The Text REtrieval Conferences (TRECs)

Donna Harman
National Institute of Standards and Technology
Gaithersburg, MD 20899

There have been four Text REtrieval Conferences (TRECs); TREC-1 in November 1992, TREC-2 in August 1993, TREC-3 in November 1994 and TREC-4 in November 1995. The number of participating systems has grown from 25 in TREC-1 to 36 in TREC-4, including most of the major text retrieval software companies and most of the universities doing research in text retrieval (see table for some of the participants). The diversity of the participating groups has ensured that TREC represents many different approaches to text retrieval, while the emphasis on individual experiments evaluated in a common setting has proven to be a major strength of TREC.

The test design and test collection used for document detection in TIPSTER was also used in TREC. The participants ran the various tasks, sent results into NIST for evaluation, presented the results at the TREC conferences, and submitted papers for a proceedings. The test collection consists of over 1 million documents from diverse full-text sources, 250 topics, and the set of relevant documents or "right answers" to those topics. A Spanish collection has been built and used during TREC-3 and TREC-4, with a total of 50 topics.

TREC-1 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 results from TREC-2 showed significant improvements over the TREC-1 results, and should be viewed as the appropriate baseline representing state-of-the-art retrieval techniques as scaled up to handling a 2 gigabyte collection.

TREC-3 therefore provided the first opportunity for more complex experimentation. The major experiments in TREC-3 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 the training information to select only the best terms for routing 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.

TREC-4 allowed a continuation of many of these complex experiments. The topics were made much shorter and this change triggered extensive investigations in automatic query expansion. There were also five new tasks, called tracks. These were added to help focus research on certain known problem areas, and included such issues as investigating searching as an interactive task by examining the process as well as the outcome, investigating techniques for merging results from the various TREC subcollections, examining the effects of corrupted data, and evaluating routing systems using a specific effectiveness measure. Additionally more groups participated in a track for Spanish retrieval.

The TREC conferences have proven to be very successful, allowing broad participation in the overall DARPA TIPSTER effort, and causing widespread use of a very large test collection. All conferences have had very open, honest discussions of technical issues, and there have been large amounts of "cross-fertilization" of ideas. This will be a continuing effort, with a TREC-5 conference scheduled in November of 1996.

A Sample of the TREC-4 Participants

CLARITECH/Carnegie Melon University
CITRI, Australia
City University, London
Cornell University
Department of Defense
Excalibur Technologies, Inc.
GE Corporate R & D/New York University
George Mason University
HNC, Inc.
Lexis-Nexis
Logicon Operating Systems
NEC Corporation
New Mexico State University
Queens College, CUNY
Rutgers University (two groups)
Siemens Corporate Research Inc.
Swiss Federal Institute of Technology (ETH)
University of California - Berkeley
University of Massachusetts at Amherst
University of Waterloo
Xerox Palo Alto Research Center

The Text REtrieval Conferences (TRECs)

Donna Harman
National Institute of Standards and Technology
Gaithersburg, MD 20899

1. INTRODUCTION

Phase two of the TIPSTER project included two workshops for evaluating document detection (information retrieval) projects: the third and fourth Text REtrieval Conferences (TRECs). These workshops were held at the National Institute of Standards and Technology (NIST) in November of 1994 and 1995 respectively. The conferences included evaluation not only of the TIPSTER contractors, but also of many information retrieval groups outside of the TIPSTER project. The conferences were run as workshops that provided a forum for participating groups to discuss their system results on the retrieval tasks done using the TIPSTER/TREC collection. As with the first two TRECs, the goals of these workshops were:

- 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
- To serve as a showcase for state-of-the-art retrieval systems for DARPA and its clients.

The number of participating systems has grown from 25 in TREC-1 to 32 in TREC-3 (see Table 1) and to 36 in TREC-4 (see Table 2). These systems include most of the major text retrieval software companies and most of the universities doing research in text retrieval. Note that whereas the universities tend to participate every year, the companies often skip years because of the amount of effort required to run the TREC tests.

By opening the evaluation to all interested groups, TIPSTER has ensured that TREC represents many different approaches to text retrieval. The emphasis on diverse 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 handling a 2 gigabyte collection.

TREC-3 provided an opportunity for 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 help systems 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). For more details on the specific system approaches, see the complete overview of the TREC-3 conference, including papers from the participating groups [1].

TREC-4 presented a continuation of many of these complex experiments, and also included a set of five focussed tasks, called tracks. Both the main tasks were more difficult -- the test topics were much shorter, and the test documents were harder to retrieve. Several groups made major changes in their retrieval algorithms, and all groups had difficulty working with the very short topics. Many interesting experiments were done in the tracks, including 10 groups that worked with Spanish

Australian National University	Bellcore
CLARITECH/Carnegie Mellon University	CITRI, Australia
City University, London	Cornell University
Dublin City University	Environment Research Institute of Michigan
Fulcrum	George Mason University
Logicon Operating Systems	Mayo Clinic/Foundation
Mead Data Central	National Security Agency
New York University	NEC Corporation
Queens College	Rutgers University (two groups)
Siemens Corporate Research Inc.	Swiss Federal Institute of Technology (ETH)
TRW/Paracel	Universitaet Dortmund, Germany
University of California - Berkeley	University of Central Florida
University of Massachusetts at Amherst	VPI&SU (Virginia Tech)
University of Minnesota	University of Toronto
Universite de Neuchatel, Switzerland	Verity Inc.
West Publishing Co.	Xerox Palo Alto Research Center

Table 1: TREC-3 Participants

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 2: TREC-4 Participants

data, and 11 groups that ran extensive experiments in interactive retrieval. Details of specific system approaches are in the proceedings of the TREC-4 conference [2].

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 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 section 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 Figure 1).

The 50 routing topics for testing are a specific subset of the training topics (selected by NIST). In TREC-3 the routing topics corresponded to the TREC-2 adhoc topics, i.e., topics 100-150. The test documents for TREC-3 were the documents on disk 3 (see section 3.2). Although this disk had been part of the general training data, there were no relevance judgments for topics 100-150 made on this disk of documents. This less-than-optimal testing was required by the last-minute unavailability of new data.

In TREC-4 a slightly different methodology was used to select the routing topics and test data. Because of the 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 in 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.

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 training documents. Fifty new topics (numbers 150-200) were generated for TREC-3, with fifty additional new topics created for TREC-4 (numbers 201-250). The known documents used in TREC-3 were on disks 1 and 2, and those 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.

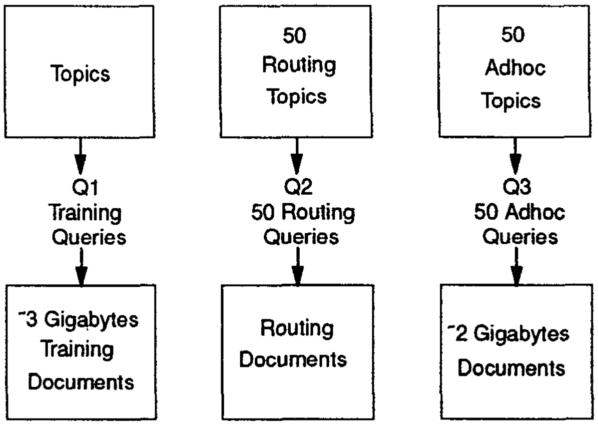

Figure 1. TREC Main Tasks

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, they are constructed to reflect an actual operational environment, and to allow as fair as possible separation among the diverse query construction approaches. Three generic categories of query construction were defined, based on the amount and kind of manual intervention used.

1. Automatic (completely automatic query construction)
2. Manual (manual query construction)
3. Interactive (use of interactive techniques to construct the queries)

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. In TREC-3, four out of the five tracks were run as preliminary investigations into the feasibility of running formal tracks in TREC-4.

