The VLC track explored how well retrieval algorithms scale to larger document collections. In contrast to the ad hoc task that uses a 2 GB document collection, the first running of the VLC track in TREC-6 used a 20 GB collection, while the TREC-7 track used a 100 GB document collection.

The TREC-6 track's corpus consisted of 7.5 million texts for a total of 20.14 GB of data, including the five TREC CDs; USENET news postings; Canadian and Australian Hansards; HTML-formatted documents including university websites, and laws and judgments from the Australian Attorney General's Department; and the *Glasgow Herald* and *Financial Times* newspapers. The TREC-6 ad hoc topics were used. with a set of relevance judgments produced by assessors at the Australian National University (ANU) for the top 20 documents retrieved (precision at 20 was the major effectiveness measures for the task). Also reported were query response time; data structure (e.g., inverted index) building time; and a cost measure of number of queries processed per minute per hardware dollar. Participants were required to submit two runs: one run over the entire VLC corpus and a second run over a baseline collection that consisted of a random 10% sample of the full corpus. The focus of the evaluation was on the ratio of the measures between the baseline and full corpus runs.

Seven groups submitted VLC track runs. All of the participants were able to complete the VLC task with the hardware available to them (i.e., no special hardware purchases were made for the track). Indeed, the major conclusion of the track was that current systems are able to obtain good (high precision) retrieval effectiveness on a 20 GB collection with reasonable resources. For example, one of the best runs, from the University of Waterloo (Cormack, Clarke, Palmer, & To, 1998), retrieved an average of 12.8 relevant documents in the top twenty processing at the rate of 2678 queries per hour using a cluster of four commodity PCs.

The TREC-7 collection consisted of World Wide Web data that was collected by the Internet Archive (`http://www.archive.org`). A 100 GB sample of this data was used, along with TREC-7 ad hoc topics (and relevance judgments by ANU as before). To more accurately measure the effect size has on the retrieval systems used by the participants, the track provided 3 collections: the original 100 GB collections plus 1% and 10% subsamples.

Seven groups participated in the TREC-7 VLC track, with six groups processing the entire 100GB corpus. The track demonstrated that processing a 100GB corpus is well within the capabilities of today's retrieval systems.

6.12 The Web track

Track reports– (Hawking, Voorhees, Craswell, & Bailey, 2000)

Like the Question Answering track, the web track was a new track for TREC-8. The purpose of the track was to provide the infrastructure required to reliably evaluate new search techniques and to perform repeatable experiments in the context of the World Wide Web. The track used a frozen snapshot of the web as its document collection. This collection, known as the VLC2 collection and used in last year's Very Large Corpus track (Hawking et al., 1999) is over 100 gigabytes and represents some 18.5 million web pages.

The track defined two subtasks, the small web and the large web tasks, based on the amount of the web data used. The small web task used a 2 gigabyte, 250,000 document subset of the VLC2 collection, while the large web task used the entire collection.

The focus of the small web task was on answering two questions:

- Do the best methods used in the TREC ad hoc task also work best on web data? and
- Can link information in web data be used to obtain more effective search rankings than can be obtained using page content alone?

The small web task was exactly the same as the TREC-8 ad hoc task except that the web documents were searched instead of the documents on Disks 4 and 5. The NIST relevance assessors who judged the ad hoc pools also judged the corresponding small web pools. Results from the 17 participants were somewhat inconclusive in that little difference was seen between performance of special web searching techniques and "normal" ad hoc techniques. There were some major questions about the effects of the size and structure of the 2-gigabyte web data and this has lead to a much larger (10 gigabytes) web track in TREC-9 that has a more controlled structure design.

The large web task was also a traditional ad hoc retrieval task. In this case, however, the full VLC2 collection of documents was searched using 10,000 queries extracted from logs from the Alta Vista and Electric Monk search engines. Eight participants submitted the top 20 documents for all 10,000 queries to the Cooperative Research Centre for Advanced Computational Systems (ACSys). ACSys selected 50 of the 10,000 queries to judge, and judged all 20 documents for each run for those 50 queries. Results again verified the ability of these systems to handle the large amount of web data.

