The Text REtrieval Conference (TREC): History and Plans for TREC-9

Ellen M. Voorhees, Donna Harman
National Institute of Standards and Technology
100 Bureau Drive, STOP 8940
Gaithersburg, MD 20899-8940
{ellen.voorhees, donna.harman}@nist.gov

Text retrieval systems have historically been refined through experimentation on test collections. In 1990 NIST was asked to build a very large test collection for use in evaluation of text retrieval technology in the DARPA TIPSTER project. This collection was to be on the order of 1 million full-text documents: a magnitude about 100 times larger than any existing non-proprietary test collection. The following year NIST proposed that this collection be made available to the full research community by the formation of the Text REtrieval Conference (TREC). The first conference took place in September, 1992 with 25 participating groups including most of the leading text retrieval research groups. Although scaling their research algorithms to handle this near-operational amount of text (for 1992) was a Herculean task, groups joined TREC to get the test data and to take part in the first major cross-system search engine evaluation. Participation in TREC has since increased each year. The most recent TREC (TREC-8 held in November, 1999) had 66 participating groups representing 16 different countries.

This paper serves as a general introduction to TREC. Detailed information about TREC including the complete Proceedings for each workshop, instructions for obtaining test collections, and the Call for Participation for TREC-9 can be found on the TREC web site at `http://trec.nist.gov`. Task descriptions for some of the tasks to be done in TREC-9 follow in this issue. To participate in TREC-9, follow the instructions given in the Call for Participation.

The Ad Hoc Task

Each of eight TREC conferences has had a main task called the Ad Hoc task. The Ad Hoc task investigates the performance of systems that search a static set of documents using new questions. 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. The Ad Hoc task is also the primary mechanism by which the TREC test collections have been built.

For each TREC, NIST has provided a set of documents and a set of 50 natural language statements of information need called *topics*. Each document set consists of approximately 2GB of text representing about 800,000 documents. The documents are primarily 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 both 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.