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 use of the interactive query construction method in TREC-3 demonstrated interest in using interactive search techniques, so a formal track was formed for TREC-4. 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). Several groups tried these techniques in TREC-3 and it was decided to form a track in this area for TREC-4. 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. This was a new track proposed during the TREC-3 conference. 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 current routing track. For several years some participants have been concerned about the definition of the routing task, and a few groups experimented in TREC-3 with an alternative definition of routing. In TREC-4 the track was formalized. 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 the TREC-4 proceedings).

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-3 and TREC-4, disks 1, 2 and 3 were all available as training material (see Table 3). In TREC-3, disks 1 and 2 were also used for the adhoc task, and disk 3 for the routing task. In TREC-4, disks 2 and 3 were used for the adhoc task, and new data (also shown in Table 3) was used for the routing task. The following shows the actual contents of each of the three CD-ROMs (disks 1, 2, and 3).

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 3: Document Statistics

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 3 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.

Sample TREC-1/TREC-2 topic

```
<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing

<desc> Description:
Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.

<narr> Narrative:
A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one or more features of the company's product.

<con> Concept(s):
1. natural language processing
2. translation, language, dictionary, font
3. software applications

<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>
```

Each topic is formatted in the same standard method to allow easier automatic construction of queries. Besides a beginning and an end marker, each topic has a number, a short title, and a one-sentence description. There is a narrative section which is aimed at providing a complete description of document relevance for the assessors. Each topic also has a concepts section with a list of concepts related to the topic. This section is designed to provide a mini-knowledgebase about a topic such as a real searcher might possess. Additionally each topic can have a definitions section and/or a factors section. The definition section has one or two of the definitions critical to a human understanding of the topic. The factors section is included to allow easier automatic query building by listing specific items from the narrative that constrain the documents that are relevant. Two particular factors were used in the TREC-1/TREC-2 topics: a time factor (current, before a given date, etc.) and a nationality factor (either involving only certain countries or excluding certain countries).

The new (adhoc) topics used in TREC-3 reflect a slight change in direction. Whereas the TREC-1/TREC-2 topics were designed to mimic a real user's need, and were written by people who are actual users of a retrieval system, they were intended to represent long-standing information needs for which a user might be willing to create elaborate topics. This made them more suited to the routing task than to the adhoc task, where users are likely to ask much shorter questions. The adhoc topics used in TREC-3 (topics 151-200) are not only much shorter, but also are missing the complex structure of the earlier topics. In particular the concepts field has been removed because it was felt that real adhoc questions would not contain this field, and because inclusion of the field discouraged research into techniques for expansion of "too short" user need expressions. The shorter topics do not create a problem for the routing task, as experience in TREC-1 and 2 has shown that the use of the training documents allows a shorter topic (or no topic at all).

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.

In addition to being shorter, the new topics were written by the same group of people that did the relevance judgments (see next section). Specifically, each of the new topics (numbers 151-200) 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).

However, 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 4 gives the average number of terms in the title, description, narrative, and concept fields (all three fields for TREC-1 and TREC-2, no concept field in TREC-3, and only a description field in TREC-4). As can be seen, the topics are indeed much shorter, particularly in going from TREC-3 to TREC-4.

	Mean	Median
TREC-1	131	127
TREC-2	157	161
TREC-3	107	105
TREC-4	16	17

Table 4: Topic Lengths

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 [3] 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 5 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 has increased by more than 20% (adhoc), the number of unique documents found has actually dropped. The percentage of relevant documents found, however, has not changed much. The more accurate results going from TREC-1 to TREC-2 mean that fewer nonrelevant documents are being found by the systems. This trend continued in TREC-3, with a major drop (particularly for the routing task) that reflects increased accuracy in rejecting nonrelevant documents. In TREC-4, the trend was reversed. In the case of the adhoc task (including most of the track runs also), there is a slight increase in the percentage of unique documents found, probably caused by the wider variety of expansion terms used by the systems to compensate for the lack of a narrative section in the topic. A larger percentage increase is seen in the routing task, due to fewer runs being pooled, i.e., a higher percentage of documents is likely to be unique. Also the TREC-4 routing task was more difficult, both because of the long Federal Register documents and because there was 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).

	Adhoc		
	Possible	Actual	Relevant
TREC-1	3300	1279 (39%)	277 (22%)
TREC-2	4000	1106 (28%)	210 (19%)
TREC-3	4800	1005 (21%)	146 (15%)
TREC-4	7300	1710 (24%)	130 (7.5%)

	Routing		
	Possible	Actual	Relevant
TREC-1	2200	1067 (49%)	371 (35%)
TREC-2	4000	1466 (37%)	210 (14%)
TREC-3	4900	703 (14%)	146 (21%)
TREC-4	3800	957 (25%)	132 (14%)

Table 5: Overlap of Submitted Results

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 3 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 well less than one new relevant document per run, since 48 runs from all systems were used in the adhoc test (49 runs in the routing test). These tests show that the levels of completeness found during the TREC-2 and TREC-3 testing 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. Table 6 shows the breakdown for the 50 adhoc topics in TREC-3. The median of 30 new relevant documents occurs for a topic with 122 original 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.

TREC-3 -- Relevant Documents Found above 100			
Percent New Rel.	No. of Topics	Average New Rel.	Average No. Rel.
0%	1	0	85
1-9%	12	3	65
10-19%	7	13	96
20-29%	22	59	237
30-36%	8	137	381
Average		50	196
Median		30	122

Table 6: Relationship between completeness and the initial number of 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.

4. EVALUATION

An important component of TREC was to provide a common evaluation forum. Standard recall/precision figures have been calculated for each TREC system, along with some single-value evaluation measures. New for TREC-3 was a histogram for each system showing performance on each topic. In general, more emphasis has been placed on a "per topic analysis' in an effort to get beyond the problems of averaging across topics.

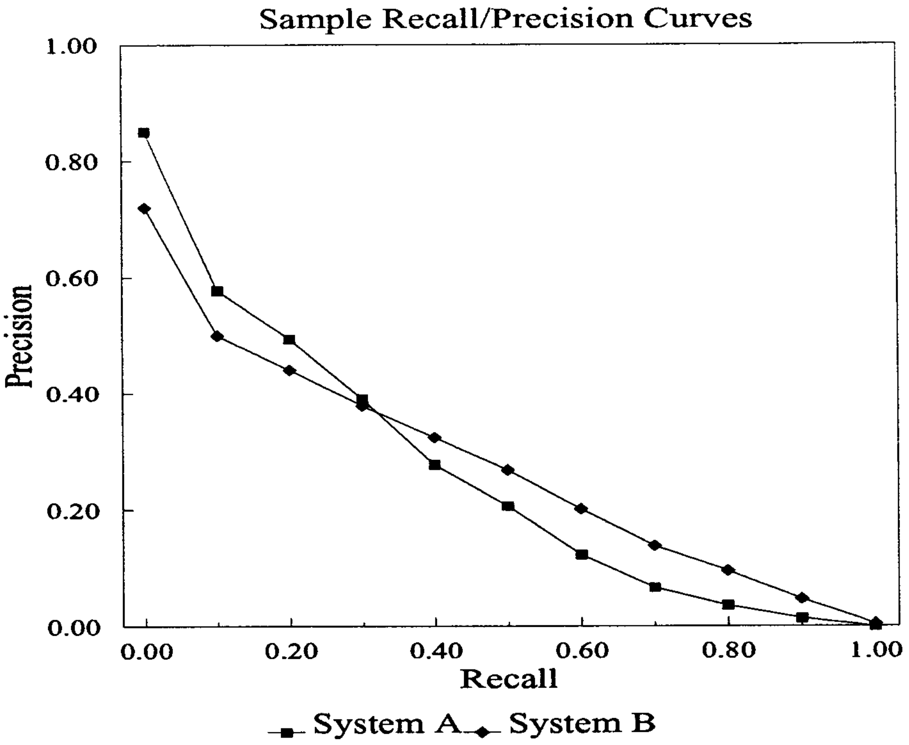

Figure 2. A Sample Recall/Precision Curve.