6.13 The Database Merging Track

(Track report– (Voorhees, 1997))

The database merging track had the goal of investigating techniques for merging results from the various TREC subcollections (as opposed to treating the collections as a single entity). This type of investigation is important for real-world collections, and also to allow researchers to take advantage of possible variations in retrieval techniques for heterogeneous collections.

The track was started in TREC-4, with 3 participating groups. running the ad hoc topics separately on each of the 10 subcollections, merging the results, and then submitting these, along with a baseline run treating the subcollections as a single collection. The 10 subcollections were defined corresponding to the various dates of the data, i.e., the three different years of the *Wall Street Journal*, the two different years of the *AP* newswire, the two sets of Ziff documents (one on each disk), and the three single subcollections (the *Federal Register*, the *San Jose Mercury News*, and the U.S. Patents).

If results are produced without use of collection information, then the merging process is trivial. Certainly this is one method of handling the problems of merging results from different databases. However this precludes using information about the collection to modify the various algorithms in the search engine, and, even more importantly, it does not deal with the issue about which collection to select. An implied question in this track was the hypothesis that one might want to bias searching towards certain collections.

There was a second running of the database merging track in TREC-5, again with only three groups participating. This time the data was split into many more (98) databases, to allow testing of database selection methods. Unfortunately this proved to be a high-overhead track and thus did not attract much participation despite a general interest in the problem. The track has not been run since TREC-5, and remains an area with many open research issues.

7 Conclusions

It is difficult to summarize all the TREC results from eight years of TREC work, comprising several thousand major experiments conducted by all the participating systems. This paper serves only as an introduction to TREC. Each of the conferences has produced a proceedings containing papers from all the participating groups giving the details of these experiments and these proceedings have an overview of the work, containing some highlights of what was accomplished.

There are many new areas for exploration, including video retrieval, retrieval in Arabic or Hindi, retrieval of domain-specific information, and retrieval of documents that contain significant amounts of usable metadata. There are also innumerable challenges to the research community in terms of filling in the gaps where more analysis is needed. Karen Sparck Jones (Sparck Jones, 1995, 2000) has raised many interesting questions in her analysis of TREC results. Additionally she has worked with the Okapi group in a series of grid experiments to consolidate ad hoc experimental results (Sparck Jones, Walker, & S.E.Robertson, 2000). Results from 7 years of TREC experiments are publically available and could serve as a rich resource for analysis on topic variability and other issues. The test collections are available not only for further system experiments but for useful "add-ons", such as user experiments with the core set of relevance judgments to examine learning effects on relevance.

TREC continues to be successful in advancing the state of the art in text retrieval, providing a forum for cross-system evaluation using common data and evaluation methods, and acting as a focal point for discussion of methodological questions on how retrieval research evaluation should be conducted. TREC-9 is currently underway!!

For more information on TREC, including how to get publications or test collections, or how to join, see the TREC web site `trec.nist.gov`. This site also has online versions of the full proceedings for each workshop.

Acknowledgments

The author gratefully acknowledges the continued support of the TREC conferences by the Intelligent Systems Office of the Defense Advanced Research Projects Agency. Thanks also go to the TREC program committee and the staff at NIST. The TREC tracks could not happen without the efforts of the track coordinators; my special thanks to them. In particular I would like to acknowledge the fact that my distillation of what has been learned in TREC is based on overviews in which Ellen Voorhees has been the primary author and on the track reports, which are authored by the track coordinators.

References

Allan, J., Ballesteros, L., Callan, J., Croft, B., & Lu, Z. (1996). Recent Experiments with INQUERY. In D. K. Harman (Ed.), (pp. 49–63). (NIST Special Publication 500-236.)

Ballerini, J.-P., Büchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P., & Wechsler, M. (1997). SPIDER Retrieval System at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 217–228). (NIST Special Publication 500-238.)

