Overview of the Fourth Text REtrieval Conference (TREC-4)

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. INTRODUCTION

The fourth Text REtrieval Conference (TREC-4) was held at the National Institute of Standards and Technology (NIST) in November 1995. The conference, co-sponsored by DARPA and NIST, is run as a workshop for participating groups to discuss their system results on the retrieval tasks done using the TIPSTER/TREC collection. As with the first three TRECs, the goals of this workshop are:

- To encourage research in text retrieval based on large-scale test collections
- To increase communication among industry, academia and government by creating an open forum for exchange of research ideas
- To speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems
- To increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems

The number of participating systems has grown from 25 in TREC-1 to 36 in TREC-4 (see Table 1), including most of the major text retrieval software 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 within a common setting has proven to be a major strength of TREC.

The research done by the participating groups in the four TREC conferences has varied, but has followed a general pattern. TREC-1 (1992) required significant system rebuilding by most groups due to the huge increase in the size of the document collection from a traditional test collection of several megabytes in size to the 2 gigabyte TIPSTER collection. The second TREC conference (TREC-2) occurred in August of 1993, less than 10 months after the first conference. The results (using new test topics) showed significant improvements over the TREC-1 results, but should be viewed as an appropriate baseline representing the 1993 state-of-the-art retrieval techniques as scaled up to a 2 gigabyte collection.

TREC-3 [Harman 1994] provided an opportunity for more complex experimentation. The experiments included the development of automatic query expansion techniques, the use of passages or subdocuments to increase the precision of retrieval results, and the use of training information to select only the best terms for queries. Some groups explored hybrid approaches (such as the use of the Rocchio methodology in systems not using a vector space model), and others tried approaches that were radically different from their original approaches. For example, experiments in manual query expansion were done by the University of California at Berkeley and experiments in combining information from three very different retrieval techniques were done by the Swiss Federal Institute of Technology (ETH).

TREC-4 represented a continuation of many of these complex experiments, and also included a set of five focussed tasks, called tracks. This paper provides a review of the TREC-4 tasks, a very brief description of the test collection being used, and an overview of the results. The papers from the individual groups should be referred to for more details on specific system approaches.

2. THE TASKS

2.1 The Main Tasks

All four TREC conferences have centered around two main tasks based on traditional information retrieval modes: a routing task and an adhoc* 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 adhoc 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

* spelled as a single word in TREC

Australian National University	CLARITECH/Carnegie Mellon University
CITRI, Australia	City University, London
Cornell University	Department of Defense
Dublin City University	Excalibur Technologies, Inc.
FS Consulting	GE Corporate R & D/New York University
George Mason University	Georgia Institute of Technology
HNC, Inc.	Information Technology Institute
InText Systems (Australia)	Lexis-Nexis
Logicon Operating Systems	National University of Singapore
NEC Corporation	New Mexico State University
Oracle Corporation	Queens College, CUNY
Rutgers University (two groups)	Siemens Corporate Research Inc.
Swiss Federal Institute of Technology (ETH)	Universite de Neuchatel
University of California - Berkeley	University of California - Los Angeles
University of Central Florida	University of Glasgow
University of Kansas	University of Massachusetts at Amherst
University of Toronto	University of Virginia
University of Waterloo	Xerox Palo Alto Research Center

Table 1: TREC-4 Participants

library, where the collection is known, but where the questions likely to be asked are unknown.

In TREC the routing task is represented by using known topics and known relevant documents for those topics, but new data for testing. The training for this task is shown in the left-hand column of Figure 1. The participants are given a set of known (or training) topics, along with a set of documents, including known relevant documents for those topics. The topics consist of natural language text describing a user's information need (see sec. 3.3 for details). The topics are used to create a set of queries (the actual input to the retrieval system) which are then used against the training documents. This is represented by Q1 in the diagram. Many sets of Q1 queries might be built to help adjust systems to this task, to create better weighting algorithms, and in general to prepare the system for testing. The results of this training are used to create Q2, the routing queries to be used against the test documents (testing task shown on the middle column of Fig. 1).

The 50 routing topics for testing are a specific subset of the training topics (selected by NIST). A new methodology was used in TREC-4 to select the routing topics and test data. Because of difficulty in getting new data, it was decided to select the new data first, and then select topics that matched the data. The ready availability of more Federal Register documents suggested the use of topics that tended to find relevant documents in the Federal Register. Twenty-five of the routing topics were picked using this criteria. This also created a subcollection of the longer, more structured Federal Register documents for later use by the research community. The second set of 25 routing topics was selected to build a subcollection in the domain of computers. The testing documents for the computer issues were documents from the Internet, plus part of the Ziff collection.

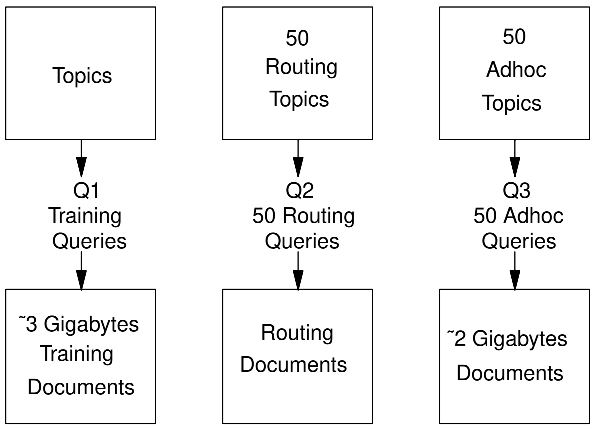

Figure 1. TREC Main Tasks.

The adhoc task is represented by new topics for known documents. This task is shown on the right-hand side of Figure 1, where the 50 new test topics are used to create Q3 as the adhoc queries for searching against the known documents. Fifty new topics (numbers 201-250) were generated for TREC-4. The known documents used in TREC-4 were on disks 2 and 3. Sections 3.2 and 3.3 give more details about the documents used and the topics that were created. The results from searches using Q2 and Q3 are the official test results sent to NIST for the routing and adhoc tasks.

In addition to clearly defining the tasks, other guidelines are provided in TREC. These guidelines deal with the methods of indexing and knowledgebase construction and with the methods of generating the queries from the supplied topics. In general, the guidelines are constructed to reflect an actual operational environment, and to allow as fair as possible separation among the diverse query construction approaches. Two generic categories of query construction were defined in TREC-4, based on the amount and kind of manual intervention used.

1. Automatic (completely automatic query construction)
2. Manual (manual query construction)

The participants were able to choose between two levels of participation: Category A, full participation, or Category B, full participation using a reduced dataset (1/4 of the full document set). Each participating group was provided the data and asked to turn in either one or two sets of results for each topic. When two sets of results were sent, they could be made using different methods of creating queries, or different methods of searching these queries. Groups could choose to do the routing task, the adhoc task, or both, and were asked to submit the top 1000 documents retrieved for each topic for evaluation.

2.2 The Tracks

One of the goals of TREC is to provide a common task evaluation that allows cross-system comparisons. This has proven to be a key strength in TREC. The second major strength is the loose definition of the two main tasks allowing a wide range of experiments. The addition of secondary tasks (tracks) in TREC-4 combines these strengths by creating a common evaluation for tasks that are either related to the main tasks, or are a more focussed implementation of those tasks.

Five formal tracks were run in TREC-4: a multilingual track, an interactive track, a database merging track, a "confusion" track, and a filtering track.

The multilingual track represents an extension of the adhoc task to a second language (Spanish). An informal Spanish test was run in TREC-3, but the data arrived late and few groups were able to take part. In TREC-4 the track was made official and 10 groups took part. There were about 200 megabytes of Spanish data (the *El Norte* newspaper from Monterey, Mexico), and 25 topics. Groups used the adhoc task guidelines, and submitted the top 1000 documents retrieved for each of the 25 Spanish topics.

The interactive track focusses the adhoc task on the process of doing searches interactively. It was felt by many groups that TREC uses evaluation for a batch retrieval environment rather than the more common interactive environments seen today. However there are few tools for evaluating interactive systems, and none that seem appropriate to TREC. The interactive track has a double goal of developing better methodologies for interactive evaluation and investigating in depth how users search the TREC topics. Eleven groups took part in this track in TREC-4. A subset of the adhoc topics was used, and many different types of experiments were run. The common thread was that all groups used the same topics, performed the same task(s), and recorded the same information about how the searches were done. Task 1 was to retrieve as many relevant documents as possible within a certain timeframe. Task 2 was to construct the best query possible.

The database merging task also represents a focussing of the adhoc task. In this case the goal was to investigate techniques for merging results from the various TREC subcollections (as opposed to treating the collections as a single entity). There were 10 subcollections 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). The 3 participating groups ran the adhoc topics separately on each of the 10 subcollections, merged the results, and submitted these results, along with a baseline run treating the subcollections as a single collection.

The "confusion" track represents an extension of the current tasks to deal with corrupted data such as would come from OCR or speech input. The track followed the adhoc task, but using only the category B data. This data 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.

The filtering track represents a variation of the routing task, and was designed to investigate concerns about the current definition of this task. It used the same topics, training documents, and test documents as the routing task. The difference was that the results submitted for the filtering runs were unranked sets of documents satisfying three "utility function" criteria. These criteria were designed to approximate a high precision run, a high recall run, and a "balanced" run. For more details on this track see the paper "The TREC-4 Filtering Track" by David Lewis (in this proceedings).

Subset of collection	WSJ (disks 1 and 2) SJMN (disk 3)	AP	ZIFF	FR (disks 1 and 2) PAT (disk 3)	DOE
Size of collection (megabytes)					
(disk 1)	270	259	245	262	186
(disk 2)	247	241	178	211	
(disk 3)	290	242	349	245	
Number of records					
(disk 1)	98,732	84,678	75,180	25,960	226,087
(disk 2)	74,520	79,919	56,920	19,860	
(disk 3)	90,257	78,321	161,021	6,711	
Median number of terms per record					
(disk 1)	182	353	181	313	82
(disk 2)	218	346	167	315	
(disk 3)	279	358	119	2896	
Average number of terms per record					
(disk 1)	329	375	412	1017	89
(disk 2)	377	370	394	1073	
(disk 3)	337	379	263	3543	

Training and Adhoc Task

Collection Source	Size in Mbytes	Terms per Record		Total Records
		Mean	Median	
Ziff (disk 3)	249	263	119	161,021
Federal Register (1994)	283	456	390	55,554
IR Digest	7	2,383	2,225	455
News Groups	237	340	235	102,598
Virtual Worlds	28	416	225	10,152

Routing Task, TREC-4

Table 2: Document Statistics

3. THE TEST COLLECTION (ENGLISH)

3.1 Introduction

Like most traditional retrieval collections, there are three distinct parts to this collection -- the documents, the questions or topics, and the relevance judgments or "right answers."

3.2 The Documents

The documents were distributed on CD-ROMs with about 1 gigabyte of data on each, compressed to fit. For TREC-4, disks 1, 2 and 3 were all available as training material (see Table 2) and disks 2 and 3 were used for the adhoc task. New data (also shown in Table 2) was used for the routing task. The following shows the actual contents of each of the three CD-ROMs (disks 1, 2, and 3).

