The Text REtrieval Conference (TREC)

Donna Harman

National Institute of Standards and Technology
Gaithersburg, Md. 20899 USA

1. Introduction

Information retrieval has a long history of experimentation, building both on research with manual indexing devices stretching backinto the 1800's and also on research in various natural language processing areas in the 1950's [Sparck Jones & Willett 1997]. The Cranfield II studies [Cleverdon et al. 1966] showed that automatic indexing was comparable to manual indexing, and this and the availability of computers created a major interest in the automatic indexing and searching of texts.

The Cranfield experiments also emphasized the importance of creating test collections and using these for comparative evaluation. The Cranfield collection, created in the late 1960's, contained 1400 documents (abstracts) and 225 queries. The now familiar recall and precision metrics were used in this study and many of the principles on which we base our evaluation today were established by this early date.

Additional test collections such as CACM and NPL were subsequently developed in the 1970's, but none of these collections were much larger than the Cranfield collection, and attempts to get funding for building a larger collection were not successful. By the early 1990's, it was obvious that progress in the text retrieval field was being hampered by the lack of realistically large test collections. The DARPA TIPSTER project [Merchant 1993] provided the means for breaking this deadlock by commissioning the National Institute of Standards and Technology (NIST) to build a new test collection with 2 gigabytes of full text documents, including newspaper and newswire text. NIST in turn requested that this new test collection be made available to the broader information retrieval research community via the establishment of the Text REtrieval Conference (TREC).

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 stories that 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 six more TREC conferences, co-sponsored by NIST and DARPA, with the latest one (TREC-7) taking place in November of 1998 [Voorhees & Harman 1999]. The number of participating systems has grown from 25 in TREC-1 to 56 in TREC-7, including participants from 18 different countries, over 20 companies and most of the universities doing research in text retrieval. The diversity of the participating groups has ensured that TREC represents many different approaches to text retrieval, while the emphasis on individual experiments evaluated in a common setting has proven to be a major strength of TREC.

All the TREC conferences have centered around two main tasks based on traditional information retrieval modes: a "routing" task and an "ad hoc" task. In the routing task (not run in TREC-7) it is assumed that the same questions are always being asked, but that new data is being searched. This task is similar to that done by news clipping services or by library profiling systems. In the ad hoc task, it is assumed that new questions are being asked against a static set of data. This task is similar to how a researcher might use a library, where the collection is known but the questions likely to be asked are unknown.

In TREC the routing task is accomplished by using known topics with known "right answers" (relevant documents) for those topics, but then using new data for testing. The topics consist of natural language text describing a user's information need (see section 2 for a sample topic). The participants use the training data to produce the "best" set of queries (the actual input to the retrieval system), and these queries are then tested using new data.

The ad hoc task is represented by using known documents, but then creating new topics for testing. For both the ad hoc and routing tasks the participating groups run 50 test topics against the test documents and turn in the top ranked 1000 documents for each topic. These results are then evaluated at NIST, with appropriate performance measures (mainly recall and precision) being used for comparison of system results.

2. The 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". The test collection components are discussed briefly here -- for a more complete description of the collection, see the TREC-7 conference proceedings.

The documents in the current test collections were selected from 11 different sources: the Wall Street Journal, AP Newswires, articles from Computer Select disks (Ziff-Davis Publishing), the Federal Register, short abstracts from DOE publications, the San Jose Mercury News, U.S. Patents, the Financial Times, the Congressional Record, the Los Angeles Times, and the Foreign Broadcast Information Service. There are currently five CD-ROM's with approximately 1 gigabyte of text per disk, with only two of these used for each TREC, i.e. only 2 gigabytes of data has generally been used in the testing.

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 more traditional queries. Starting in TREC-3, different lengths (and component parts) of topics have used in each TREC to explore the effects of topic length. By TREC-5, the format had stabilized to include a title, a sentence-length description and a longer narrative. The following is one of the topics used in TREC-6.

> <num> Number: 302
> <title> Poliomyelitis and Post-Polio
>
> <desc> Description:
> Is the disease of Poliomyelitis (polio) under control in the world?
>
> <narr> Narrative:
> Relevant documents should contain data or outbreaks of the polio disease (large or small scale), medical protection against the disease, reports on what has been labeled as "post-polio" problems. Of interest would be location of the cases, how severe, as well as what is being done in the "post-polio" area.

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 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-7 there was an average of 1605 documents judged per topic, with about 6% or 93 of these found relevant.

