[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

The Text REtrieval Conferences (TRECs) and the Cross-Language Track

Donna Harman

National Institute of Standards and Technology
Gaithersburg, M D. 20899, USA

Abstract

The Text REtrieval Conferences (TRECs) are a series of workshops that encourage research in information retrieval from realistic applications by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing results. The first workshop was held in November, 1992, and workshops (following evaluations) have been held yearly since then. In addition to the main task, six additional subtasks, called "tracks" allow participants to focus on particular common subproblems in retrieval, such as retrieval across languages, retrieval of speech, and retrieval at high levels of accuracy. TREC is co-sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA).

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 workshop 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 five more TREC conferences, co-sponsored
by NIST and DARPA, with the latest one (TREC-6) taking place in November of 1997. The number of participating systems has grown from 25 in TREC-1 to 51 in TREC-6, including participants from 12 different countries, 21 companies and most of the universities doing research in text retrieval (see Table 1). The diversity of the participating groups has ensured that TREC represents many different approaches to information retrieval, while the emphasis on individual experiments evaluated in a common setting has proven to be a major strength of TREC.

All six 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 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 training with known questions (called topics in TREC) and some 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. The test collection components are discussed briefly here — for a more complete description of the collection, see the TREC-6 conference proceedings [Voorhees & Harman 1998].

Apple Computer	MIT/IBM Almaden Research Center
AT & T Labs Research	NEC Corporation
Australian National University	New Mexico State U. (2 groups)
CEA (France)	NSA (Speech Research Branch)
Carnegie Mellon University	Open Text Corporation
Center for Information Research, Russia	Oregon Health Sciences U.
City University, London	Queens College, CUNY
CLARITECH Corporation	Rutgers University (two groups)
Cornell U./SaBIR Research, Inc.	Siemens AG
CSIRO (Australia)	SRI International
Daimler Benz Research Center Ulm	Swiss Federal Inst. of Tech. (ETH)
Dublin City University	TNO/U. of Twente
Duke U./U. of Colorado/Bellcore	U. of California, Berkeley
FS Consulting	U. of California, San Diego
GE Corp./Rutgers University	U. of Glasgow
George Mason U./NCR Corp.	U. of Maryland, College Park
Harris Corp.	U. of Massachusetts, Amherst
IBM T. J. Watson Research (2 groups)	U. of Montreal
ITI (Singapore)	U. of North Carolina (2 groups)
MSI/IRIT/U. Toulouse (France)	U. of Sheffield/U. of Cambridge
ISS (Singapore)	U. of Waterloo
APL, Johns Hopkins University	Verity, Inc.
Lexis-Nexis	Xerox Palo Alto Research Center

Table 1: TREC-6 Participants

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,* the U.S. Patents, the *Financial Times,* the *Congressional Record,* the *Los Angeles Times,* and material from the Foreign Broadcast Information Service. The document selection criteria is based on availability and 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.

These documents are currently stored on five CD-ROM's with approximately 1 gigabyte of text per disk. Only two gigabytes of data have generally been used in the testing for each TREC.

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.

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. This document pool is given to human assessors for making relevance judgments, with each topic judged by a single assessor to insure the best consistency of judgment. For TREC-6 there was an average of 1445 documents judged per topic, with about 6% or 92 of these found relevant.

3. TREC Results

It is difficult to summarize all the TREC results from six years of work, comprising over a thousand major experiments conducted by all the participating systems. Each of the conferences has produced a proceedings [NIST 1998] containing papers from the participating

groups giving the details of these experiments. These proceedings also contain an overview of the work, providing some highlights of what was accomplished.

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

The test collections, currently five gigabytes in size and containing 350 topics with relevance judgments, are heavily used throughout the text retrieval community. The availability of these collections has allowed existing text retrieval research groups in academia to scale their systems up to near operational dimensions; additionally it has allowed many new research groups to test radically different methods within a realistic environment, and to compare their results with those from more traditional methods. Commercial search engines use these collections as one part of their in-house performance testing, and companies such as Lexis-Nexis, CLARITECH, and Verity have reported major improvements based on TREC and its collections.

The system results in TREC itself show a steady progression to more complex retrieval techniques that result in higher performance. Existing research groups (such as the Cornell SMART system) report a doubling in performance over the six 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 workshop itself encourages transfer of new methods into many different types of basic search techniques. For example, in TREC-2 the OKAPI system from City University, London introduced some new term weighting methods. By TREC-4 these methods had been picked up by several groups, including the INQUERY system and a modified version of the Cornell SMART system. These groups in turn added to the methodology and by TREC-6 most of the other groups had incorporated these superior weighting techniques into their own systems.