Disk 1

- WSJ -- *Wall Street Journal* (1987, 1988, 1989)
- AP -- *AP Newswire* (1989)
- ZIFF -- Articles from *Computer Select* disks (Ziff-Davis Publishing)
- FR -- *Federal Register* (1989)
- DOE -- Short abstracts from DOE publications

Disk 2

- WSJ -- *Wall Street Journal* (1990, 1991, 1992)
- AP -- *AP Newswire* (1988)

- ZIFF -- Articles from *Computer Select* disks
- FR -- *Federal Register* (1988)

Disk 3

- SJMN -- *San Jose Mercury News* (1991)
- AP -- *AP Newswire* (1990)
- ZIFF -- Articles from *Computer Select* disks
- PAT -- U.S. Patents (1993)

Table 2 shows some basic document collection statistics. Although the collection sizes are roughly equivalent in megabytes, there is a range of document lengths across collections, from very short documents (DOE) to very long (FR). Also, the range of document lengths within a collection varies. For example, the documents from the AP are similar in length, but the WSJ, the ZIFF and especially the FR documents have much wider range of lengths within their collections.

The documents are uniformly formatted into SGML, with a DTD included for each collection to allow easy parsing.

<DOC>
<DOCNO> WSJ880406-0090 </DOCNO>
<HL> AT&T Unveils Services to Upgrade Phone Networks Under Global Plan </HL>
<AUTHOR> Janet Guyon (WSJ Staff) </AUTHOR>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
American Telephone & Telegraph Co. introduced the first of a new generation of phone services with broad
.
.
</TEXT>
</DOC>

3.3 The Topics

In designing the TREC 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 criteria make a document relevant.

The adhoc topics used in TREC-3 reflected a slight change in direction from earliers TRECs. They were not only much shorter, but also were missing the complex structure of the earlier topics. In addition to being shorter and less complex, the TREC-3 (and TREC-4) topics were written by the same group of people that did the relevance judgments (see sec. 3.4). Specifically, each of the new topics (numbers 151-250) was developed from a genuine need for information brought in by the assessors. Each assessor constructed his/her own topics from some initial statements of interest, and performed all the relevance assessments on these topics (with a few exceptions).

Sample TREC-3 topic

<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

<narr> Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

Participants in TREC-3 felt that the topics were still too long compared with what users normally submit to operational retrieval systems. Therefore the TREC-4 topics were made even shorter. Only one field was used (i.e., there is no title field and no narrative field).

Sample TREC-4 Topic

<num> Number: 207

<desc> What are the prospects of the Quebec separatists achieving independence from the rest of Canada?

Table 3 gives the average number of terms for the adhoc topics for each of the TRECs. The averages are broken down by field (title, description, narrative, and concept), with all four fields for TREC-1 and TREC-2, no concept field in TREC-3, and only a description field in TREC-4. The counts are shown both including and excluding the 23 standard stopwords used by the NIST ZPRISE system.

	Stopwords	W/O Stopwords
TREC-1 (51-100)	149	99
title	6	5
description	44	27
narrative	71	41
concepts	28	26
TREC-2 (101-150)	178	125
title	7	7
description	47	30
narrative	87	54
concepts	37	34
TREC-3 (151-200)	119	70
title	6	6
description	30	18
narrative	83	46
TREC-4 (201-250)	16	10
description	16	10

Table 3: Topic Lengths

Three different topic characteristics can be observed from this table. First, there is a length difference between the topics in TREC-1 and TREC-2. The narrative and concept fields are shorter on average for the TREC-1 topics, due to the presence of many short topics. The TREC-1 topics were produced quickly, without guidelines, by several different people, whereas the TREC-2 topics were produced by a single person. This person constructed elaborate topics that are more standarized in length, and have longer narrative and concept fields.

Second, the TREC-3 topics are not only missing the concept fields (by design), but also contain significantly shorter description fields. The TREC-3 topics were written by the 10 TREC-3 assessors who made the relevance judgments for those topics. The types of questions being asked by these assessors were less complex than the more studied questions in TREC-2, and this resulted in shorter description fields. The narrative fields are about the same length, however, probably because the TREC-2 topics were used as an example of how to write narratives. The shorter description fields, and lack of concept fields, led to topics that are about two-thirds the length of the TREC-2 topics.

The third noticeable topic characteristic is that the TREC-4 topics are much shorter than the TREC-3 topics. Not only are the narrative fields removed, but the title field is also gone. In addition, the description fields turned out to be significantly shorter going from TREC-3 to TREC-4. This was not expected, but resulted from a change in the way the topics were built. In TREC-3 the assessors brought in "seeds" of topics, i.e., ideas of issues on which to build a topic. These seeds were then expanded by each assessor, based on looking at the items that were retrieved. The resulting topics were therefore "tuned" to the collections, and were still "artificial" topics.

To avoid this tuning in TREC-4, the assessors were asked to bring in completed topics, i.e., one-sentence descriptions that were used for the actual searching. The final set of 50 topics in TREC-4 were selected by NIST from approximately 150 of these initial searches. The selection was based on how many "relevant" documents were found during sample searching. The candidate topics that retrieved too many "relevant" documents were rejected; topics were also rejected that seemed ambiguous. This different method of constructing topics resulted in the much shorter descriptions that tended to resemble the "seeds" of the TREC-3 topics rather than the TREC-3 description section. The very short topics in TREC-4 had a major effect on the results.

3.4 The Relevance Judgments

The relevance judgments are 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. All four TRECs have used the pooling method [Sparck Jones & van Rijsbergen 1975] to assemble the relevance assessments. In this method a pool of possible relevant documents is created by taking a sample of documents selected by the various participating systems. This sample is then shown to the human assessors. The particular sampling method used in TREC is to take the top 100 documents retrieved in each submitted run for a given topic and merge them into the pool for assessment. This is a valid sampling technique since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first.

A measure of the effect of pooling can be seen by examining the overlap of retrieved documents. Table 4 shows the statistics from the merging operations in the four TREC conferences. For example, in TREC-1 and TREC-2 the top 100 documents from each run (33 runs in TREC-1 and 40 runs in TREC-2) could have produced a total of 3300 and 4000 documents to be judged (for the adhoc task). The average number of documents actually judged per topic (those that were unique) was 1279 (39%) for TREC-1 and 1106 (28%) for TREC-2. Note that even though the number of runs increased in TREC-2 by more than 20%, the number of unique documents found actually dropped. The percentage of relevant documents found, however, has not changed much for the adhoc task. (The TREC-2 routing task had many fewer relevant documents because the topics were designed to be much narrower in scope.) The more accurate results going from TREC-1 to TREC-2 mean that fewer nonrelevant documents were being found by the systems. This trend continued in TREC-3, with a drop in the number of unique documents being found (particularly for the routing task) that reflects increased accuracy in rejecting nonrelevant documents. (Since only one run per system was judged, a higher percentage of documents were unique. Note a correction from the TREC-3 proceedings.)

	Adhoc		
	Possible	Actual	Relevant
TREC-1	3300	1279 (39%)	277 (22%)
TREC-2	4000	1106 (28%)	210 (19%)
TREC-3	2700	1005 (37%)	146 (15%)
TREC-4	7300	1711 (24%)	130 (7.5%)
adhoc	4000	1345 (34%)	115 (8.5%)
confusion	900	205	0
dbmerge	800	77	2
interactive	1600	84	13

	Routing		
	Possible	Actual	Relevant
TREC-1	2200	1067 (49%)	371 (35%)
TREC-2	4000	1466 (37%)	210 (14%)
TREC-3	2300	703 (31%)	146 (21%)
TREC-4	3800	957 (25%)	132 (14%)
routing	2600	930 (35%)	131 (14%)
filtering	1200	27	1 (14%)

Table 4: Overlap of Submitted Results

In TREC-4, however, the trend was reversed. Table 4 presents the statistics from the main tasks (adhoc and routing) and the associated tracks. Note that the numbers given for the tracks are in addition the main tasks, e.g., there was an average of 205 additional unique documents found from runs in the confusion track, but no new relevant documents were found. The numbers of unique documents are affected by the order of merging; that is, the average number of unique documents found by the interactive track does not count documents already found by the confusion and dbmerge tracks. The average number of relevant documents found per task or track is the actual average of unique relevant documents for that track or task.

In the case of the adhoc runs (including most of the track runs), there is a slight increase in the percentage of unique documents found. Looking at the adhoc task alone, a relatively high percentage of unique (mostly nonrelevant) documents were found. This is likely caused by the wide variety of expansion terms used by the systems to compensate for the lack of a narrative section in the topic. The additional unique documents found by the tracks appears to be characteristic of the type of methodology being tested within that track. The confusion track, where the data is corrupted for most of the runs, turned in many unique (and all nonrelevant) documents. The data merging track turned in far fewer unique documents, and some of these were additional relevant documents. The interactive track (the last to be merged) still found additional unique documents, with a relatively high percentage of those documents being relevant.

Slightly more unique documents were found for the routing task in TREC-4 than in TREC-3, probably resulting from the increased difficulty of the TREC-4 routing task. This increased difficulty stems from 1) the concentration of long Federal Register documents, which have consistently been harder to retrieve, and 2) a mismatch of the testing data to the training data (for the computer topics). Both these factors led to less accurate filtering of nonrelevant documents.

The total number of relevant documents found has dropped with each TREC, and that drop has been caused by a deliberate tightening of the topics each year to better guarantee completeness of the relevance judgments (see below for more details on this).

Evaluation of retrieval results using the assessments from this sampling method is based on the assumption that the vast majority of relevant documents have been found and that documents that have not been judged can be assumed to be not relevant. A test of this assumption was made using TREC-2 results, and again during the TREC-3 evaluation. In both cases, a second set of 100 documents was examined from each system, using only a sample of topics and systems in TREC-2, and using all topics and systems in TREC-3.

For the TREC-2 completeness tests, a median of 21 new relevant documents per topic was found (11% increase in total relevant documents). This averages to three new relevant documents found in the second 100 documents for each system, and this is a high estimate for all systems since the 7 runs sampled for additional judgments were from the better systems. Similar results were found for the more complete TREC-3 testing, with a median of 30 new relevant documents per topic for the adhoc task, and 13 new ones for the routing task. This averages to about one new relevant document per run, since 27 runs from all systems were used in the adhoc test (23 runs in the routing test). These levels of completeness are quite acceptable for this type of evaluation.

The number of new relevant documents found was shown to be correlated with the original number of relevant documents. Topics with many more relevant documents initially tend to have more new ones found, and this has led to a greater emphasis on using topics with fewer relevant documents.

In addition to the completeness issue, relevance judgments need to be checked for consistency. In each of the TREC evaluations, each topic was judged by a single assessor to ensure the best consistency of judgment. Some testing of this consistency was done after TREC-2, when a sample of the topics and documents was rejudged by a second assessor. The results showed an average agreement between the two judges of about 80%. In TREC-4 all the adhoc topics had samples rejudged by two additional assessors, with the results being about 72% agreement among all three judges, and 88% agreement between the initial judge and either one of the two additional judges. This is a remarkably high level of agreement in relevance assessment, and probably is due to the general lack of ambiguity in the topics. More consistency checking will be done before TREC-5, particularly investigating known inconsistencies and topics with major disagreements.

4. Evaluation

An important element of TREC is to provide a common evaluation forum. Standard recall/precision and recall/fallout figures have been calculated for each TREC system and are shown in Appendix A, along with some single evaluation measures for each system. A detailed explanation of the measures is also included in the appendix.

Additional data about each system was collected that describes system features and system timing, and allows some primitive comparison of the amount of effort needed to produce the results. The individual system descriptions are given in Appendix B.

5. Results

5.1 Introduction

One of the important goals of the TREC conferences is that the participating groups freely devise their own experiments within the TREC task. For some groups this means doing the routing and/or adhoc task with the goal of achieving high retrieval effectiveness performance. For other groups, however, the goals are more diverse and may mean experiments in efficiency, unusual ways of using the data, or experiments in how "users" would view the TREC paradigm.