Work has been done, however, to find statistical differences among the systems (see paper "A Statistical Analysis of the TREC-3 Data" by Jean Tague-Sutcliffe and James Blustein in the TREC-3 proceedings.) Additionally charts have been published in the proceedings that consolidate information provided by the systems describing features and system timing, and allowing some primitive comparison of the amount of effort needed to produce the results.

4.1 Definition of Recall/Precision

Figure 2 shows typical recall/precision curves. The x axis plots the recall values at fixed levels of recall, where

$$Recall = \frac{number\ of\ relevant\ items\ retrieved}{total\ number\ of\ relevant\ items\ in\ collection}$$

The y axis plots the average precision values at those given recall values, where precision is calculated by

$$Precision = \frac{number\ of\ relevant\ items\ retrieved}{total\ number\ of\ items\ retrieved}$$

These curves represent averages over the 50 topics. The averaging method was developed many years ago [4] and is well accepted by the information retrieval community. The curves show system performance across the full range of retrieval, i.e., at the early stage of retrieval where the highly-ranked documents give high accuracy or precision, and at the final stage of retrieval where there is usually a low accuracy, but more complete retrieval. The use of these curves assumes a ranked output from a system. Systems that provide an unranked set of documents are known to be less effective and therefore were not tested in the TREC program.

The curves in figure 2 show that system A has a much higher precision at the low recall end of the graph and therefore is more accurate. System B however has higher precision at the high recall end of the curve and therefore will give a more complete set of relevant documents, assuming that the user is willing to look further in the ranked list.

4.2 Single-Value Evaluation Measures

In addition to recall/precision curves, there are 2 single-value measures used in TREC.

The first measure, the non-interpolated average precision, corresponds to the area under an ideal (non-interpolated) recall/precision curve. To compute this average, a precision average for each topic is first calculated. This is done by computing the precision after every retrieved relevant document and then averaging these precisions over the total number of retrieved relevant documents for that topic. These topic averages are then combined (averaged) across all topics in the appropriate set to create the non-interpolated average precision for that set.

The second measure used is an average of the precision for each topic after 100 documents have been retrieved for that topic. This measure is useful because it reflects a clearly comprehended retrieval point. It took on added importance in the TREC environment because only the top 100 documents retrieved for each topic were actually assessed. For this reason it produces a guaranteed evaluation point for each system.

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. It points to some of the other experiments run in TREC-3 where results cannot be measured completely using recall/precision measures, and discusses the tracks in TREC-4.

In all cases, readers are referred to the system papers in the TREC-3 and TREC-4 proceedings for more details.

5.2 TREC-3 Adhoc Results

The TREC-3 adhoc evaluation used new topics (topics 151-200) against two disks of training documents (disks 1 and 2). A dominant feature of the adhoc task in TREC-3 was the removal of the concepts field in the topics (see more on this in the discussion of the topics, section 3.3) Many of the participating groups designed their experiments around techniques to expand the shorter and less "rich" topics.

There were 48 sets of results for adhoc evaluation in TREC-3, with 42 of them based on runs for the full data set. Of these, 28 used automatic construction of queries, 12 used manual construction, and 2 used interactive construction.

Figure 3 shows the recall/precision curves for the 6 TREC-3 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. For more details on the various runs and procedures, please see the cited paper in the TREC-3 proceedings.

cityal -- City University, London ("Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used a probabilistic term weighting scheme similar to that used in TREC-2, but expanded the topics by up to 40 terms (average around 20) automatically selected from the top 30 documents retrieved. They also used dynamic passage retrieval in addition to the whole document retrieval in their final ranking.

INQ101 -- University of Massachusetts at Amherst ("Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used a version of probabilistic weighting that allows easy combining of evidence (an inference net). Their basic term weighting formula (and query processing) was simplified from that used in TREC-2, and they also used passage retrieval and whole document information in their ranking. The topics were expanded by 30 phrases that were automatically selected from a phrase "thesaurus" that had been previously built automatically from the entire corpus of documents.

CrnlEA -- Cornell University ("Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used the vector-space SMART system, with term weighting similar to that done in TREC-2. The top 30 documents were used in a Rocchio relevance feedback technique to massively expand (500 terms + 10 phrases) the topics. No passage retrieval was done in this run; the second Cornell run (*CrnlLA*) used their local/global weighting schemes (with no topic expansion).

westpl -- West Publishing Company ("TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System" by Paul Thompson, Howard Turtle, Bokyung Yang and James Flood) used their commercial product (WIN) which is based on the same inference method used in *INQ101*. Both passages and whole documents were

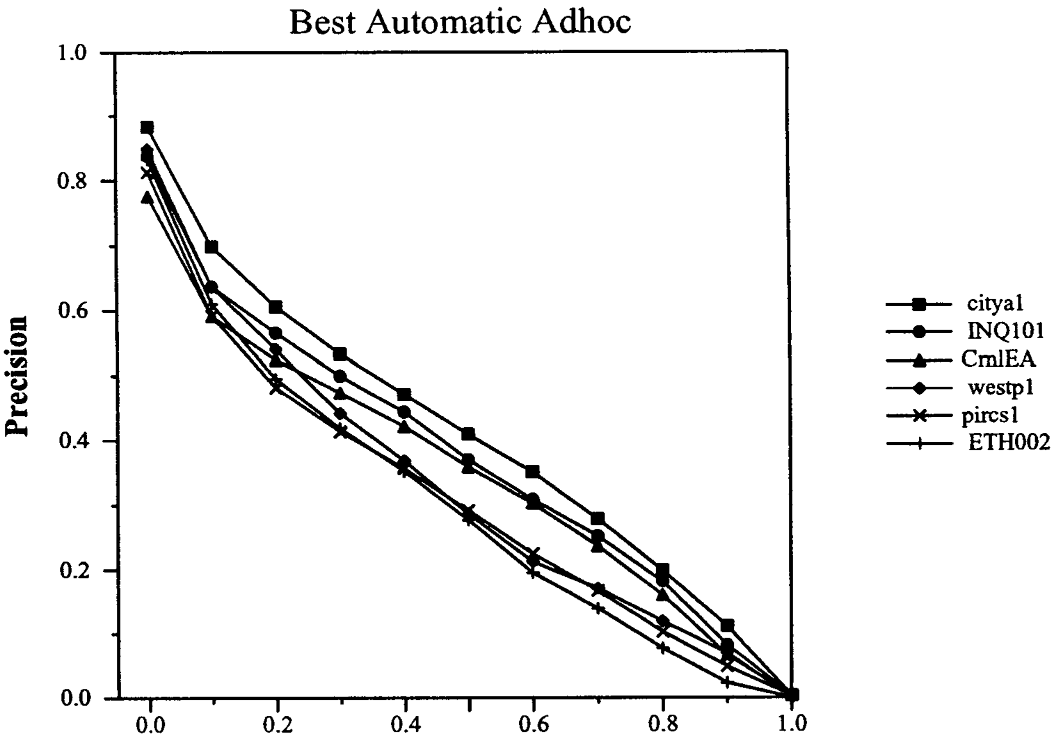

Figure 3. Best TREC-3 Automatic Adhoc Results.

used in document ranking, but only minimal topic expansion was used, with that expansion based on preconstructed general-purpose synonym classes for abbreviations and other exact synonyms.

pircs1 -- Queens College, CUNY ("TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis) used a spreading activation model on subdocuments (550-word chunks). Topic expansion was done by allowing activation from the top 6 documents in addition to the terms in the original topic. The highest 30 terms were chosen, with an average of 11 of those not in the original topic.