Braschler, M., Krause, J., Peters, C., & Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In E. Voorhees & D. Harman (Eds.), (p. 25-32). (NIST Special Publication 500-242.)

Braschler, M., Schäuble, P., & Peters, C. (2000). Cross-Language Information Retrieval (CLIR) Track Overview.

Brown, E. (1995). Fast evaluation of structured queries for information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 30–38).

Buckley, C. (1998). TREC-6 High-Precision Track. In E. Voorhees & D. Harman (Eds.), (p. 69-72). (NIST Special Publication 500-240.)

Buckley, C. (1999a). TREC-7 High-Precision Track. In E. Voorhees & D. Harman (Eds.), (p. 57-64). (NIST Special Publication 500-242.)

Buckley, C. (1999b). TREC-7 Query Track. In E. Voorhees & D. Harman (Eds.), (p. 73-78). (NIST Special Publication 500-242.)

Buckley, C. (2000). TREC-8 Query Track.

Buckley, C., Mitra, M., Walz, J., & Cardie, C. (1999). SMART High Precision: TREC 7. In E. Voorhees & D. Harman (Eds.), (p. 285-298). (NIST Special Publication 500-242.)

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic Query Expansion Using SMART: TREC-3. In D. K. Harman (Ed.), (pp. 69–80). (NIST Special Publication 500-225.)

Cormack, G. V., Clarke, C. L., Palmer, C. R., & To, S. S. L. (1998). Passage-based refinement (MultiText experiments for TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 303–319). (NIST Special Publication 500-240.)

Garofolo, J., Fiscus, J., & Fisher, W. (1997). Design and preparation of the 1996 Hub-4 broadcast news benchmark test corpora. In *Proceedings of the DARPA speech recognition workshop* (pp. 15–21).

Garofolo, J., Voorhees, E., & Auzanne, C. (2000). 1999 TREC-8 Spoken Document Retrieval Track Overview and Results.

Garofolo, J., Voorhees, E., Auzanne, C., Stanford, V., & Lund, B. (1999). 1998 TREC-7 Spoken Document Retrieval Track Overview and Results. In E. Voorhees & D. Harman (Eds.), (p. 79-90). (NIST Special Publication 500-242.)

Garofolo, J., Voorhees, E., Stanford, V., & Jones, K. S. (1998). 1997 TREC-6 Spoken Document Retrieval Track Overview and Results. In E. Voorhees & D. Harman (Eds.), (p. 83-92). (NIST Special Publication 500-240.)

Harman, D. K. (Ed.). (1994, March). *Proceedings of the second text REtrieval conference (TREC-2).* (NIST Special Publication 500-215.)

Hawking, D., Craswell, N., & Thistlewaite, P. (1999). Overview of TREC-7 Very Large Collection Track. In E. Voorhees & D. Harman (Eds.), (p. 91-104). (NIST Special Publication 500-242.)

Hawking, D., & Thistlewaite, P. (1998). Overview of TREC-6 Very Large Collection Track. In E. Voorhees & D. Harman (Eds.), (p. 93-106). (NIST Special Publication 500-240.)

Hawking, D., Voorhees, E., Craswell, N., & Bailey, P. (2000). Overview of TREC-8 Web Track.

Hiemstra, D., & Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and Cross-language Track. In E. Voorhees & D. Harman (Eds.), (p. 227-238). (NIST Special Publication 500-242.)

Hull, D. A. (1998). The TREC-6 Filtering Track: Description and Analysis. In E. Voorhees & D. Harman (Eds.), (p. 45-68). (NIST Special Publication 500-240.)

Hull, D. A. (1999). The TREC-7 Filtering Track: Description and Analysis. In E. Voorhees & D. Harman (Eds.), (p. 33-56). (NIST Special Publication 500-242.)

Hull, D. A., & Robertson, S. (2000). The TREC-8 Filtering Track Final Report.

Kantor, P. B., & Voorhees, E. M. (2000). The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*.

Kwok, K. (1996). A new method of weighting query terms. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 187–196).