Table 2 shows some of the different techniques that have come out of TREC experiments. Work in TREC-1 is not shown because it involved mostly the massive system engineering effort of scaling up to search gigabytes of data. Six different new research areas evolving within TREC are shown in Table 2, including the new term weighting techniques described earlier. Many of these areas have been 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 TRECs 4 and 5, but it resurfaced in TREC-6 to facilitate better input to relevance feedback.

Similarly the query expansion techniques shown in the third and fourth lines were started when the topics were substantially shortened in TREC-3. Groups that were building queries using automatic methods revived an old technique of assuming that the top retrieved documents 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. Groups that built their queries manually also looked into better query expansion techniques, and these techniques have evolved into the very extensive user-in-the-loop experiments seen in TREC-6.

Data fusion has been used in TREC by many groups in various ways, but has increased in complexity over the years. In TREC-6, for example, several groups such as Lexis-Nexis used multiple stages of data fusion, including merging results from different term weighting schemes and from different query expansion schemes.

The final major research area shown in this table started in TREC-5. This area is illustrated in the experiments by several groups to "mine" more information from the initial topic, rather than simply treating the topic as a bag of potential keywords for input to the system. The INQUERY system from the University of Massachusetts has worked in all TRECs to automatically build more structure into their queries, based on information they have mined from the topic. In an effort to further improve performance, more groups have experimented with other information in the initial topic, including making more use of term proximity features, clustering potential query expansion terms to maintain the initial topic balance, and looking for clues that would suggest a need for more emphasis on certain topic terms.

The main tasks in TREC have been very successful in advancing the state of the art in text retrieval. It is expected that many of the research areas shown in Table *2* will continue to attract interest in TREC-7, both as more groups adapt these methods for use in their retrieval models, and and as more experiments are done to further enhance the techniques already discovered. Additionally it is highly likely that new research areas will be investigated by many groups, leading to better tools for search engines and for potential search engine users.

4. 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. Eight tracks were run in TREC-6:

Chinese — an ad hoc task with topics and documents in Chinese.

Cross-Language — an ad hoc task in which documents were in English, German, and French. Each topic was in all three languages, and the focus of the track was to retrieve documents that pertain to the topic regardless of language.

TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
"baseline" for most systems beginning of OKAPI weighting experiments	OKAPI perfects BM25 algorithm	new SMART weighting algorithm newINQUERY weighting algorithm	use of OKAPI/SMART weighting algorithms by other groups	adaptions of of OKAPI/SMART algorithms in most systems
use of subdocuments by the PIRCS system	heavy use of passages/ subdocuments			use of passages in relevance feedback
	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
	beginning of manual expansion using other sources	major experiments in manual editing/ user-in-the-loop	continued user-in-the-loop experiments	extensive user-in-the-loop experiments
	initial use of "data fusion"	continued use of "data fusion"	continued use of "data fusion"	more complex use of "data fusion"
			beginning of more concentration on initial topic	continued focus on initial topic, including title

Table 2: Progress in Ad Hoc Retrieval Techniques

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, etc.).

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.

NLP - an ad hoc task that investigated the contribution natural language processing techniques can make to IR systems.

Spoken Document Retrieval -- a "known-item" retrieval task that used 50 hours of speech "documents" taken from news broadcasts.

Very Large Corpus (VLC) -- an ad hoc task that investigated the ability of retrieval systems to handle larger amounts of data. For TREC-6 the corpus size was approximately 20 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 joining TREC to specifically tackle some of the tracks.

The introduction of the tracks has led to research in new areas of information 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, just 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. Further Details on the Cross-Language Track

The availability of electronic data in many languages, particularly on the Web, has created increased interest in searching across languages, i.e. asking a question in one language and automatically retrieving documents in many languages. The use of automatic translation, even at only a rudimentary "gisting" level, is becoming more prevalent, allowing users to "read" documents in different languages. An example of this is the translate facility currently available on the Altavista web service. Additionally many users are able to understand documents in another language, even if they are not fluent enough in that language to construct a question. The current level of automatic translation, however, is not sufficient to suc-

cessfully serve as input to a search engine, and therein lies the challenge to the information retrieval community.

This challenge led to a workshop on cross-language retrieval at the 1996 ACM SIGIR conference. Several groups presented their current research in this area, including the use of bilingual dictionaries and parallel corpora to generate the necessary translation facility. Of particular interest at this workshop was an effort by the Swiss Federal Institute of Technology (ETH) [Sheridan & Ballerini 1996] in which they automatically built a cross-language thesaurus using "comparable" corpus as input. This corpus, consisting of Swiss newswire in three languages, is a naturally-occurring set of documents that are not translations of each other (as in parallel corpora), but are independently-produced articles dealing with the same time period. As evaluation for this research, a test collection was built consisting of roughly 100,000 news stories and 65 topics [Sheridan et al 1996]. It was the potential availability of this test collection, and the obvious interest in having a major evaluation effort for cross-language retrieval, that led to the formation of the TREC-6 cross-language track.