The overview of the results discusses the effectiveness of the systems and analyzes some of the similarities and differences in the approaches that were taken. In all cases, readers are referred to the system papers in this proceedings for more details.

5.2 TREC-4 Adhoc Results

The TREC-4 adhoc evaluation used new topics (topics 201-250) against two disks of training documents (disks 2 and 3). Only 49 topics were used in the actual evaluation as topic 201 retrieved no relevant documents. A dominant feature of the adhoc task was the much shorter topics (see more on this in the discussion of the topics, section 3.3). Many groups tried their automatic query expansion methods on the shorter topics (with good success); other groups also did manual query construction experiments to contrast these methods for the very short topics.

There were 39 sets of results for adhoc evaluation in TREC-4, with 33 of them based on runs for the full data set. Of these, 14 used automatic construction of queries, and 19 used manual construction. All of the category B groups used automatic construction of the queries.

Figure 2 shows the recall/precision curves for the 6 TREC-4 groups with the highest non-interpolated average precision using automatic construction of queries. The runs are ranked by the average precision and only one run is shown per group (both official Cornell runs would have qualified for this set).

A short summary of the techniques used in these runs shows the breadth of the approaches and the changes in approach from TREC-3. For more details on the various runs and procedures, please see the cited papers in this proceedings.

CrnlEA -- Cornell University ("New Retrieval Approaches Using SMART: TREC-4" by Chris Buckley, Amit Singhal, Mandar Mitra, (Gerald Salton)) used the SMART system, but with a non-cosine length normalization method. The top 20 documents were used to locate 50 terms and 10 phrases for expansion, as contrasted with using the top 30 documents to massively expand (500 terms + 10 phrases) the topics as in TREC-3. This change in expansion techniques was mostly due to the major change in the basic algorithm. However, additional care was taken not to overexpand the very short topics.

pircs1 -- Queens College, CUNY ("TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) used a spreading activation model on subdocuments (550-word chunks). It was expected that this type of model would be particularly affected by the shorter topics, and experiments were run trying several methods of topic expansion. For this automatic run, expansion was done by selecting 50 terms from the top 40 subdocuments in addition to the terms in the original topic. Several other experiments were made using manual modifications/expansions of the topics and these are reported with the manual adhoc results.

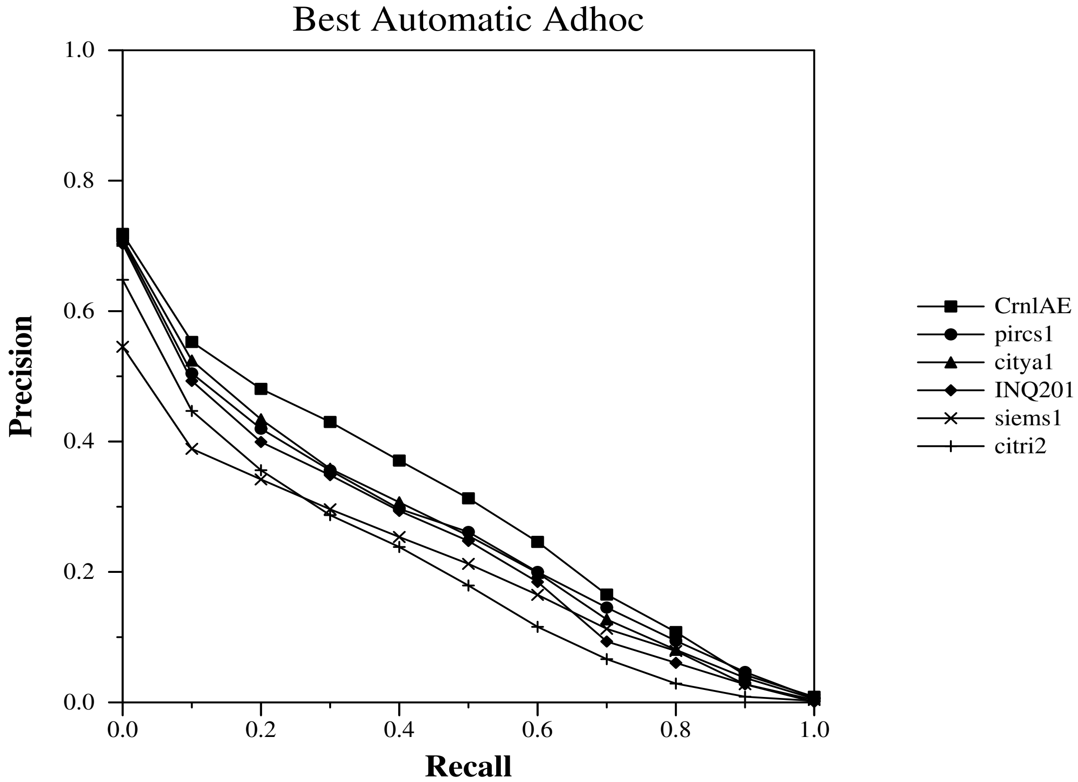

Figure 2. Best TREC-4 Automatic Adhoc Results.

citya1 -- City University, London ("Okapi at TREC-4" by S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford and A. Payne") used a probabilistic term weighting scheme similar to that used in TREC-3. An average of 20 terms were automatically selected from the top 50 documents retrieved (only initial and final passages of these documents were used for term selection). The use of passages seemed to have little effect. This run was a base run for their experiments in manual query editing.

INQ201 -- University of Massachusetts at Amherst ("Recent Experiments with INQUERY" by James Allan, Lisa Bellesteros, James P. Callan, W. Bruce Croft and Zhihong Lu) used a version of probabilistic weighting that allows easy combining of evidence (an inference net). Their basic term weighting formula underwent a major change between TREC-3 and TREC-4 that combined the TREC-3 INQUERY weighting with the OKAPI (City University) weighting. They also used passage retrieval as in TREC-3, but found it detrimental in TREC-4. The topics were expanded by 30 phrases that were automatically selected from a phrase "thesaurus" (InFinder) that had previously been built automatically from the entire corpus of documents. Expansion did not work as well as in TREC-3.

siems1 -- Siemens Corporate Research ("Siemens TREC-4 Report: Further Experiments with Database Merging" by Ellen M. Voorhees) used the SMART retrieval strategies from TREC-3 in this run (their base run for the database merging track). The standard vector normalization was used, and query expansion was done using the Rocchio method to select up to 100 terms and 10 phrases from the top 15 documents retrieved.

citri2 -- RMIT, Australia ("Similarity Measures for Short Queries" by Ross Wilkinson, Justin Zobel, and Ron Sacks-Davis) was the result of a series of investigations into similarity measures. The best of these measures combined the standard cosine measure with the OKAPI measure. No topic expansion was done for this run.

It is interesting to note that many of the systems did critical work on their term weighting/similarity measures between TREC-3 and TREC-4. Three of the top 6 runs were results of major revisions in the basic ranking algorithms, revisions that were the outcome of extensive analysis work on previous TREC results. At Cornell they investigated the problems with using the cosine normalization on the long documents in TREC. This investigation resulted in a completely new term weighting/similarity strategy that performs well for all lengths of documents. The University of Massachusetts examined the issue of dealing with terms having a high frequency in documents (which is also related to document length).

The result of their investigation was a term weighting algorithm that combined the OKAPI algorithm (City University) for high frequency terms with the old INQUERY algorithm for lower frequency terms. The work at RMIT (the *citri2* run) was part on their ongoing effort to test various term weighting schemes.

These experiments in more sophisticated term weighting and matching algorithms are yet another step in the adaptation of retrieval systems to a full-text environment. The issues of long documents, with their higher frequency terms, mean that the algorithms originally built for abstract-length documents need rethinking. This did not happen in earlier TRECs because the problem seemed less important than, for example, discovering automatic query expansion methods in TREC-3.

The dominant new feature in TREC-4 was the very short topics. These topics were much shorter than any previous TREC topics (an average reduction from 119 terms in TREC-3 to 16 terms in TREC-4). In general the participating groups took two approaches: 1) they used roughly the same techniques that they would have on the longer topics, and 2) most of them tried some investigative manual experiments. Of the 6 runs shown in Figure 2, two runs (*INQ201* and *citya1*) used a similar number and source of expansion terms as for the longer queries. The SMART group (*CrnlAE*) used many fewer terms because of their new algorithms. The *pircs1* run was a result of more expansion, but this was due to corrections of problems in TREC-3 as opposed to changes needed for the shorter topics. The run from Siemens *siems1* was made as a baseline for database merging, and therefore had less expansion. There was no expansion in the *citri21* run.

Figure 3 shows the comparison of results between TREC-3 and TREC-4 for 4 of the groups that did well in each evaluation. As expected, all groups had worse performance. The performance for City University, where similar algorithms were used in TREC-3 and TREC-4, dropped by 36%. A similar drop (34%) was true for the INQUERY results, even though the new algorithm resulted an almost 5% improvement in results (for the TREC-4 topics). Whereas the Cornell results represented a major improvement in performance over the TREC-3 algorithms, their overall performance dropped by 14%.

This points to several issues that need further investigation in TREC-5. First, experiments must still continue on the shorter topics, since this represents the typical initial input query. The results from the shorter topics may be so poor that the top documents provide misleading expansion terms. This was a major concern in TREC-3 and analysis of this issue is clearly needed. The fact that passage retrieval, which provided substantial improvement of results in TREC-3, did not help with the shorter TREC-4 topics indicates that other types of "noise" control may be needed for short topics. It may be that the statistical "clues" presented by these shorter topics are simply not enough to provide good retrieval performance and that better human-aided systems need to be tested.

However, the manual systems also suffered major drops in performance (see Figure 4). This leads to a second issue, i.e., a need for further investigation into the causes of the generally poorer performance in the TREC-4 adhoc task. It may be that the narrative section of the topic is necessary to make the intent of the user clear to both the manual query builder and the automatic systems. The fact that machine performance mirrored human performance in TREC-4 makes the decrease in automatic system performance more acceptable, but still requires further analysis into why both types of query construction were so affected by the very short topics.

Figure 5 shows the recall/precision curves for the 6 TREC-4 groups with the highest non-interpolated average precision using manual construction of queries. A short summary of the techniques used in these runs follows. Again, for more details on the various runs and procedures, see the cited papers in this proceedings.