In designing the TREC Ad Hoc task, there was a conscious decision made to provide "user need" statements rather than more traditional queries. Two major issues were involved in this decision. First, there was a desire to allow a wide range of query construction methods by keeping the topic (the need statement) distinct from the query (the actual text submitted to the system). The second issue was the ability to increase the amount of information available about each topic, in particular to include with each topic a clear statement of what makes a document relevant. The amount of information included in topic statements has

```
<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.

changed over the course of TREC, but has included three main fields (title, description, and narrative) for the past several years. An example topic statement is given in Figure 1.

Since TREC-3, the topics have 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.

TREC participants use their retrieval systems to produce rankings of the top 1000 documents for each of the 50 topics and submit their rankings to NIST. For each topic, NIST forms a pool of the unique documents in the top 100 documents across all submissions. The assessor who created the topic makes a binary decision for each document in the pool as to whether the document is relevant to the topic or not. Documents that are not in the pool (because no system retrieved the document in its top 100) are assumed to be not relevant. On average, assessors judge approximately 1500–2000 documents per topic.

NIST evaluates the rankings submitted by the participants using the relevance judgments produced from the pools and the `trec_eval` evaluation package. The evaluation package, written by Chris Buckley and available on the TREC web site, reports some 85 different numbers based on roughly 20 different measures to give a broad picture of each run. Of these numbers, the single-valued measure most frequently reported in TREC is the mean (non-interpolated) average precision.

TREC has successfully met its goals of improving the state-of-the-art in information retrieval and facilitating technology transfer. The TREC test collections are large enough that they realistically model operational settings. Indeed, most of the commercial search engines now include technology first developed in TREC. In addition, the effectiveness of the top-performing systems has doubled since the beginning of TREC. This means, for example, that retrieval engines that could retrieve three good documents in 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 TREC-9 Tracks

TREC remains a vigorous research program by focusing on different retrieval subtasks called "tracks". While the main Ad Hoc task provides an entry point for new participants and provides a baseline of retrieval performance, the tracks encourage and support research in new areas of information retrieval. To the extent the same retrieval techniques are used for the different tasks, the tracks also validate the findings of the Ad Hoc task. Tracks were first introduced into TREC in TREC-3 and have been the primary source of growth in TREC. Figure 2 shows how the number of experiments has grown in each TREC, where the set of retrieval runs submitted for a track by one group is counted as one experiment.

The set of tracks run in any particular year depends on the interests of the participants and sponsors, as well as on the suitability of the problem to the TREC environment. Some initial tracks have been discontinued because the goals of the track were met. For example, the Spanish track, an ad hoc task in which both topics and documents are in Spanish, was discontinued when the results demonstrated that

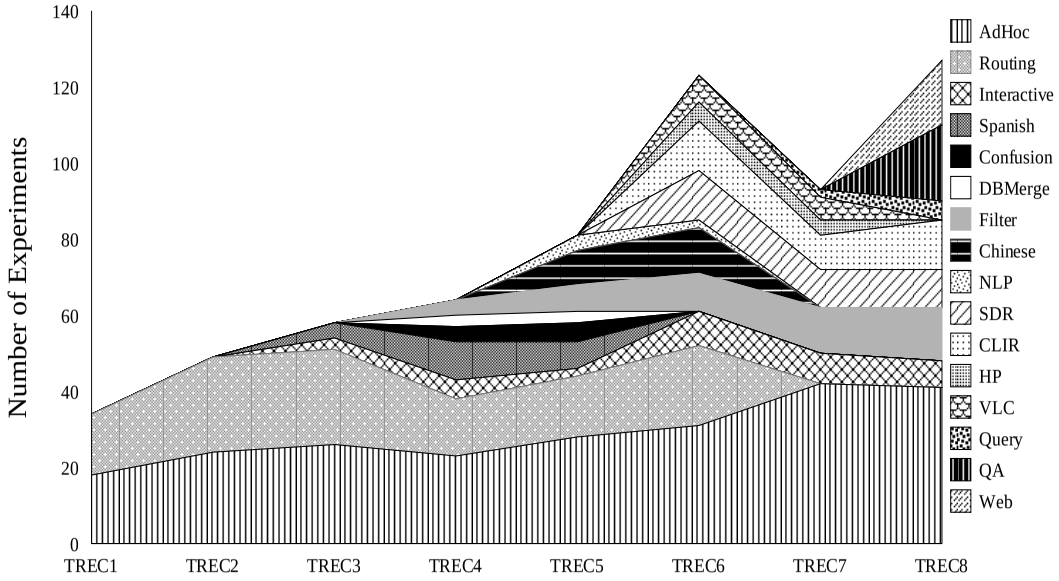

Figure 2: Number of TREC experiments by TREC task.

current retrieval systems can retrieve Spanish documents as effectively as English documents. Other tracks, such as the interactive track, have been run each year, but have changed their focus in different years.

Each track has a set of guidelines developed under the direction of the track coordinator. Participants are free to choose which of the tracks they will join. A short definition of each of the tracks to be offered in TREC-9 is given below. (There will be no main Ad Hoc task in TREC-9 so that we can concentrate efforts on building a good test collection in the web track.) More detailed task descriptions for some of the tracks also follow in this issue.

Cross-Language – a track that investigates the ability of retrieval systems to find documents that pertain to a topic regardless of the language in which the document is written. In TREC-9, the cross-language track will use Mandarin documents and English topics.

Filtering – a task in which the user's information need is stable and some relevant documents are known but there is a stream of new documents. For each document, the system must make a binary decision as to whether the document should be retrieved as opposed to forming a ranked list.

Interactive – a track studying user interaction with text retrieval systems. All participating groups follow a common experimental protocol that provides insights into user searching.

Query – a track designed to foster research on the effects of query variability and analysis on retrieval performance. Each participant constructs several different versions of existing TREC topics. All groups then run all versions of the topics.

Question Answering – a track designed to take a step closer to *information* retrieval rather than *document* retrieval. For each of a set of 500 questions, systems produce a text extract that answers the question.

Spoken Document Retrieval – A track that investigates the effects of speech recognition errors on retrieval performance. Participants retrieve segments of audio broadcasts using text topics.

Web – A track featuring ad hoc search tasks on a document set that is a snapshot of the World Wide Web.

Summary

Evaluating competing technologies on a common problem set is a powerful way to improve the state of the art and hasten technology transfer. TREC has been able to build on the text retrieval field's tradition

of experimentation to significantly improve retrieval effectiveness and extend the experimentation to new subproblems. By defining a common set of tasks, TREC focuses retrieval research on problems that have a significant impact throughout the community. The conference itself provides a forum in which researchers can efficiently learn from one another and thus facilitates technology transfer. TREC also provides a forum in which methodological issues can be raised and discussed, resulting in improved text retrieval research.

Acknowledgments

NIST gratefully acknowledges the continued support of the TREC conferences by the Intelligent Systems Office of the Defense Advanced Research Projects Agency. TREC participants commit time and resources to perform the TREC tasks and thus enable the test collecitons to be built. Thanks to the TREC program committee, which provides oversight regarding all spects of TREC. The TREC tracks could not happen without the efforts of the track coordinators; our special thanks to them.