ETH002 -- Swiss Federal Institute of Technology (ETH) ("Improving a Basic Retrieval Method by Links and Passage Level Evidence" by Daniel Knaus, Elke Mittendorf and Peter Schäuble) used a completely new method in TREC-3 based on combining information from three very different retrieval techniques. The three techniques are a vector-space system, a passage retrieval method using a Hidden Markov model, and a "topic expansion" method based on document links generated automatically from analysis of common phrases.

The dominant new themes in the automatic adhoc runs are the use of some type of term expansion beyond the terms contained in the "less rich" (TREC-3) topics, and some form of passage or subdocument retrieval element. Note that term expansion is mostly a recall device; adding new terms to a topic increases the chances of matching the wide variation of terms usually found in relevant documents. But adding terms also increases the "noise" factor, so accuracy may need to be improved via a precision device, and hence the use of passages, subdocuments, or more local weighting.

Two main types of term expansion were used by these top groups: term expansion based on a pre-constructed thesaurus (for example the INQUERY PhraseFinder) and term expansion based on selected terms from the top X documents (as done by City, Cornell, and PIRCS). Both techniques worked well. The top 3 runs (*cityal*, *INQ101*, and *CrnlEA*) have excellent performance (see Figure 3) in the "middle" recall range (30 to 80%), with this performance likely coming from the query expansion.

The use of the top 30 documents as a source of terms, as opposed to using the entire corpus, should be sensitive to the quality of the documents in this initial set. Notably, for 6 of the 8 topics in which the *INQ101* run was superior (a 20% or more improvement in average precision) to the *cityal* run, the *INQ101* run was also superior to the *CrnlEA* run. These topics tended to have

	base run	passages	expansion	both
City	0.337	-	0.388 (15%)	0.401 (19%)
INQUERY (11 pt. average)	0.318	0.368 (16%)	0.348 (9%)	0.381 (20%)
Cornell	0.2842	0.3302 (16%)	0.3419 (20%)	-
ETH	0.2578	0.2853 (11%)	0.2737 (6%)	0.2916 (13%)
PIRCS	-	0.2764	-	0.3001 (9%)

Table 7: Comparison of Performance (Average Precision) for Passage Retrieval and Topic Expansion

fewer relevant documents, but also tended to be topics for which the systems bringing terms in manually (such as by manually selecting from a thesaurus or outside sources) also did well.

Another factor in topic expansion is the number of terms being added to the topics. The average number of terms in the queries is widely varied, with the City group averaging around 50 terms (20 terms from expansion), the INQUERY system using around 100 terms on average, and the Cornell system using 550 terms on average. This huge variation seemed to have little effect on results, largely because each group found the level of topic expansion appropriate for their retrieval techniques. The *cityal* run tended to "miss" more relevant documents than the *CrnlEA* run (7 topics were seriously hurt by this problem), but was better able to rank relevant documents within the 1000 document cutoff so that more relevant documents appeared in the top 100 documents. This better ranking could have happened because of the many fewer terms that were used, or could be caused by the use of passage retrieval in the City run.

The use of passages or subdocuments to reduce the noise effect of large documents has been used for several years in the PIRCS system. City, INQUERY and Cornell all did many experiments for TREC-3 to first determine the correct length of a passage, and then to find the appropriate use of passages in their ranking schemes. INQUERY and Cornell use overlapped passages of fixed length (200 words) as compared to City's non-overlapped passages of 4 to 30 paragraphs in length. All three systems use information from passages and whole documents retrieved rather than passage retrieval alone. (Cornell's version of this is called local/global weighting.) Both INQUERY and City combined the passage retrieval with query expansion; Cornell did two separate runs.

The *westpl* run did not use topic expansion, although a mixture of passages and whole documents was used in the final ranking of documents. The performance has suffered for this in the middle recall range. West Publishing used their production system to see how far it differed from the research systems and therefore did not want to use more radical topic expansion methods. Additionally they used a shortened topic (title + description + first sentence of narrative) because it was more similar in length to the topics submitted by their users. The *INQ101* run had 18 topics with superior performance to the *westpl* run, mostly because of new relevant documents being retrieved to the top 1000 document set. The *westpl* run was superior to the *INQ101* run for 11 topics, mostly caused by better ranking for those topics.

The *pircs1* system used both passage retrieval (subdocuments) and topic expansion. This system used far fewer top documents for expansion (the top 6 as opposed to the top 30), and this may have hurt performance. There were 22 topics in which the *INQ101* run was superior to the *pircs2* run, and these were mostly because of missed relevant documents. Even though both systems added about the same number of expansion terms, using only the top 6 documents as a source of terms for spreading activation might have provided too much focussing of the concepts.

The *ETH001* run used both topic expansion and passages, in addition to a baseline vector-space system. Both the topic expansion and the passage determination were completely new (untried) techniques; additionally there are known difficulties in combining multiple methods. In comparison to the Cornell expansion results (*CrnlEA*), the main problems appear to be missed relevant documents for all 17 of the topics where the Cornell results were superior. The ETH results were superior for 8 topics, mostly because of better ranking. Clearly this is a very promising approach and more experimentation is needed.

Table 7 shows a breakdown of improvements from expansion and passage retrieval that combines information from the non-official runs given in the individual papers. In general groups seem to be getting about 20% improvement over their own baselines (less for ETH and PIRCS), with that improvement coming in different percentages from passage retrieval or expansion, depending on the specific retrieval techniques being used.

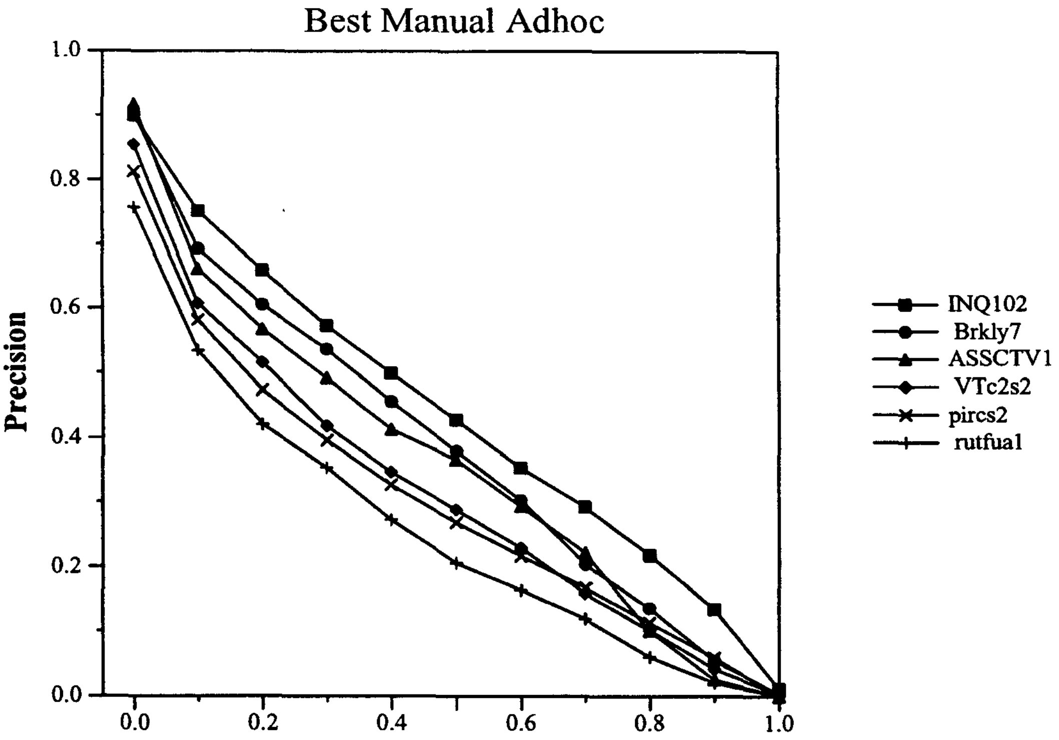

Figure 4. Best TREC-3 Manual Adhoc Results.