CnQst2 -- Excalibur Corporation ("The Excalibur TREC-4 System, Preparations and Results" by Paul E. Nelson) used manually built queries. This system uses a two-level searching scheme in which the documents are first ranked via coarse-grain methods, and then the resulting subset is further refined. There are thesaurus tools available for expansion, and this run was the result of many experiments into such issues as term groupings and assignment of term strengths.

pircs2 -- Queens College, CUNY ("TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) is a manual modification of the automatic queries in *pircs1*. The modification was to replicate words (this increases the weight) and to add a few associated words (an average of 1.73 words per query or at most 3 content words). The simple replication of words led to a 12% increase in performance; adding the associated words (the *pircs2* run) upped this increase to 30% improvement over the initial automatic query.

uwgcl1 -- University of Waterloo ("Shortest Substring Ranking (MultiText Experiments for TREC-4)" by Charles L.A. Clarke, Gordon V. Cormack, and Forbes J. Burkowski) used queries that were manually built in a special query language called GCL. This query language uses Boolean operators and proximity constraints to

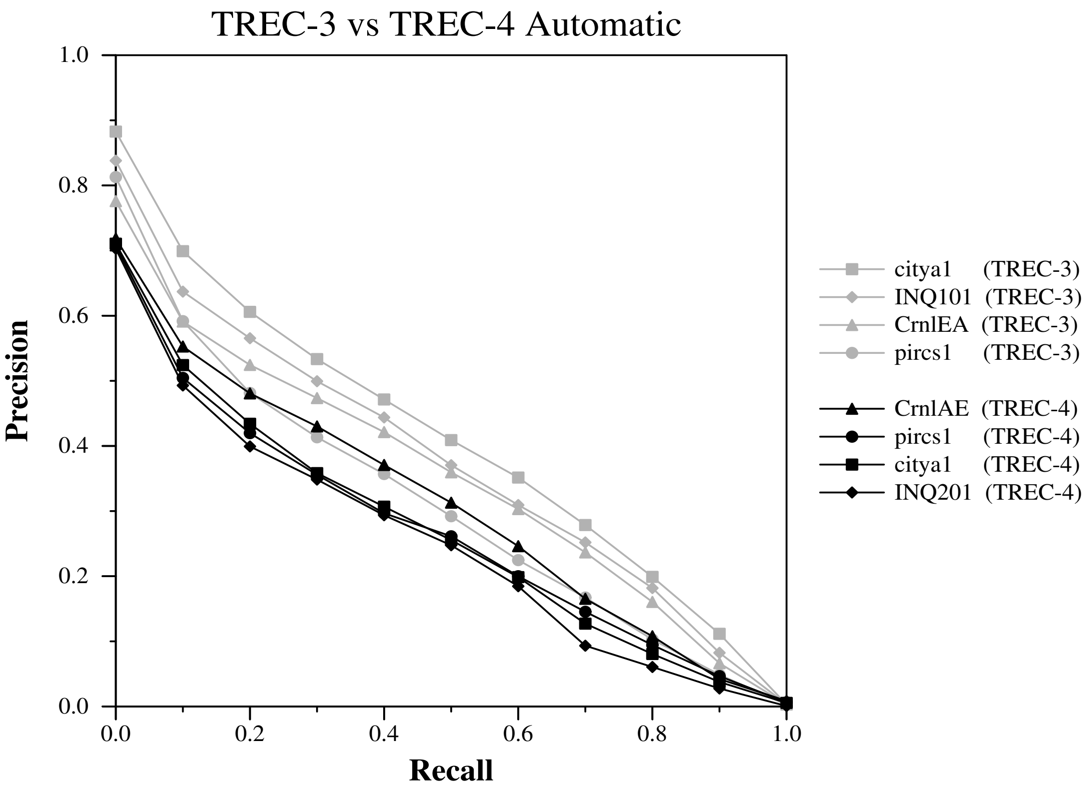

Figure 3. Comparison of Automatic Adhoc Results for TREC-3 and TREC-4.

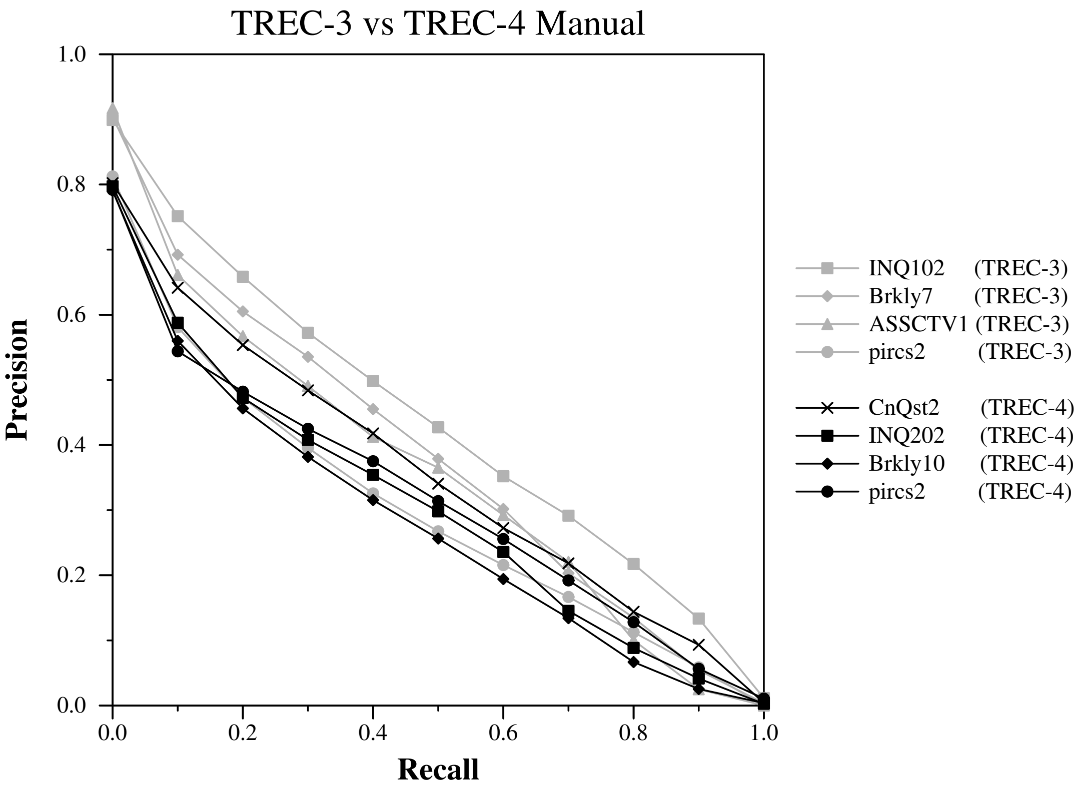

Figure 4. Comparison of Manual Adhoc Results for TREC-3 and TREC-4.

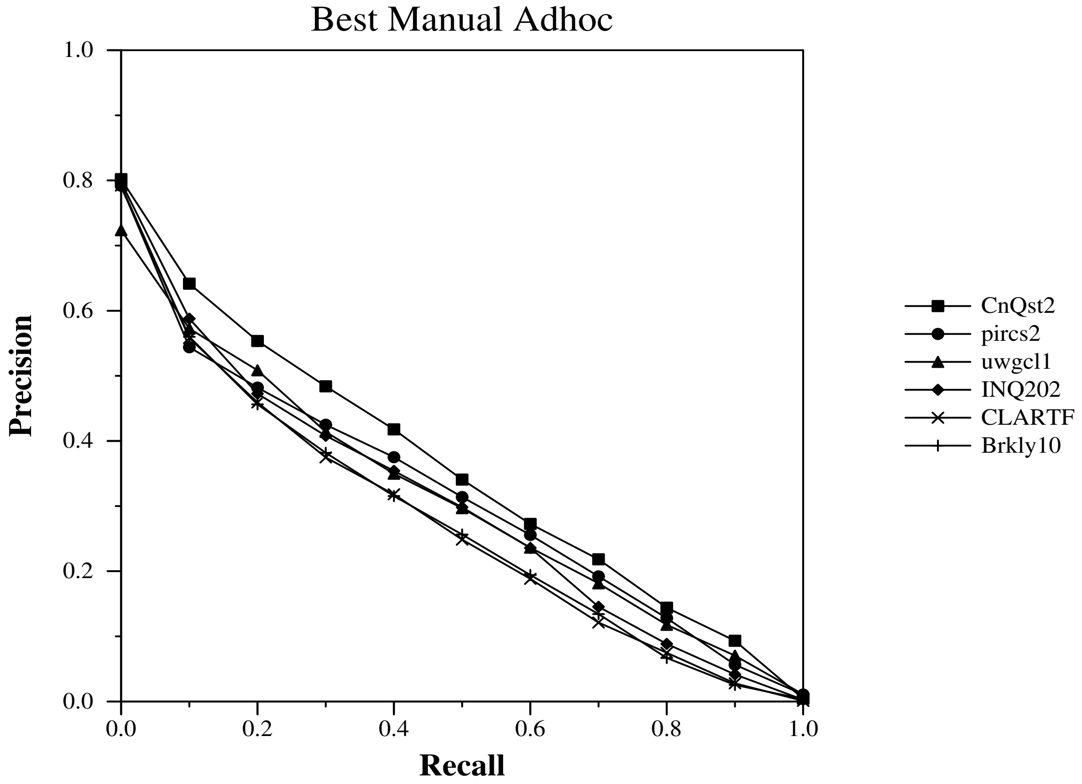

Figure 5. Best TREC-4 Manual Adhoc Results.

create intervals of text that satisfy specific conditions. The ranking algorithms rely on combining the results of increasingly less restrictive queries until the 1000 document list is created.

INQ202 -- University of Massachusetts at Amherst ("Recent Experiments with INQUERY" by James Allan, Lisa Bellesteros, James P. Callan, W. Bruce Croft and Zhihong Lu) This run is a manual modification of the *INQ201* run, with strict rules for the modifications that only allow removal of words and phrases, modification of weights, and addition of proximity restrictions. This type of manual modification increased overall average precision by 21%. The same types of modification gained only 15.5% in TREC-3.

CLARTF -- CLARITECH Corporation ("CLARIT TREC-4 Experiments" by David A. Evans, Natasa Milic-Frayling, and Robert G. Lefferts) used the CLARIT system in a machine-aided manual query construction process. The initial query terms were manually modified and weighted, and then terms were manually selected for addition to the query based on an automatic thesaurus extraction process. This particular run used a manually-built "required terms filter" to locate the best document windows for use in the thesaurus extraction process.

Brkly10 -- University of California, Berkeley ("Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections" by Fredric C. Gey, Aitao Chen, Jianzhang He and Jason Meggs) used manually-reformulated queries including expansion using the News database of the MELVYL electronic catalog to either add specific instances or synonyms and related terms. The basic retrieval system is a logistic regression model that combines information from 6 measures of document relevancy based on term matches and term distribution. The coefficients were learned from the training data.

These 6 runs (and most of the other manual runs) can be divided into three different styles of manual query construction. The first group uses an automatic query construction method as a starting point, and then manually modifies the results. The *INQ202* run is a good example of this, where words and phrases were removed, term weights were modified, and proximity restrictions were added to the initial automatic query. The *pircs2* results were based on reweighting of the automatically generated terms and then adding a few new terms. The *citym1* (not shown) results were based on pre-editing the automatically generated query, and then post-editing the automatic expansion of that query.

The results of these manual modifications were highly varied. The manual edits performed by City University were only marginally effective. Manual modification of term weights seemed to have more impact, as is illustrated by the 12% improvement in the *pircs2* run, and also by some unknown percentage of the INQUERY manual results. However the addition of a few expansion terms in the *pircs2* run, or the use of proximity restrictions (*INQ202*) look to be the most promising manual modifications. Note that several of the runs in this top 6 make heavy use of some type of proximity restrictions. The ConQuest group found major improvements from term grouping, and the Multitext system from the University of Waterloo relies on proximity restrictions for their results. Since proximity restrictions are related to the use of phrases (either statistical or syntactic) or the use of additional local information, this area is clearly a focus for further research.

The second type of manual query construction, exemplified by *uwgcl1* and *Brkly10*, used queries completely manually generated using some type of auxiliary information resource such as online dictionaries (*uwgcl1*) or news databases (*Brkly10*). The query generated for *uwgcl1* uses Boolean-type restrictors, whereas the query generated for *Brkly10* uses natural language.

The third type of manual query construction involves a more complex type of human-machine interaction. Both the *CnQst2* run and the *CLARTF* run are results of experiments examining a multi-stage process of query construction. The ConQuest group starts with a manual query, and then expands this query semi-automatically by manually choosing the correct senses of terms to expand. Then they manually modify the term weights and term grouping. The CLARITECH group manually modifies queries that are automatically generated, and then provides various levels of user control of an automatic expansion process (see the CLARITECH paper for several experiments involving this user control).