The TREC-6 cross-language track was a joint effort by ETH (in particular Peter Schäuble and Páraic Sheridan) and NIST. ETH was able to secure permission for use of the Swiss newswire in TREC. This newswire, from the Swiss news agency (Schweizerische Depeschen Agentur) consists of 185,099 documents in German and 141,656 in French, all taken from the years 1988-1990. Old TREC data consisting of the AP newswires on disks 1-3 (1988-1990, 242,918 documents) was used as the English documents. In addition to this data, a Swiss newspaper in German from Zurich (Neue Zuercher Zeitung (NZZ)) was used to enlarge the collection and to offer a different type of data (newspaper) and different dates (1994) for additional experimentation.

There were 25 topics created by NIST for this collection. These topics were built in 3 languages, English, French and German, by people familiar with all three languages. The topics were constructed such that they would retrieve documents from all languages, and each topic contained all three languages for use as the input topic. The task posed to the researchers was to do pairs of runs, one run in a monolingual manner, e.g. the French version of the topics against the French documents, and one run in a cross-language manner, e.g. the French version of the topics against the German documents. The evaluation was to measure the difference in retrieval performance between the monolingual baselines and the cross-lingual results.

Thirteen groups took part in the TREC-6 Cross-Language track (CLIR). Ten of these groups submitted cross-language results, with 3 others doing a only monolingual run in either French or German (this was allowed to help enlarge the pool for relevance judgments). Many very different approaches were taken, including machine translation of all documents, use of bilingual dictionaries, latent semantic indexing, and corpus-based similarity thesauri. A complete summary of the track [Schäuble & Sheridan 1998], including an overview of the results, is available in the TREC-6 proceedings [Voorhees & Harman 1998].

In general the results were very good. Cross-language retrieval performance was between 50% to 75% as good as the appropriate monolingual baseline, although there was a wide range of performance across groups. Many interesting issues were raised during the workshop, such as the difficulty in finding bilingual resources, and the problems of properly understanding issues that are more cross-cultural rather than cross-lingual.

The track will be run again in TREC-7, this time as a joint effort by four different countries. ETH will again help NIST administer the track, but three other organizations will help build the topics and create the relevance judgments. In particular, EPFL in Lausanne, Switzerland will be providing topics and relevance judgments in French; Informationszentrum Sozialwissenschaften in Bonn, Germany will be providing topics and relevance judgments in German; IEI-CNR in Pisa, Italy will be providing topics and relevance judgments in Italian; and NIST will be providing topics and relevance judgments in English. These topics (28 in all) will be translated into all four languages by EPFL, and participating groups in TREC will make runs using their chosen single topic language against the full multilingual document set. New for TREC-7 will be cross-language retrieval in Italian, again using the SDA Swiss newswire. Additionally there will be a special subtask to do cross-language retrieval on a structured text file of 31,000 documents in the field of social science. These documents are in German and will be provided by the Informationszentrum Sozialwissenschaften, along with 28 topics specifically created for this data.

CEA (France)
Cornell U./SaBIR Research, Inc.
Dublin City University
Duke U./U. of Colorado/Bellcore
MSI/IRIT/U. Toulouse (France)
New Mexico State U.
Swiss Federal Inst. of Tech. (ETH)
TNO/U. of Twente
U. of California, Berkeley
U. of Maryland, College Park
U. of Massachusetts, Amherst
U. of Montreal
Xerox Palo Alto Research Center

Table 3: TREC-6 Cross-Language Participants

6. References

NIST (1998). TREC web site-trec.nist.gov

For more information on TREC, including how to participate and how to obtain the test collections, visit the TREC web site. This site also contains online versions of the proceedings from past conferences and pointers to sources of hard-copy versions of the same.

Schäuble P. and Sheridan P. (1998). Cross-Language Information Retrieval (CLIR) Track Overview. In: *Proceedings of Sixth Text REtrieval Conference (TREC-6),* in press (and also at TREC web site).

Sheridan P. and Ballerini J.P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER system. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 58-65.

Sheridan P., Ballerini J.P. and Schäuble P. (1996). Building a Large Multilingual Test Collection from Comparable News Documents. In: *Proceedings of the Workshop on Cross-Linguistic Information Retrieval, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Voorhees E. and Harman D. (1998). Overview of the sixth Text REtrieval Conference (TREC-6). *Proceedings of Sixth Text REtrieval Conference (TREC-6),* in press (and also at TREC web site).