Figure 4 shows the recall/precision curves for the 6 TREC-3 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 the TREC-3 proceedings.

INQ102 -- University of Massachusetts at Amherst. This run is a manual modification of the *INQ101* run, with strict rules for the modifications to only allow removal of words and phrases, modification of weights, and addition of proximity restrictions.

Brkly7 -- University of California, Berkeley ("Experiments in the Probabilistic Retrieval of Full Text Documents" by William S. Cooper, Aitao Chen and Fredric C. Gey) is a modification of the *Brkly6* run, with that modification being the manual expansion of the queries by adding synonyms found from other sources. The *Brkly6* run uses a logistic regression model to combine information from 6 measures of document relevancy based on term matches and term distribution. The coefficients were learned from the training data in a manner similar to that done in TREC-2, but the specific set of measures used has been expanded and modified for TREC-3. No passage retrieval was done.

ASSCTV1 -- Mead Data Central, Inc ("Query Expansion/Reduction and its Impact on Retrieval Effectiveness" by X. Allan Lu and Robert B Keefer) is also a manual expansion of queries using an associative thesaurus built from the TREC data. The retrieval system used in *ASSCTV1* is the SMART system.

VTc2s2 -- Virginia Tech ("Combination of Multiple Searches" by Joseph A. Shaw and Edward A. Fox) used a combination of multiple types of queries, with 2 types of natural language vector-space queries and 3 types of manually constructed P-Norm (soft Boolean) queries.

pircs2 -- Queens College, CUNY. This run is a modification of the base PIRCS system to use manually constructed soft Boolean queries.

rutfual -- Rutgers University ("Decision Level Data Fusion for Routing of Documents in the TREC3 Context: A Best Cases Analysis of Worst Case Results" by Paul B. Kantor) used data fusion methods to combine the retrieval ranks from three different retrieval schemes all using the INQUERY system. Two of the schemes used Boolean queries (one with ranking and one without) and the third used the same queries without operators.

The three dominant themes in the runs using manually constructed queries are manual modification of automatically generated queries (*INQ102*), manual expansion of queries (*Brkly7* and *ASSCTV1*) and combining of multiple retrieval techniques or queries. Three runs can be compared to a "baseline" run to check the effects of manual versus automatic query construction.

INQ102, the manually modified version of *INQ101*, had a 15% improvement in average precision over *INQ101*, and 17 topics that were superior in performance for the manual system (as opposed to only 3 for the automatic system). An analysis of those topics shows that many more relevant documents were in the top 1000 documents and the top 100 documents, probably caused by manually eliminating much of the noise that was producing higher ranks for nonrelevant documents. This noise elimination could have happened because many spurious terms had been manually removed from the queries (*INQ102* had an average of about 30 terms as opposed to nearly 100 terms in *INQ101*), or could have come from the use of the proximity operators.

The *Brkly7* run, a manually expanded version of *Brkly6*, used about the same number of terms as the *INQ102* run (around 36 terms on average), but the terms had been manually pulled from multiple sources (as opposed to editing an automatic expansion as done by INQUERY). The improvement from *Brkly6* to *Brkly7* is a 34% gain in average precision, with 25 topics having superior performance in the manually expanded run. Note however that there was no topic expansion done in the automatic *Brkly6* run, so this improvement represents the results of a good manual topic expansion over no expansion at all.

The INQUERY system outperforms the Berkeley system by 14% in average precision, with much of that difference coming in the high recall end of the graph (see Figure 4). This is consistent with the difference in their topic expansion techniques in that the automatic expansion (even manually edited) is likely to bring in terms that users might not select from "non-focussed" sources.

The *ASSCTV1* run also represents a manual expansion effort, but using a pre-built thesaurus as opposed to using textual sources for the expansion. The topics were expanded to create a query averaging around 135 terms and then were run using the default Cornell SMART system. A comparison of the automatically expanded *CrnlEA* run and the manually expanded *ASSCTV1* run shows minimal difference in average precision, but superior performance in 18 of the topics for the manual expansion (as opposed to only 10 of the topics having superior performance for the automatic Cornell run). In both cases, the improvements come from finding more relevant documents because of the expansions, but different expansion methods help different topics.

The *pircs2* run is a manual query version of the baseline PIRCS system. A soft Boolean query is created from the topic, but no topic expansion is done. There is minimal difference in average precision between the two PIRCS runs, but more topics show superior performance for the soft Boolean query *pircs2* run (8 superior topics versus 4 superior topics for the topic expansion *pircs1* run). It is not clear whether this difference comes from the increased precision of the soft Boolean approach or from the relatively poor performance of the PIRCS term expansion results.

In TREC-3, as opposed to TRECs 1 and 2, the manual query construction methods perform better than their automatic counterparts. The removal of some of the topic structure (the concepts) has allowed differences to appear that could not be seen in earlier TRECs. Since topic expansion was necessary to produce top scores, the superiority of the manual expansion over no expansion in the Berkeley runs should not be surprising. Less clear is why the manual modifications in the *INQ102* run showed superior performance to the automatic run with no modifications. The likely explanation is that the automatic term expansion methods are relatively uncontrolled in TREC-3 and manual intervention plays an important role.

The last two groups in the top six systems using manual query construction used some form of combination of retrieval techniques. The Virginia Tech group (*VTc2s2*) combined the results of up to 5 different types of query construction (3 P-Norms with different P values and 2 vector-space, one short and one manually expanded) to create their results. They used a simple combination method (adding all the similarity values) and tested various combinations of query types. Their best result combined only two of the query types, one a P-Norm and one a vector-space. A series of additional runs (see paper for details) confirmed that the best method was to combine the results of the best two query techniques (the "long" vector-space and the P=2 P-Norm). They concluded that improvements from combining results only occurred when the input techniques were sufficiently different.

Although the Rutgers group (*rutfual*) used more elaborate combining techniques, they came to the same conclusion. Combining different retrieval techniques offers improvements over a single technique (over 30% for the Virginia Tech group), but the input techniques need to be more varied to get further improvements. But the more varied the individual techniques, the more