Note that these three styles of manual query construction require various levels of user effort and training. Simple edits of automatic queries, user term weighting, and (less likely) proximity restrictions can be done by a relatively untrained user. The performance of these users is not apt to be as good as the *INQ202* or *pircs2* results, however, since both of these runs were the results of the primary system developers functioning as users.

The complete manual generation of queries (such as the *uwgcl1* or *Brkly10* efforts) require the types of skills currently seen in search intermediaries. Using specific query languages takes lots of training, and learning to find reasonable terms to expand topics is an art acquired only after lots of practice. This should be contrasted with the third type of query construction. The complex interaction with the user exemplified by the *CnQst2* and *CLARTF* runs requires a different type (and possibly level) of skills and training. These systems are a completely new model of search engine, and it will be necessary to develop different skills and new "mental models" in order that users can become proficient in searching.

The amount of effort and training required to achieve these improvements over automatic results should not preclude using these techniques. Indeed the major improvements shown by these methods illustrate the importance of continuing investigation into the best places for human intervention. Many studies have shown that users feel a need for more control of their searching and this control is absent from current automatic systems.

5.5 TREC-4 Routing Results

The routing evaluation used a specifically selected subset of the training topics, with that selection guided by the availability of new testing data. The ease of obtaining more Federal Register documents suggested the use of topics that tended to find relevant documents in the Federal Register and 25 of the routing topics were picked using this criterion. The second set of 25 routing topics were selected to build a subcollection in the domain of computers. The testing documents for the computer issues were documents from the Internet, plus part of the Ziff collection (see table 3).

There were a total of 28 sets of results for routing evaluation, with 26 of them based on runs for the full data set. Of the 26 systems using the full data set, 23 used automatic construction of queries, and 3 used manual construction. There were 2 sets of category B routing results, both using automatic construction of queries.

Figure 6 shows the recall/precision curves for the 6 TREC-4 groups with the highest non-interpolated average precision for the routing queries. The runs are ranked by the average precision. A short summary of the techniques used in these runs follows. For more details on the various runs and procedures, please see the cited papers in this proceedings.

INQ203 -- University of Massachusetts at Amherst ("Recent Experiments with INQUERY" by James Allan, Lisa Bellesteros, James P. Callan, W. Bruce Croft and Zhihong Lu) used the inference net engine (same as for the adhoc task), but made major refinements of the algorithms used in TREC-3. The queries were constructed using a Rocchio weighting approach for terms in relevant and non-relevant training documents, and then these queries were expanded by 250 new concepts (adjacent term pairs) found in the 200-word best-matching windows in the relevant documents. Further experiments were made in weighting terms, including use of the

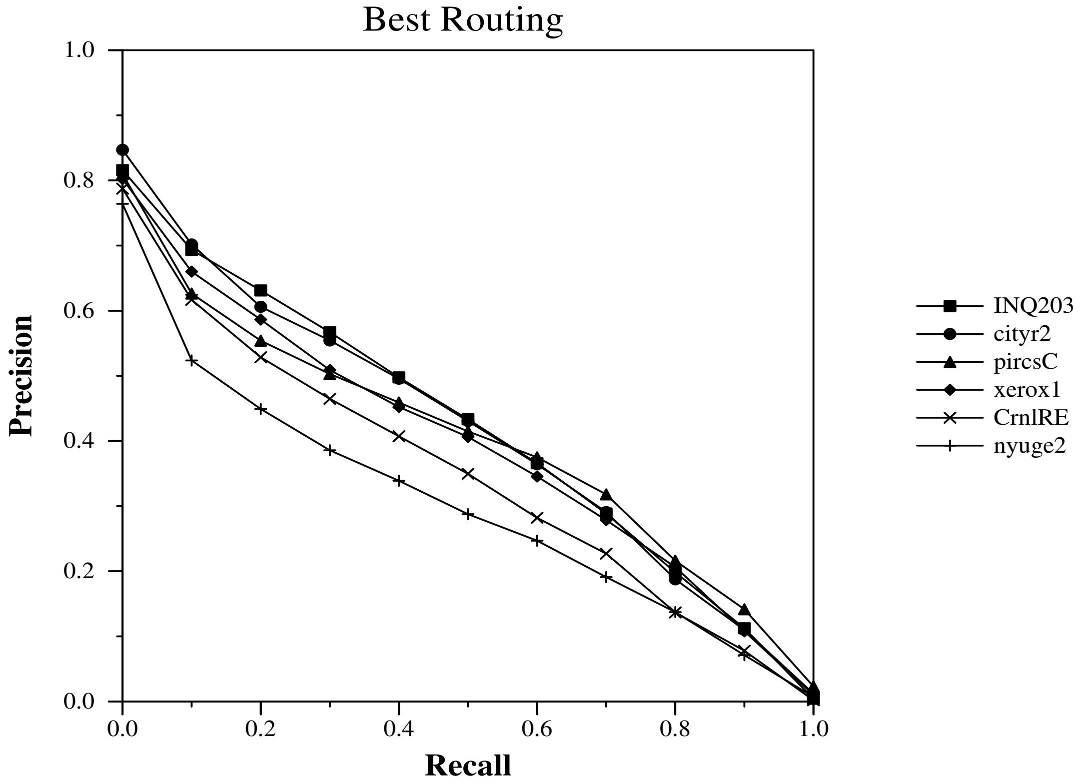

Figure 6. Best TREC-4 Routing Results.

Dynamic Feedback Optimization from Cornell (and City University).

cityr2 -- City University, London ("Okapi at TREC-4" by S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford and A. Payne") used the same probabilistic techniques as for the adhoc task, but constructed the query using a very selective set of terms (36 on average) from the relevant documents, similar to their TREC-3 approach. The method used for term selection involved optimizing the query based on trying different combinations of terms from the relevant documents. Since this is a very compute-intensive method, the work for TREC-4 looked for more efficient methods.

pircsC -- Queens College, CUNY ("TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS" by K.L. Kwok and L. Grunfeld) used the same spreading activation model used in the adhoc task, but combined the results of four different query experts. Two of these query experts used different levels of topic expansion (80 terms and 350 terms), and the other two were trained on specific subsets of the data (FR and Ziff vs WSJ, AP and SJMN).

xerox1 -- Xerox Research Center ("Xerox Site Report: Four TREC-4 Tracks" by Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schutze, Gregory Grefenstette and David Hull) used a complex routing algorithm that involved using LSI techniques to discover the best features, and then used three different classification techniques (combined) to rank the documents selected by these features.

CrnlRE -- Cornell University ("New Retrieval Approaches Using SMART: TREC-4" by Chris Buckley, Amit Singhal, Mandar Mitra, (Gerald Salton)) worked with the same new SMART algorithms used in the adhoc task. Because of inexperience with these new algorithms, minimal query expansion was used (only 50 single terms, as opposed to the TREC-4 300 terms). Dynamic query optimization was tried, but did not help.

nyuge2 -- GE Corporate Research and New York University ("Natural Language Information Retrieval: TREC-4 Report" by Tomek Strzalkowski and Jose Perez Carballo) used NLP techniques to discover syntactic phrases in the documents. Both single terms and phrases were indexed and specially weighted. The *nyuge2* run used topic expansion of up to 200 terms and phrases based on the relevant documents.

The issue of what features of documents should be used for retrieval was the paramount issue for all these groups (plus most of the other groups doing the routing task). It

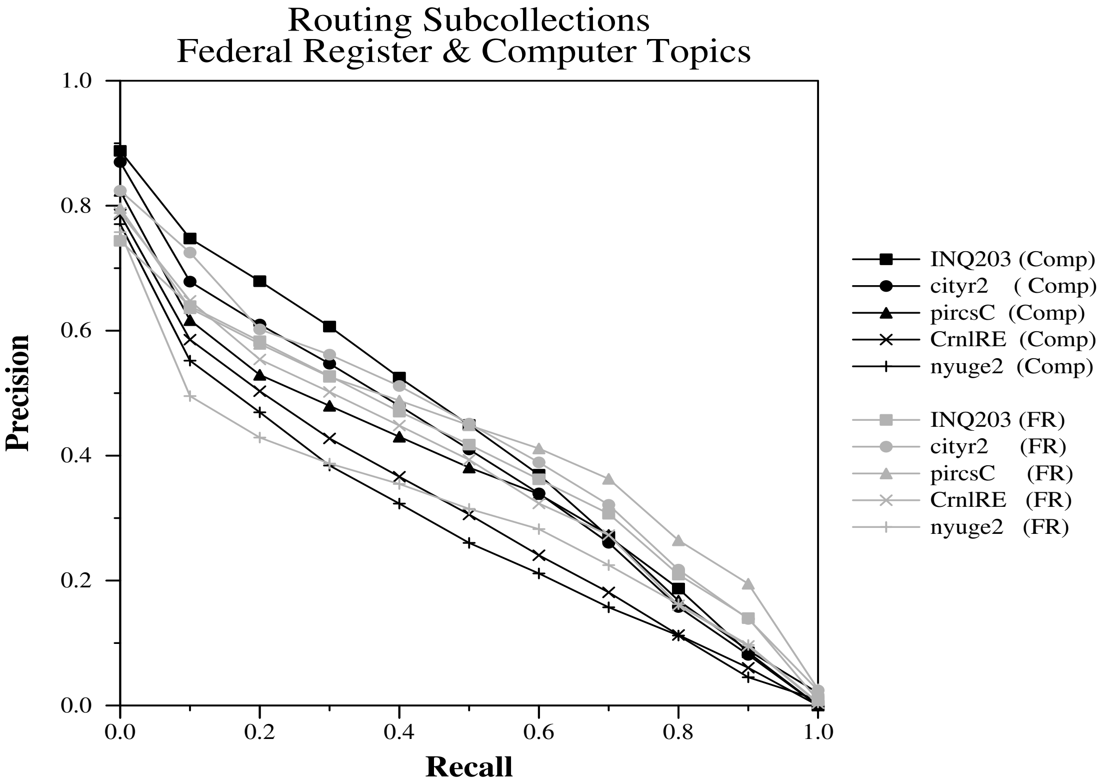

Figure 7. Comparison of Results for Federal Register Topics.

is interesting that the six groups shown in Figure 6 have used very different methods. The Cornell group used traditional Rocchio relevance feedback methods to locate and weight 50 terms and 10 statistical phrases. The statistical phrases are based on term co-occurance information for the whole collection, not just the relevant and expansion using 200 terms and syntactic phrases, with those phrases created from a full parse of the entire collection of documents. These methods can be contrasted with the INQUERY group, who started with a traditional Rocchio approach to select and weight 50 terms, but then expanded the query by 250 word pairs selected from only portions of the relevant documents.

The other three groups used less traditional methods. The group from City University repeated their very successful technique from TREC-3, in which they first used an ordering function to produce a list of terms as candidate terms for the query. This list was then optimized by repeatedly trying different sets of terms. The final term set in the *cityr2* run used an average of 36 terms per query, with the number varying across queries. The Xerox group started by expanding the query using Rocchio techniques, and used this expanded query to select 2000 documents. These 2000 documents were then fed into a LSI process to reduce the dimensionality of the final feature set. The final group, the *pircsC* run from Queens College, CUNY, was the result of four different expansions, two using different levels of expansion and two using different subcollections of documents for the expansion.

In addition to using different methods to select the features for the queries, two of the groups experimented with different ways of combining these features. The group from Xerox used three different classification techniques, combining the results from these three "experts." The *pircsC* group combined the results of their four query expansion experts. Both groups found that the combination of experts outperformed using a single method, even when one method (large expansion in the *pircs2* case and neural networks in the *xerox1* case) was generally superior. Also both groups found that there was a huge variation in performance across topics, with some topics performing best for each of the various experts.