Kwok, K., & Grunfeld, L. (1994). TREC-2 Document Retrieval Experiments using PIRCS. In D. K. Harman (Ed.), (pp. 233–242). (NIST Special Publication 500-215.)

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The trec-6 interactive track matrix experiment. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 164–172).

Lewis, D. (1997). The TREC-5 Filtering Track. In E. Voorhees & D. Harman (Eds.), (pp. 75–96). (NIST Special Publication 500-238.)

Milic-Frayling, N., Evans, D., Tong, X., & Zhai, C. (1997). CLARIT Compound Queries and Constraint-Controlled Feedback in TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 315–334). (NIST Special Publication 500-238.)

Miller, D., Leek, T., & Schwartz, R. (1999). A hidden markov model information retrieval system. In *Proceedings of the 22th annual international ACM SIGIR conference on research and development in information retrieval* (p. 214-221).

Over, P. (1997). TREC-5 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (pp. 29–56). (NIST Special Publication 500-238.)

Over, P. (1998). TREC-6 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (pp. 73–81). (NIST Special Publication 500-240.)

Over, P. (1999). TREC-7 Interactive Track Report. In E. Voorhees & D. Harman (Eds.), (p. 65-72). (NIST Special Publication 500-242.)

Over, P. (2000). TREC-8 Interactive Track Report.

Robertson, S., Walker, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi and TREC-2. In D. K. Harman (Ed.), (pp. 21–34). (NIST Special Publication 500-215.)

Schäuble, P., & Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In E. Voorhees & D. Harman (Eds.), (pp. 31–43). (NIST Special Publication 500-240.)

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 21–29).

Smeaton, A., & Wilkinson, R. (1997). Spanish and Chinese Document Retrieval in TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 57–64). (NIST Special Publication 500-238.)

Sparck Jones, K. (1995). Reflections on TREC. *Information Processing and Management, 31*(3), 291–314.

Sparck Jones, K. (2000). Further Reflections on TREC. *Information Processing and Management, 36*(1), 37–86.

Sparck Jones, K., & van Rijsbergen, C. (1975). *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.

Sparck Jones, K., Walker, S., & S.E.Robertson. (2000). A probabilistic model of information retrieval: Development and comparative experiments part i and part ii. *Information Processing and Management.*

Strzalkowski, T., & Jones, K. S. (1997). NLP Track at TREC-5. In E. Voorhees & D. Harman (Eds.), (pp. 97–102). (NIST Special Publication 500-238.)

Strzalkowski, T., Lin, F., Wang, J., Guthrie, L., Leistensnider, J., Wilding, J., Karlgren, J., Straszheim, T., & Perez-Carballo, J. (1997). Natural Language Information Retrieval: TREC-5 Report. In E. Voorhees & D. Harman (Eds.), (pp. 291–314). (NIST Special Publication 500-238.)

Swan, R., & Allan, J. (1998). Aspect windows, 3-d visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 173–181).

Voorhees, E. (1997). The TREC-5 Database Merging Track. In E. Voorhees & D. Harman (Eds.), (pp. 103–104). (NIST Special Publication 500-238.)

Voorhees, E., & Harman, D. (Eds.). (1997, November). *Proceedings of the fifth Text REtrieval Conference (TREC-5).* (NIST Special Publication 500-238.)

Voorhees, E., & Harman, D. (Eds.). (1998, August). *Proceedings of the sixth Text REtrieval Conference (TREC-6).* (NIST Special Publication 500-240.)

Voorhees, E., & Harman, D. (Eds.). (1999, April). *Proceedings of the seventh Text REtrieval Conference (TREC-7).* (NIST Special Publication 500-242.)

Voorhees, E., & Harman, D. (Eds.). (2000). *Proceedings of the eighth Text REtrieval Conference (TREC-8).*

Voorhees, E., & Tice, D. (2000). The TREC-8 Question Answering Track Evaluation.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management.*

Wilkinson, R. (1998). Chinese Document Retrieval at TREC-6. In E. Voorhees & D. Harman (Eds.), (pp. 25–30). (NIST Special Publication 500-240.)