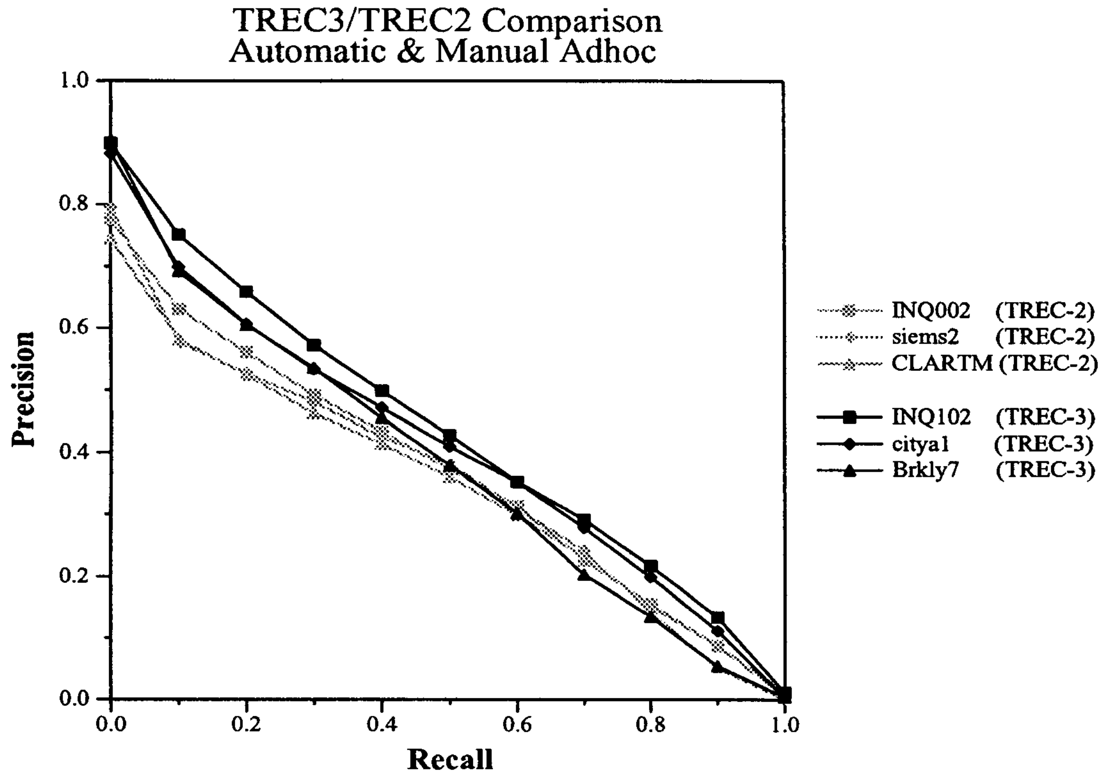

Figure 5. Comparison of Adhoc Results for TREC-2 and TREC-3

need for elaborate combining methods such as used in the *rutfual* run. The automatic *ETH001* run best exemplifies the direction needed here; first getting "good" performance for three very different but complementary techniques and then discovering the best ways of combining results.

Several comments should be made with respect to the overall adhoc recall/precision averages. First, the better results are very similar and it is unlikely that there is any statistical difference between them. The Scheffe′ tests run by Jean Tague-Sutcliffe (see paper "A Statistical Analysis of the TREC-3 Data" by Jean Tague-Sutcliffe and James Blustein in the TREC-3 proceedings) show that the top 20 category A runs (manual and automatic mixed) are all statistically equivalent at the α=0.05 level. This lack of system differentiation comes from the very wide performance variation across topics (the cross-topic variance is much greater than the cross-system variance) and points to the need for more research into how to statistically characterize the TREC results.

As a second point, it should be noted that these adhoc results represent significant improvements over TREC-2. Figure 5 shows the top three systems in TREC-3 and the top three systems in TREC-2. This improvement was unexpected as the removal of the concepts section seemed likely to cause a considerable performance drop (up to 30% was predicted). Instead the advance of topic expansion techniques caused major improvements in performance with less "user" input (the concepts). Because of the different sets of topics involved, the exact amount of improvement cannot be computed. However the Cornell group has run older systems (those used in TREC-1 and TREC-2) against the TREC-3 topics. This shows an improvement of 20% for their expansion run (*CrnlEA*) over the TREC-2 system, and this is likely to be typical for many of the systems this year.

5.3 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). A dominant feature of the adhoc task in TREC-4 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,

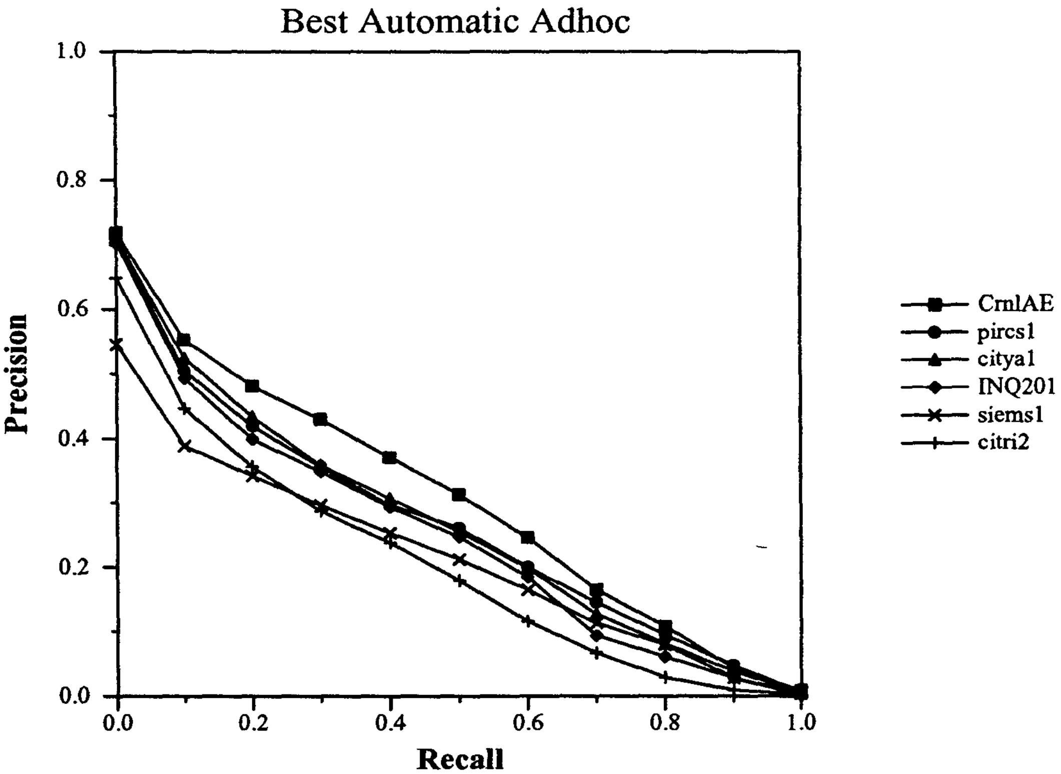

Figure 6. Best TREC-4 Automatic Adhoc Results.

and 19 used manual construction. All of the category B groups used automatic construction of the queries.

Figure 6 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 the TREC-4 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. Work has continued at Cornell in improving their radical new matching algorithm, and further information can be found in [5].

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. The experiments with short topics has continued and further results can be seen in [6].

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, and additional work comparing the use of InFinder and the use of the top documents for expansion is reported in [7].

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 a 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 107 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 6, two runs (*INQ201* and *cityal*) 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 7 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