The use of the two different subcollections of topics (25 in each set) for the routing task was, in general, not utilized by the various groups. However, it is very interesting to examine the results of the 6 groups shown in Figure 6 when broken into the two subsets. This is shown in Figure 7. The most prominant feature of these graphs is the difference in the shape of the curves. The *Federal Register* subcollection results (shown in grey) have a sharper drop in precision early in the curve, but better performance in general in the high recall end of the curve. Two

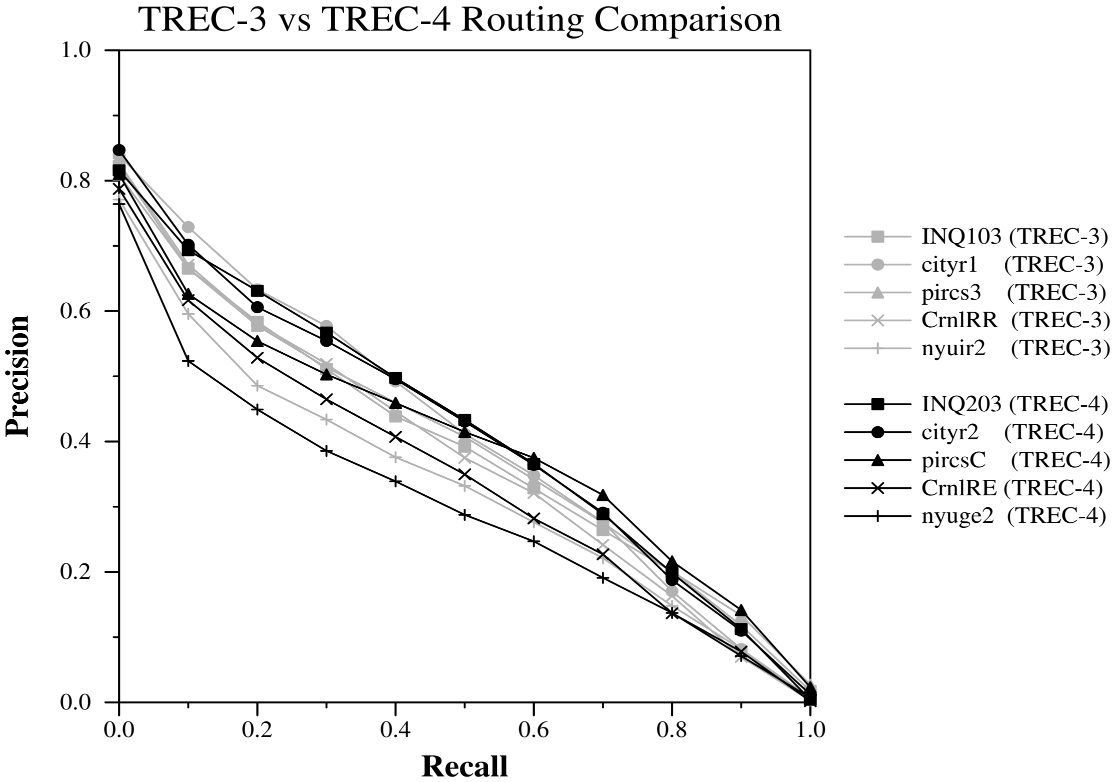

Figure 8. Comparison of Routing Results for TREC-3 and TREC-4.

differences in the subcollections account for this. First, the 25 topics in the FR subcollection retrieved significantly fewer relevant documents, an average of 99 relevant documents, as opposed to an average of 164 relevant documents for the computer topics. Additionally most of these relevant documents are *Federal Register* documents, which are very long and traditionally have been difficult to retrieve. These differences account for the sharp drop in precision in the low recall end of the curve. The higher performance of most of these 6 systems at the high recall end of the curve is somewhat more puzzling. It may be that the types of terminology in these subcollections are such that training is more effective in the FR subcollection.

Note that certain of the 6 systems seem more affected by the two subcollections. For example, the *pircsC* run is actually better for the FR subcollection than for the computer collection. This is likely because this system chunks all documents into 550 word segments, and therefore is less affected by the long FR documents. In contrast, the INQUERY system has excellent results for the computer topics, but a sharp drop in high precision results for the FR collection

There looks to be minimal improvement in overall routing results compared with those from TREC-3 (Figure 8). However, the TREC-4 topics were more difficult, particularly the FR topics. Despite the harder topics, many of the systems achieved performance improvements, especially at the high recall end of the curves. This indicates that the ability to find useful features that can retrieve the "hard-to-find" documents is growing. Such techniques as the use of word pairs from highly ranked sections of relevant documents by the INQUERY system, and the use of multiple experts in the *pircsC* and *xerox1* runs are showing promise.

6. TREC-4 TRACKS

Starting with TREC-1, there have always been groups that have pursued different goals than achieving high recall/precision performances on the adhoc and routing tasks. For example, the group from CITRI, Royal Melbourne Institute of Technology, has investigated efficiency issues in several of the TREC evaluations. By TREC-3 some of these areas had attracted several groups, all working towards the same goal. These became informal working groups, and in TREC-4 these working groups were formalized into "tracks," with specific guidelines.

6.1 The Multilingual Track

One of these tracks investigated the issues of retrieval in languages other than English. An preliminary Spanish test was run in TREC-3, with a formal track in TREC-4

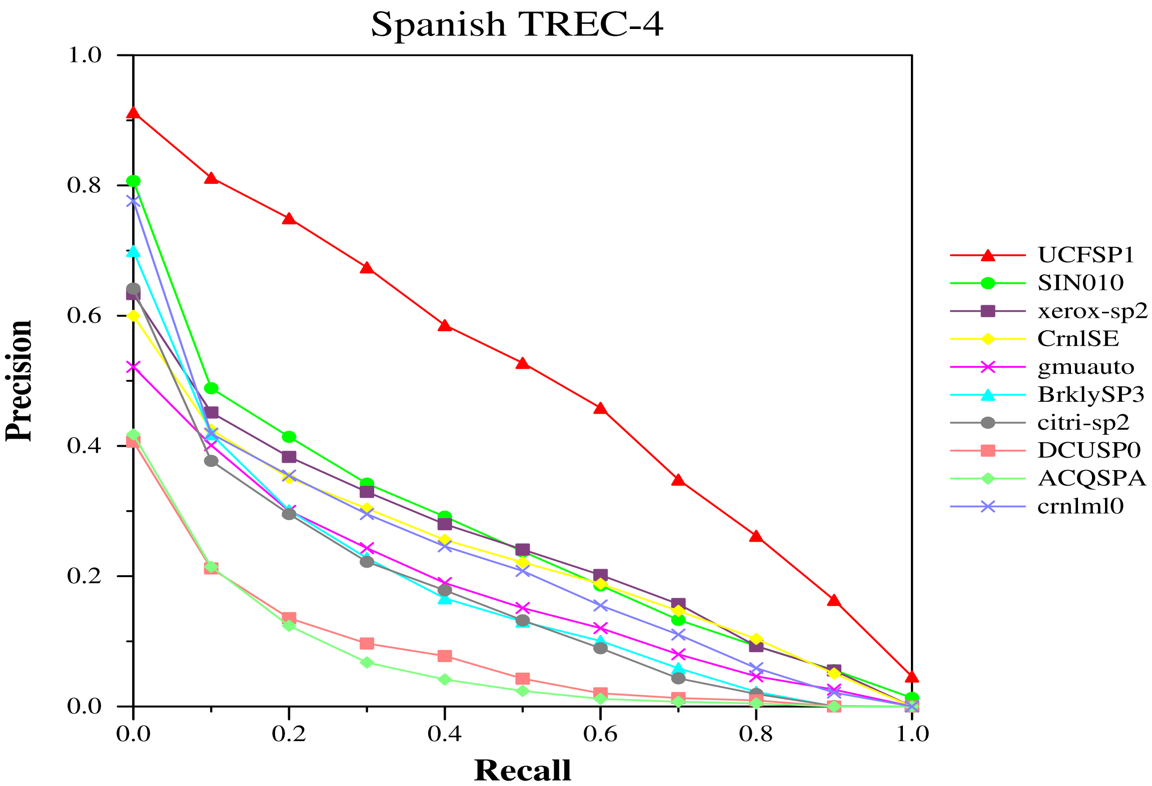

Figure 9. Results of TREC-4 Spanish Track.

that attracted 10 groups. Both TREC-3 and TREC-4 used the same documents, about 200 megabytes of the *El Norte* newspaper from Monterey, Mexico, but there were 25 different topics for each evaluation. Groups used the adhoc task guidelines, and submitted the top 1000 documents retrieved for each of the 25 Spanish topics.

The major result from TREC-3 was the ease of porting the retrieval techniques across languages. Cornell reported that only 5 hours to 6 hours of system changes were necessary (beyond creation of any stemmers or stopword lists). In TREC-4 there was training data (the results of TREC-3), and groups were able to do more elaborate testing. Figure 9 shows the recall/precision curves for these 10 TREC-4 groups, ordered by non-interpolated average precision. The cited papers are in this proceedings.

UCFSP1 -- University of Central Florida ("Multi-lingual Text Filtering Using Semantic Modeling" by James R. Driscoll, Sara Abbott, Kai-Lin Hu, Michael Miller and Gary Theis) used semantic modeling of the topics. A profile (entity-relationship schema) was manually built for each topic and lists of synonyms were constructed, including the use of an automatic Spanish verb form generator. The synonym list and domain list (instances of entities) were carefully built by Sara Abbott as part of a student summer project.

SINQ010 -- University of Massachusetts at Amherst ("Recent Experiments with INQUERY" by James Allan, Lisa Bellesteros, James P. Callan, W. Bruce Croft and Zhihong Lu) was a Spanish version of the automatic TREC-4 INQ201 run for the adhoc tests. The Spanish stemmer from TREC-3 was used, and terms were expanded using the basic InFinder technique (with a new noun phrase recognizer for Spanish).

xerox-sp2 -- Xerox Research Center ("Xerox Site Report: Four TREC-4 Tracks" by Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schutze, Gregory Grefenstette and David Hull) tested several Spanish language analysis tools, including a finite-state morphology and a hidden-Markov part-of-speech tagger to produce correct stemmed forms and to identify verbs and noun phrases. The SMART system was used as the basic search engine. Expansion was done using the top 20 retrieved documents.

CrnlSE -- Cornell University ("New Retrieval Approaches Using SMART: TREC-4" by Chris Buckley, Amit Singhal, Mandar Mitra, (Gerald Salton)) is a repeat of the TREC-3 work, using a simple stemmer and stopword list, and expanding by 50 terms from the top 20 documents. The TREC-3 version of SMART was used.

gmuauto -- George Mason University ("Improving Accuracy and Run-Time Performance for TREC-4" by David A. Grossman, David O. Holmes, Ophir Frieder, Matthew D. Nguyen and Christopher E. Kingsbury) used 5-grams with a vector-space type system for ranking. A Spanish stopword list was constructed using a Spanish linguist to prune a list of the most frequent 500 terms in the text.

BrklySP3 -- University of California, Berkeley ("Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections" by Fredric C. Gey, Aitao Chen, Jianzhang He and Jason Meggs) trained their logistic regression method on the Spanish results from TREC-3. They also built a rule-based Spanish stemmer, including a borrowed file of all verb forms for irregular verbs. The queries were formed manually by translating them into English, searching the MELVYL NEWS database, reformulating the English queries based on these searches, and then translating the queries back into Spanish.

citri-sp2 -- RMIT, Australia ("Similarity Measures for Short Queries" by Ross Wilkinson, Justin Zobel, and Ron Sacks-Davis) tried the combination methods used for their English results. A stop-list of 316 words was created, along with a Spanish stemmer that principally removed regular verb suffixes. Experiments were done using combinations of stopped and stemmed results.