3. TREC Tracks

Starting in TREC-4, secondary tasks (tracks) have been added to TREC. These tasks have been either related to the main tasks, or provide a more focussed implementation of those tasks. Seven tracks were run in TREC-7:

> Cross-Language -- an ad hoc task in which the documents could be in one of four different languages (English, German, French or Italian). The topics were produced in one of these languages, and translated to the other languages. The focus of the track was to use only a single language version of the topics, but then retrieve documents that pertain to that topic in all languages.
>
> Filtering -- a task similar to the routing task but one in which the systems made a binary decision as to whether the current document should be retrieved (as opposed to forming a ranked list).
>
> High Precision User Track -- an ad hoc task in which participants were given five minutes per topic to produce a retrieved set using any means desired (e.g., through user interaction, completely automatically).
>
> Interactive -- a task used to study user interaction with text retrieval systems. In TREC-6 this track examined ways of statistically comparing systems running "user-in-the-loop" experiments, and a modification of this method was used in TREC-7.
>
> Query -- this task was designed to help factor out the variation in system performance caused by variations in the input query. Systems were requested to turn in specific types of queries (very short, sentence-length, structured, etc.) and all participating groups ran all the different sets of input queries.
>
> Spoken Document Retrieval -- an ad hoc task in which the input data consisted of approximately 100 hours of speech from broadcast news (i.e. spoken documents) that had been automatically transcribed by speech recognition systems.
>
> Very Large Corpus (VLC) -- an ad hoc task that

investigated the ability of retrieval systems to handle larger amounts of data. For TREC-7 the corpus size was approximately 100 gigabytes.

Groups could participate in some or all of the tracks, in addition to running the two main tasks. Almost all the tracks had at least 10 participating groups, with new groups specifically joining TREC to tackle some of the tracks.

4. TREC Results

It is difficult to summarize all the TREC results from seven years of work, comprising thousands of major experiments conducted by all the participating systems. Each of the conferences has produced a proceedings containing papers from all the participating groups giving the details of these experiments. These proceedings additionally have an overview of the work, containing some highlights of what was accomplished.

The impact of TREC on text retrieval can be seen in three separate areas: the impact of the TREC test collections, the impact of the common evaluation forum and the conference itself, and the impact of extending traditional text retrieval research to new areas as represented by the tracks.

The test collections, currently five gigabytes in size and containing 400 topics with relevance judgments, are heavily used throughout the text retrieval community. The system results in TREC itself show both a steady progression to more complex retrieval techniques and the resulting higher performance. Existing research groups (such as the Cornell SMART system) report a doubling in performance over the seven years of TREC, whereas systems new to TREC typically double their performance in the first year as they move their techniques into current state of the art. The conference itself encourages transfer of new methods into many different types of basic search techniques.

The introduction of the tracks has led to research in new areas of text retrieval. The Chinese track (and the earlier Spanish track) were the first (large-scale) formal testing of retrieval systems in languages other than English. The Spoken Document track has joined the speech recognition community to the text retrieval community, allowing many kinds of rich interaction between these groups. The Cross-Language track, started in TREC-6, exploits the current high interest in cross-language retrieval and serves as a testing platform both in the United States and Europe.

5. The Future of TREC

TREC-8 is currently underway. Five of the TREC-7 tracks (interactive, filtering, spoken document, query, and cross-language retrieval) are being run, along with 3 new tracks. The first new track, question-answering, challenges the retrieval community to move towards actual question-answering, as opposed to document retrieval. This track distributed 200 questions, with results to consist of a 50-byte text string containing the answers to those questions. The TREC-7 Very Large Corpus track has split into a large web track (similar to the TREC-7 task) and a small web track using a 2 gigabyte subset of the web data as alternative data for the ad hoc task. This will allow testing of web-specific retrieval techniques.

TREC is likely to continue as long as there is a need in the retrieval community for such an evaluation. However, it must constantly change to reflect the current interests of the community and the needs of the sponsoring agencies. The TREC-8 question-answering track will continue to evolve, as this should be a major goal of information retrieval. Another area of change might be further moves into cross-lingual tasks, but using a wider variety of languages, such as Mandarin or Arabic. A different area of investigation could be a move into multimedia retrieval, either retrieval from video or retrieval from "mixed" sources such text with metadata tags, database input, etc.

6. More information

For more information on TREC, including how to participate and how to obtain the test collections, visit the TREC web site at:

trec.nist.gov

This site also contains online versions of the proceedings from past conferences and pointers to sources of hard-copy versions of the same.

7. References

Cleverdon C.W., Mills, J. and Keen E.M. (1966). Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test Results. Aslib Cranfield Research Project, Cranfield, England, 1966.

Merchant R. (1993). The Proceedings of the TIPSTER Text Program - Phase I. Morgan Kaufmann Publishers, Inc., San Mateo, California.

Sparck Jones K. and Willett P. (1997). Readings in Information Retrieval. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Voorhees E.M. and Harman D.K. (1999). Proceedings of the Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242.