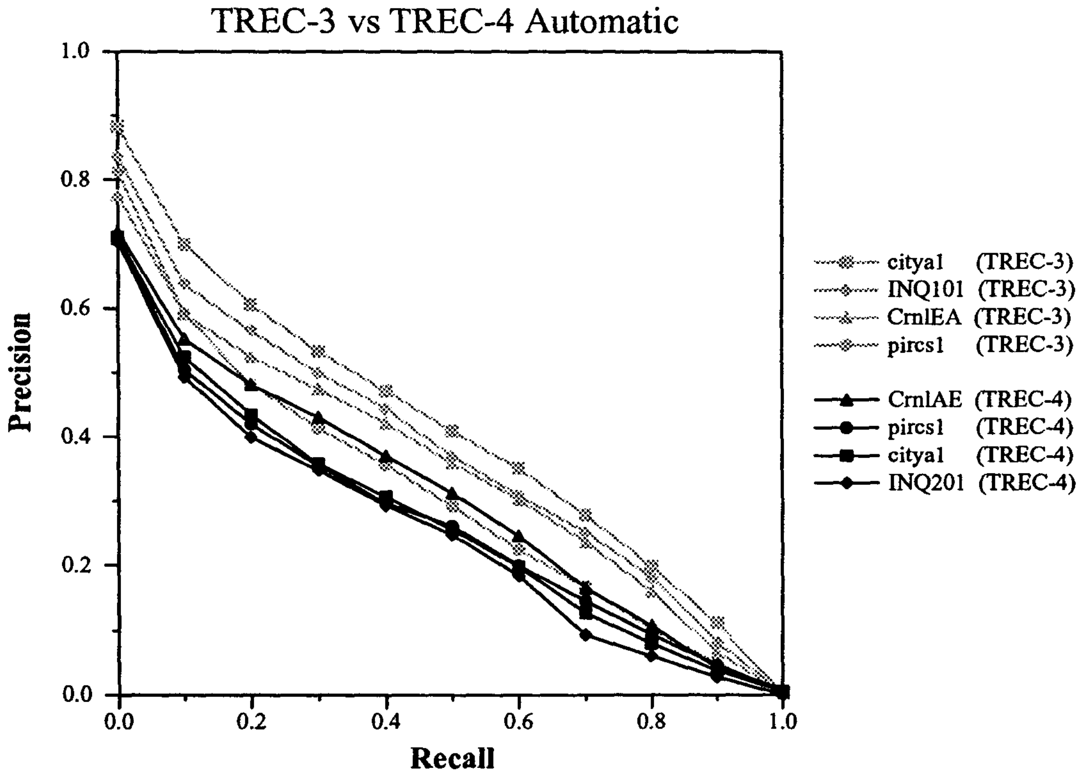

Figure 7. Comparison of Automatic Adhoc Results for TREC-3 and TREC-4

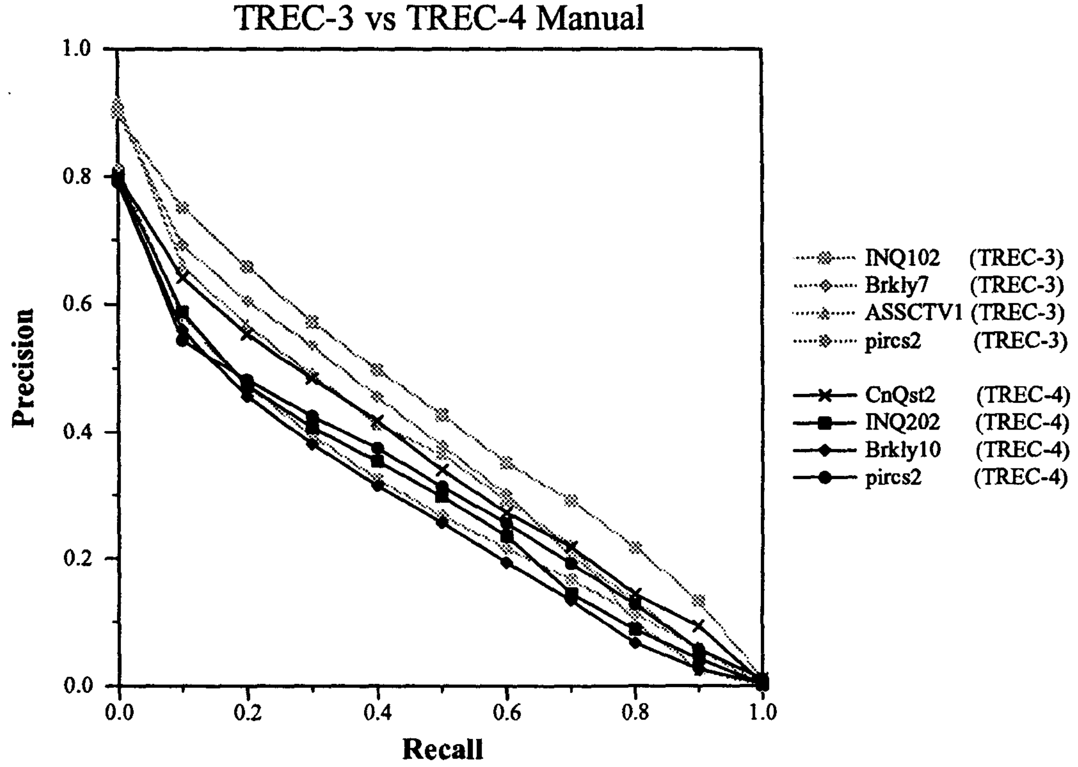

Figure 8. Comparison of Manual Adhoc Results for TREC-3 and TREC-4

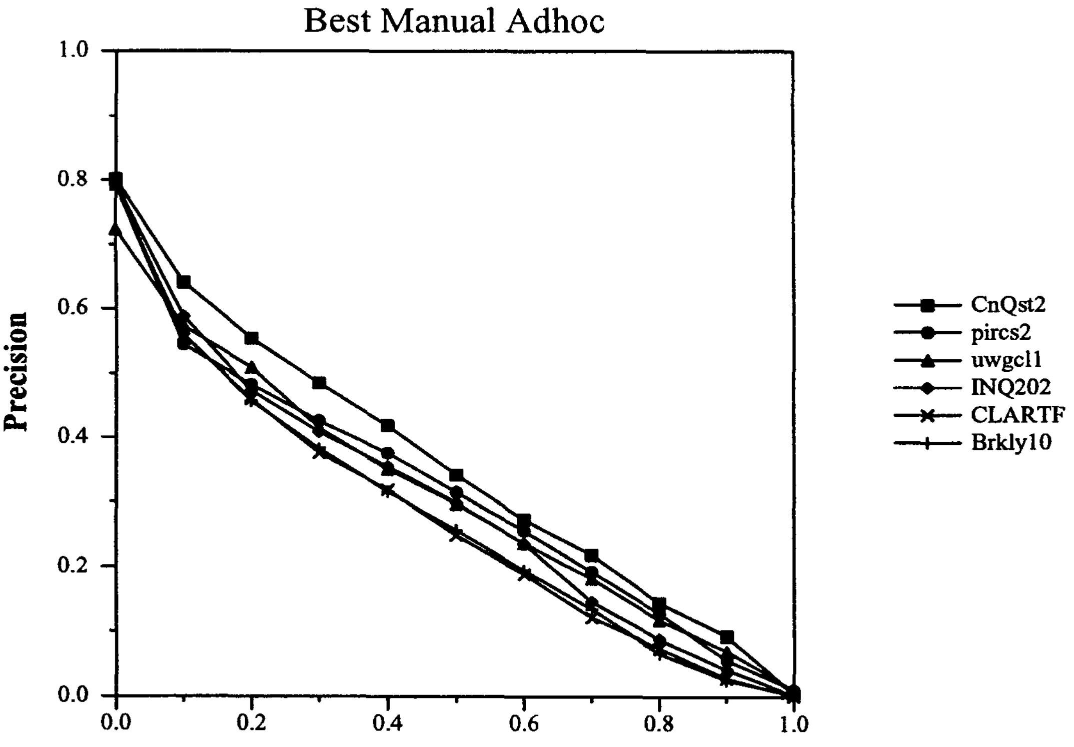

Figure 9. Best TREC-4 Manual Adhoc Results.

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 8). 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 9 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 the TREC-4 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 create intervals of text that satisfy specific conditions. The ranking algorithms rely on combining the results of increasing 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.