DCUSP0 -- Dublin City University ("TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging in Spanish" by Alan F. Smeaton, Fergus Kelledy and Ruairi O'Donnell) used the NMSU part-of-speech tagger (at NMSU) as input to the SMART system. This method also produced the base forms of the terms. The traditional *tf*IDF* weighting was used, but adjectives were double-weighted.

ACQSPA -- Department of Defense ("Acquaintance: Language-Independent Document Categorization by N-Grams" by Stephen Huffman) used a 5-gram method which normalizes the resulting document vectors by subtracting a "collection" centroid vector. Minimal topic expansion was done.

crnlml0 -- New Mexico State University ("A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval" by Mark Davis and Ted Dunning) investigated five different methods of query translation. The Spanish topics were first manually translated into English for use in these tests. Then five different methods were used to automatically translate the topics into Spanish. The five methods were 1) a term-by-term translation using a bilingual dictionary, 2) use of the parallel corpus (UN corpus) for high-frequency terms, 3) use of a parallel corpus to locate statistically significant terms, 4) optimization of 2) and 5) an LSI technique on the parallel corpus.

In general the groups participating in the Spanish task were using the same techniques as for English. This is consistent with the philosophy that the basic search engine techniques are language-independent. Only the auxiliary techniques, such as stopword lists and stemmers, need to be language dependent. Several of the groups did major linguistic work on these auxiliary files, such as the noun-phrase identifier necessary for expansion using InFinder (the INQUERY system) and the two new Spanish stemmers (*BrklySP3* and *citri-sp2*). Two groups used n-gram methods, as did two of the groups in TREC-3.

Several other issues unique to this track should be mentioned. First, the outstanding results from the University of Central Florida indicate the benefits of very careful building of the manual queries, in this case by building extensive synonym sets and other such lists. The utility of this technique outside the rather limited domain of the TREC-4 topic set is an open question however. The group from Xerox did extensive work with Spanish language tools, but the effort had the same type of minimal effects generally seen in English. As a final point, the query translation experiments by New Mexico State University demonstrated a very interesting approach to the problem of multilingual retrieval, and hopefully will be followed by better results in TREC-5.

This track will be run again in TREC-5, with new Spanish data and 25 new Spanish topics. Also new for TREC-5 will be a Chinese retrieval task, with Chinese data and 25 Chinese topics.

6.2 The Confusion Track

The "confusion" track represents an extension of the current tasks to deal with corrupted data such as would come from OCR or speech input. The track followed the adhoc task, but using only the category B data. This data 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. Figure 10 shows the recall/precision curves for the confusion track, ordered by non-interpolated average precision. Two or three runs are shown for each group, the base run (no corruption), the 10% corruption level, and (sometimes) the 20% corruption level. The cited papers are in this proceedings.

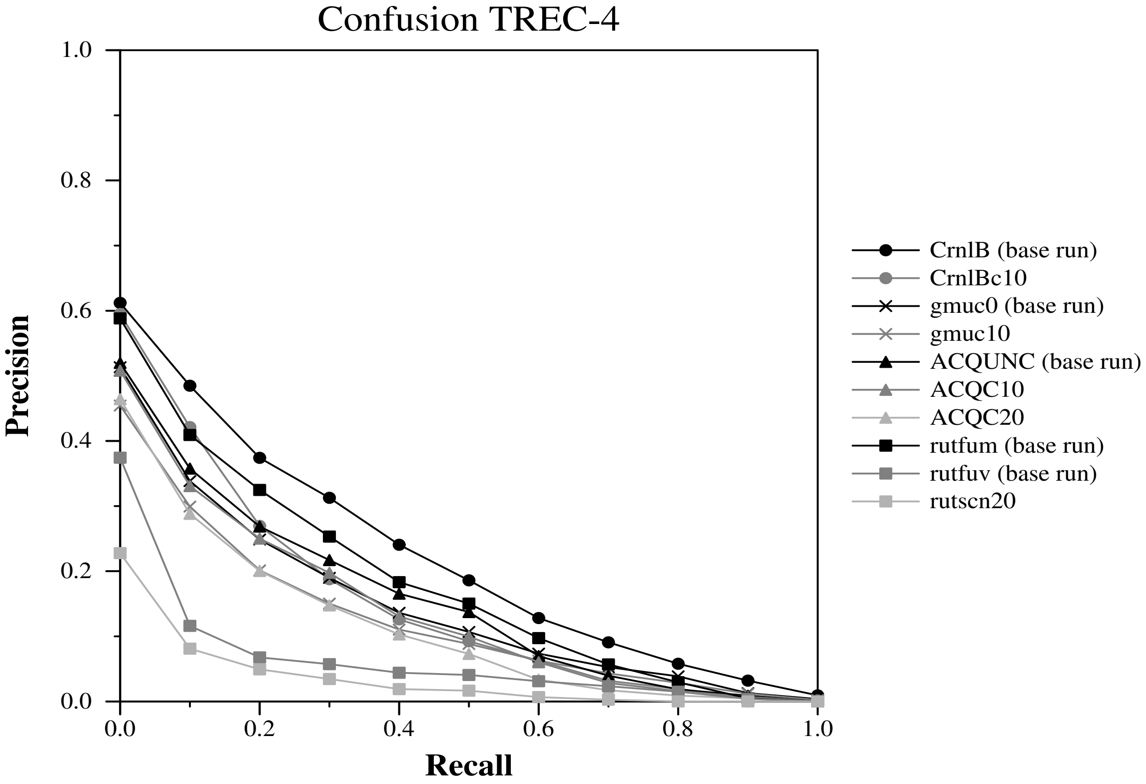

Figure 10. Results of TREC-4 Confusion Track.

CrnlB, CrnlBc10 -- Cornell University ("New Retrieval Approaches Using SMART: TREC-4" by Chris Buckley, Amit Singhal, Mandar Mitra, (Gerald Salton)) used a two-pass correction technique (only one-pass is implemented for this run). In the first pass, the query is expanded by all variants that are one transformation from the query word. The second pass improves the documents. This method avoids the use of a dictionary for correction of corrupted text.

ACQUNC, ACQC10, ACQC20 -- Department of Defense ("Acquaintance: Language-Independent Document Categorization by N-Grams" by Stephen Huffman) used an n-gram method which normalizes the resulting document vectors by subtracting a "collection" centroid vector. A 5-gram was used for the 10% corruption level and a 4-gram for the 20% level.

gmuc0, gmuc10 -- George Mason University ("Improving Accuracy and Run-Time Performance for TREC-4" by David A. Grossman, David O. Holmes, Ophir Frieder, Matthew D. Nguyen and Christopher E. Kingsbury) used a 4-gram method with a vector-space type system for ranking. A thresholding technique was tried that only worked with the best 75% of the 4-gram query in order to improve efficiency.

rutfum, rutfuv, rutscn20 -- Rutgers University ("Two Experiments on Retrieval with Corrupted Data and Clean Queries in the TREC-4 Adhoc Task Environment: Data Fusion and Pattern Scanning" by Kwong Bor Ng and Paul B. Kantor) tried the use of 5-grams and data fusion. The first experiment merged the results of two runs, one using 5-grams and one using words. The second experiment was a pattern scanning scheme called dotted 5-grams.

Since this was the first time this task had been tried, and since also there were very few participating groups, not much can be said about the results. Three of the four groups used N-grams, a method that is not known for the best results on uncorrupted data. The fourth group was unable to implement their full algorithms in time for the results. The track will be run again in TREC-5. Actual OCR output will be used at that time, as opposed to the randomly corrupted data used in TREC-4.

6.3 The Database Merging Track

A third area, that of properly handling heterogeneous collections such as the five main "subcollections" in TREC, was examined by the database merging track. 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.

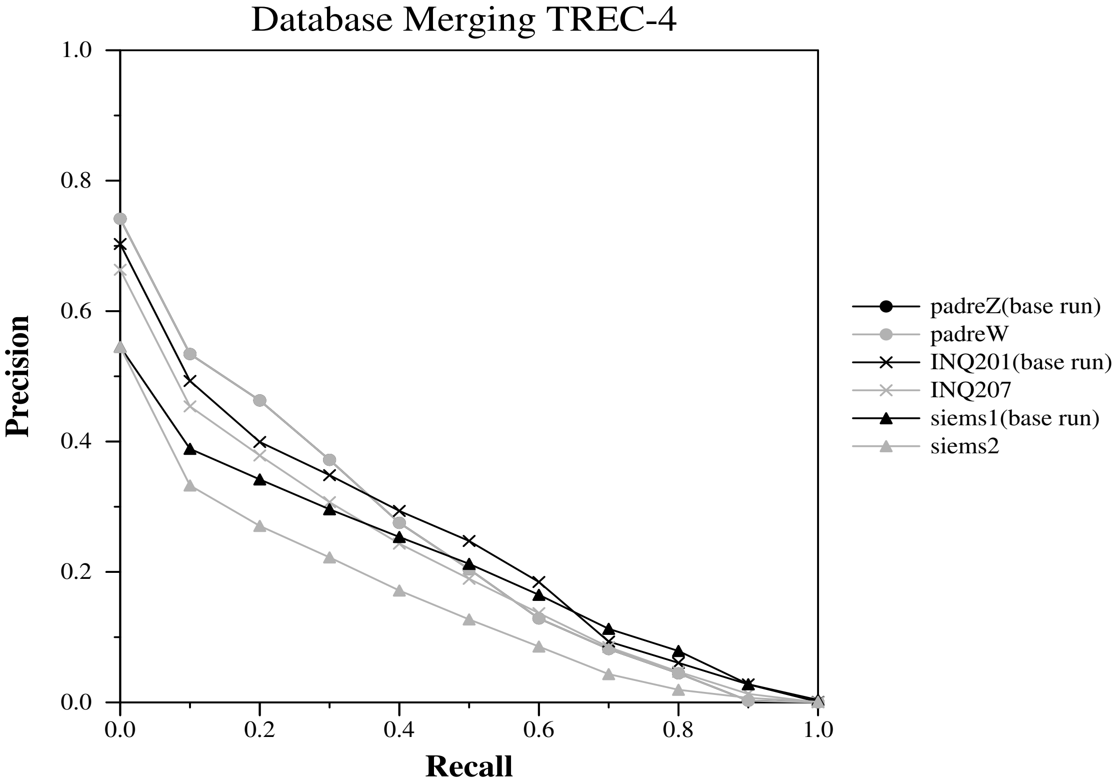

Figure 11. Results of TREC-4 Database Merging Track.

There were 10 subcollections 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). The 3 participating groups ran the adhoc topics separately on each of the 10 subcollections, merged the results, and submitted these results, along with a baseline run treating the subcollections as a single collection.

Figure 11 shows the recall/precision curves for this track, ordered by non-interpolated average precision. Two runs are shown for each group, the base run (indexed as a single database), and the best of their merged runs. The cited papers are in this proceedings.

padreZ, padreW -- Australian National University ("Proximity Operators -- So Near and Yet So Far" by David Hawking and Paul Thistlewaite) used manual queries with proximity operators. Since there are no collection-dependent variables in this system, the run using the 10 separate collections is equivalent to the run using the entire collection.

INQ201, INQ207 -- University of Massachusetts at Amherst ("Recent Experiments with INQUERY" by James Allan, Lisa Bellesteros, James P. Callan, W. Bruce Croft and Zhihong Lu) tried five variations of a basic method of collection merging [Callan et al. 1996]. The basic method scored each collection against the topic, and then weighted the document results by their collection score.

siems1, siems2 -- Siemens Corporate Research ("Siemens TREC-4 Report: Further Experiments with Database Merging" by Ellen M. Voorhees) tried two different methods, both based on information about the previous queries (training topics) as opposed to using information about the document collection itself.

If results are produced without use of collection information, then the merging process is trivial, as illustrated by the *padre* runs. 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 is the hypothesis that one might want to bias searching towards certain collections, either by developing collection scores (such as the INQUERY work) or by developing a sense of history from previous queries (the Siemens work).

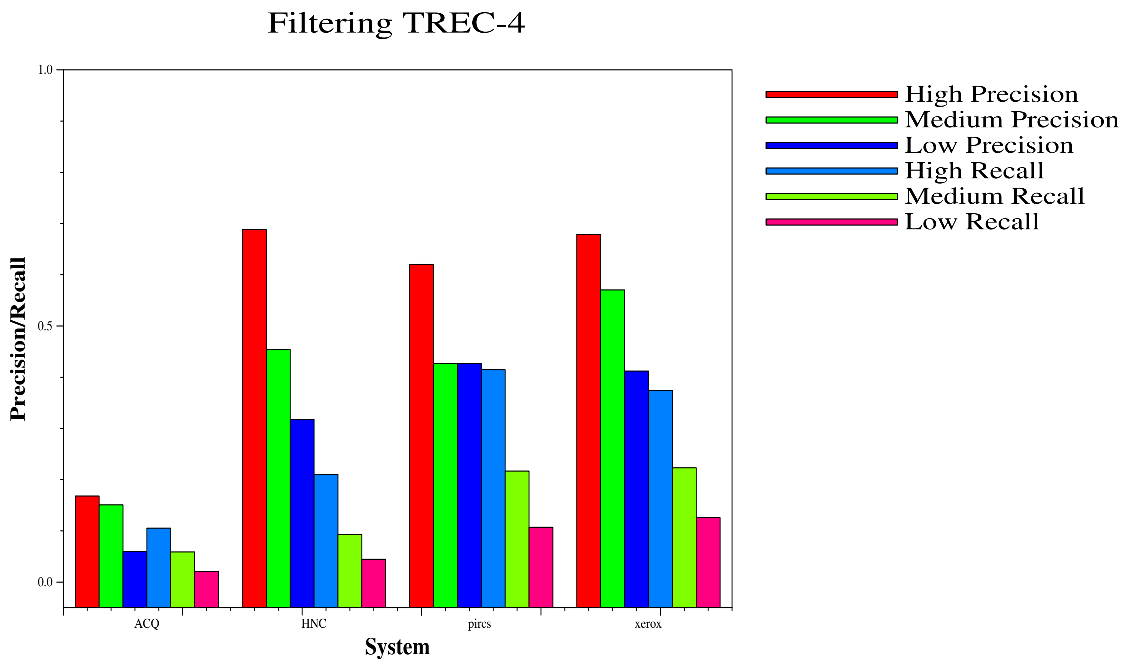

Figure 12. Results of TREC-4 Filtering Track.

6.4 The Filtering Track

For several years some participants have been concerned about the definition of the routing task, and the filtering track represents a new variation of this task. In TREC-4 this track documents, and test documents as the routing task. The difference was that the results submitted for the filtering runs were unranked sets of documents satisfying three "utility function" criteria. These criteria were designed to approximate a high precision run, a high recall run, and a "balanced" run. For more details, see the paper "The TREC-4 Filtering Track" by David Lewis (in this proceedings).

Figure 12 shows the results of the four groups that tried this track. There are 3 pairs of bars for each system, one pair corresponding to each of the three utility function criteria. The first of the pairs (the left-most and the right-most bars) correspond to the high precision/low recall run. The second pair (the second and fifth bars) correspond to the balanced (medium precision/medium recall) run, and the third pair (high recall/low precision run) are shown in the middle two bars.

One desired type of system behavior is the "stairstep" effect seen, for example, in the run from HNC Software Inc. (see paper "Using CONVECTIS, A Context Vector-Based Indexing System for TREC-4" by Joel L. Carleton, William R. Caid and Robert V. Sasseen in this proceedings). When this system is compared with the next two systems (*pircs* and *xerox*) , it can be seen that while the HNC system got a better separation of the runs, the other two groups got better results in general, particularly for the balanced run.

This was the first time this track had been tried, and the development of evaluation techniques was the most critical area. Now that these techniques are in place, it is expected that more groups will take part in the track in TREC-5.

6.5 The Interactive Track

An interactive track was formed for TREC-4, with the double goal of developing better methodologies for interactive evaluation and investigating in depth how users search the TREC topics. Eleven groups took part in this track in TREC-4, using a subset of the adhoc topics. Many different types of experiments were run, but the common thread was that all groups used the same topics, performed the same task(s), and recorded the same information about how the searches were done. Task 1 was to retrieve as many relevant documents as possible within a certain timeframe. Task 2 was to construct the best query possible. The cited papers are in this proceedings.

rutint1, *rutint2* -- Rutgers University ("Using Relevance

Feedback and Ranking in Interactive Searching" by Nicholas J. Belkin, Colleen Cool, Jurgen Koenemann, Kwong Bor Ng and Soyeon Park) recruited 50 searchers for this task. The INQUERY search engine was used, and the particular emphasis was on studying the use of ranking and relevance feedback by these searchers.

cityi1 -- City University, London ("Okapi at TREC-4" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used members of their team to evaluate their new GUI interface to OKAPI. They concentrated on examining the various stages of searching, and kept notes on items of interest, such as how many titles were examined, how many iterations were run, and how the queries were edited at various times in the search process.

UofTo1 -- University of Toronto ("Is Recall Relevant? An Analysis of How User Interface Conditions affect Strategies and Performance in Large Scale Text Retrieval" by Nipon Charoenkitkarn, Mark H. Chignell and Gene Golovchinsky) used 36 searchers on a new version of their system called BrowsIR. The goal of their experiments was to compare three different strategies for constructing queries: a text markup (similar to that done by this group in TREC-3), a query typing method, and a hybrid method. Both experts and novices were used.

ETHI01 -- Swiss Federal Institute of Technology (ETH) ("Highlighting Relevant Passages for Users of Interactive SPIDER Retrieval System" by Daniel Knaus, Elke Mittendorf and Peter Schäuble and Paraic Sheridan) experimented with several algorithms to highlight the most relevant passages, and tested this on 11 users as an aid to relevance feedback.

XERINT1, XEROXINT2 -- Xerox Research Center ("Xerox Site Report: Four TREC-4 Tracks" by Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schutze, Gregory Grefenstette and David Hull) tried three different modes of searching interfaces. The first was the Scatter/Gather method of visualizing the document space, the second was the TileBars to visualize the documents, and the third was the more traditional ranked list of titles from a vector space search engine.

CLARTI -- CLARITECH Corporation ("CLARIT TREC-4 Interactive Experiments" by Natasa Milic-Frayling, Cheng-Xiang Zhai, Xiang Tong, Michael P. Mastroianni, David A. Evans and Robert G. Lefferts) used the CLARIT system interactively to study the effects of the quality of a user's relevance judgments, the effects of time constraints on searching, and the effects of relevance feedback on the final results of queries.

LNBOOL -- Lexis-Nexis ("Interactive Boolean Search in TREC4" by David James Miller, John D. Hold and X. Allan Lu) used expert Boolean searchers and the commercial Lexis-Nexis software to compare retrieval performance between Boolean and non-Boolean systems.

gatin1, gatin2 -- Georgia Institute of Technology ("Interactive TREC-4 at Georgia Tech" by Aravindan Veerasamy) investigated the effectiveness of a new visualization tool that shows the distribution of query terms across the document space.

ACQINT -- Department of Defense ("Acquaintance: Language-Independent Document Categorization by N-Grams" by Stephen Huffman) used the Parentage information visualization system which shows clusters of documents, along with the terms which characterize those clusters.

CrnlI1, CrnlI2 -- Cornell University ("New Retrieval Approaches Using SMART: TREC-4" by Chris Buckley, Amit Singhal, Mandar Mitra, (Gerald Salton)) did an experiment to test how much of the document needed to be read in order to determine document relevancy for input to relevance feedback. They tested quick scans vs full reading.

Whereas all participants found the track very interesting and useful, there were difficulties in comparing results. One of the major outcomes of this track in TREC-4, therefore, was a general awareness of the large number of variables that need to be controlled in order to compare results. Some of these, such as the variation in performance across topics, affect all the TREC tasks, but the human element in the interactive track compounds the problem immensely. The emphasis for TREC-5 work will be on learning to control or monitor some of these variables as a first step to providing better evaluation methodology.

7. Summary

The main conclusions that can be drawn from TREC-4 are as follows:

- The much shorter topics in the adhoc task caused all systems trouble. The expansion methods used in TREC-3 continued to work, but obviously needed modifications. The types of passage retrieval used in TREC-3 did not work. The fact that the performance of the manually built queries was also hurt by the short topics implies that there are some issues involving the use of very short topics in TREC that need further investigation. It may be that the statistical "clues" presented by these shorter topics are simply not enough to provide good retrieval performance in the batch testing environment of TREC. The topics to be used in TREC-5 will contain both a short and a long version to aid in these further investigations.

- Despite the problems with the short topics, many of the systems made major modifications to their term weighting algorithms. In particular, the SMART group from Cornell University and the INQUERY group from the University of Massachusetts at Amherst produced new algorithms that yielded much better results (on the longer TREC-3 queries), and their TREC-4 results were not lowered as much as they would have been.
- There were five tracks run in TREC-4.
  - Interactive -- 11 groups investigated searching as an interactive task by examining the process as well as the outcome. The major result of this track, in addition to interesting experiments, was an awareness of the difficulties of comparing results in an interactive testing environment.
  - Multilingual -- 10 groups working with 250 megabytes of Spanish and 25 topics verified the ease of porting to a new language (at least in a language with no problems in locating word boundaries). Additionally some improved Spanish stemmers were built.
  - Multiple database merging -- 3 groups investigated techniques for merging results from the various TREC subcollections.
  - Data corruption -- 4 groups examined the effects of corrupted data (such as would come from an OCR environment) by using corrupted versions of the category B TREC data.
  - Filtering -- 4 groups evaluated routing systems on the basis of retrieving an unranked set of documents optimizing a specific effectiveness measure.

  The results from these last 3 tracks were inconclusive, and should be viewed as a first-pass at these focussed tasks.

There will be a fifth TREC conference in 1996, and most of the systems that participated in TREC-4 will be back, along with additional groups. The routing and adhoc tasks will be done again, with different data, and new topics similar in length to the TREC-3 topics. In addition, all five tracks will be run again, with new data. The Multilingual track will be run with Spanish and, as a first time, with Chinese data and topics.

Acknowledgments

The author would like to gratefully acknowledge the continued support of the Intelligent Systems Office of the Defense Advanced Research Projects Agency for the TREC conferences. Special thanks also go to the TREC program committee and the staff at NIST.

7. REFERENCES

Callan J.P., Lu Z., and Croft W.B. (1996). Searching Distributed Collections with Inference Networks. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29.

Harman D. (Ed.). (1994). *Overview of the Third Text REtrieval Conference (TREC-3)*. National Institute of Standards and Technology Special Publication 500-225, Gaithersburg, Md. 20899.

Sparck Jones K. and 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.