Brklyl0 -- 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) uses 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 *cityml* (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 group, exemplified by *uwgcll* and *Brklyl0*, used queries completely manually generated using some type of auxiliary information resource such as online dictionaries (*uwgcll*) or news databases (*Brklyl0*). The query generated for *uwgcll* uses Boolean-type restrictors, whereas the query generated for *Brklyl0* 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 *uwgcll* or *Brklyl0* 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

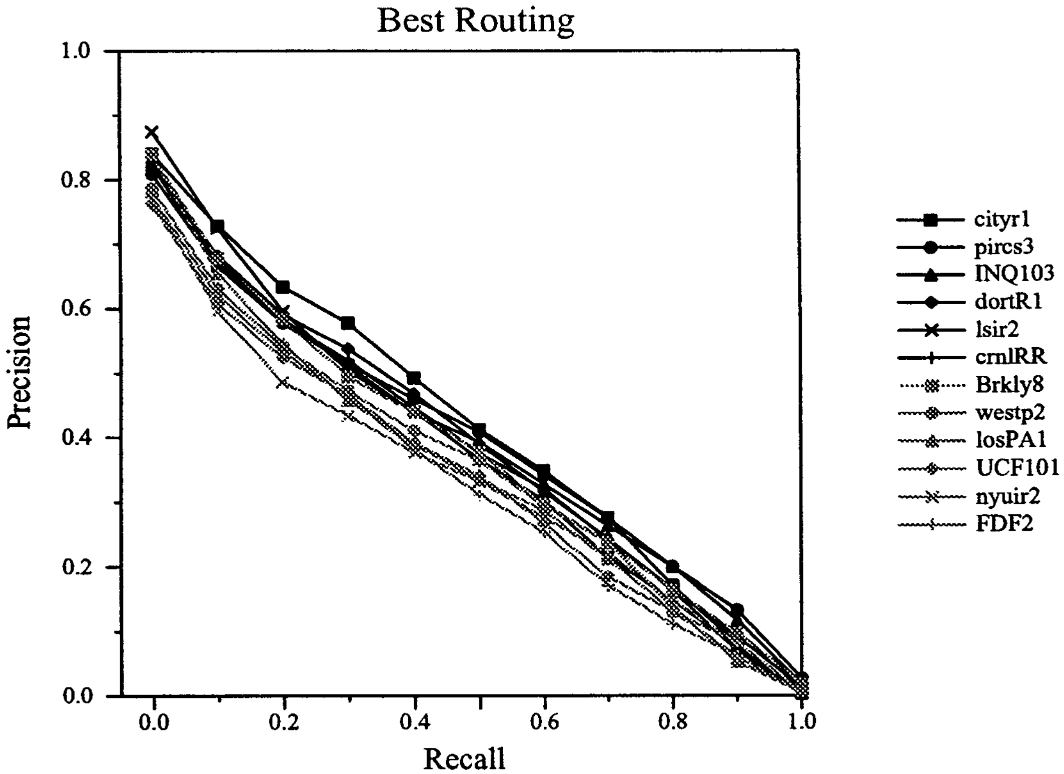

Figure 10. Best TREC-3 Routing Results.

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 in 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. Furthermore, studies have shown that users feel a need for more control of their searching and this control is absent from current automatic systems.

5.4 TREC-3 Routing Results

The routing evaluation used a subset of the training topics (topics 101-150 were used) against the disk of test documents (disk 3). Although this disk had been used in TREC-2, its use in TREC-3 was unexpected as new data had been promised. The last minute unavailability of this new data made the reuse of disk 3 necessary, but since groups had not been training with this disk (and no relevance judgments were available for this disk against topics 101-150), the routing results should not be biased by the reuse of old material.

The routing task in TREC has remained constant; however there has been a major evolution in the thrust of the research for this task. There was minimal training data for TREC-1, and most groups felt that their results were even more preliminary than for the adhoc results because the training data that was available was incomplete and inconsistent. This means that routing became a particularly interesting challenge in TREC-2 when adequate training data (the results from TREC-1 adhoc topics) became available.

The TREC-2 results therefore represent an excellent baseline of what could be achieved using traditional algorithms with large amounts of relevance information. Most notable was the effective use of the Rocchio feedback algorithm in SMART, where up to 500 new terms were added to the routing topics from the training data. Equally good results were achieved by a probabilistic system from the University of Dortmund, where only 30 terms were added, but very precise term weighting was learned from the training data. Manual construction of queries consistently gave poorer performance as the availability of training data allowed an automatic tuning of the queries that would be difficult to duplicate manually without extensive analysis.

For TREC-3, many groups made only minor modifications to their TREC-2 techniques (and concentrated on the adhoc task). There were a total of 49 sets of results for routing evaluation, with 46 of them based on runs for the full data set. Of the 46 systems using the full data set, 24 used automatic construction of queries, 18* used manual construction, and 4 used interactive query construction.

Figure 10 shows the recall/precision curves for the 12 TREC-3 groups with the highest non-interpolated average precision for the routing queries. (The grey lines only serve to allow more systems to be shown.) The runs are ranked by the average precision and only one run per group is shown (both official runs sometimes would have qualified for this set). 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 the TREC-3 proceedings.

cityr1 -- City University, London ("Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used the same probabilistic techniques as for the adhoc task, but constructed the query using a very selective set of terms (17 on average) from the relevant documents.

pircs3 -- Queens College, CUNY ("TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis) used a spreading activation model based on the topic and on terms selected from about 35% of the relevant material.

INQ103 -- University of Massachusetts at Amherst ("Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used the inference net engine (same as for the adhoc task), with topic expansion of about 60 terms selected from the relevant documents.

dortR1 -- University of Dortmund ("Routing and Adhoc Retrieval with the TREC-3 Collection in a Distributed Loosely Federated Environment" by Nikolaus Walczuch, Norbert Fuhr, Michael Pollmann and Birgit Sievers) used the SMART retrieval system with a Rocchio relevance feedback expansion adding 12% new terms and 4% new phrases from the training documents.

lsir2 -- Bellcore ("Latent Semantic Indexing (LSI): TREC-3 Report" by Susan Dumais) used the latent semantic indexing system to construct a reduced dimension vector centroid of the relevant documents (no use was made of the topics).

CrnlRR -- Cornell University ("Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used the vector-space SMART system and a basic Rocchio relevance feedback algorithm adding about 300 terms and 30 phrases to the topic.

Brkly8 -- University of California, Berkeley ("Experiments in the Probabilistic Retrieval of Full Text Documents" by William S. Cooper, Aitao Chen and Fredric C. Gey) used only the relevant documents to select a large number of terms (average 1,357 terms/topic) which were combined and weighted using a logodds formula. A chi-square test was used to select the terms.

westp2 -- West Publishing Company ("TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System" by Paul Thompson, Howard Turtle, Bokyung Yang and James Flood) used their commercial product (WIN), but expanded the topics using up to 50 terms from specially selected parts of relevant documents.

losPA1 -- Logicon, Inc. ("Research in Automatic Profile Creation and Relevance Ranking with LMDS" by Julian A. Yochum) constructed profiles based on the top 10 selected terms from the relevant documents, with term selection based on binomial probability distributions. The profile was used to select all documents containing any of those terms and the documents were then ranked using a weighting formula.

UCF101 -- University of Central Florida ("Using Database Schemas to Detect Relevant Information" by James Driscoll, Gary Theis and Gene Billings) manually constructed entity-relationship (ER) schemas for each topic and also manually created synonym lists for each labelled component in the ER schema. These schemas and lists were then used to select and rank documents.

nyuir2 -- New York University ("Natural Language Information Retrieval: TREC-3 Report" by Tomek Strzalkowski, Jose Carballo and Mihnea Marinescu) used NLP techniques to discover syntactic phrases in the documents. Both single terms and phrases were indexed and specially weighted. The *nyuir2* run used topic expansion based on the relevant documents.

FDF2 -- Paracel, Inc. ("The FDF Query Generation

* 11 of these runs were abbreviated runs from one group

Workbench" by K.I. Yu, P. Scheibe and F. Nordby) used a series of tools to generate profiles. These tools used statistical methods to create several alternative queries, and automatically evaluated the queries against the training data to select the best query for each topic.

The recall/precision curves shown in Figure 10 are very close in performance for the routing, with the Scheffe´ tests done by Jean Tague-Sutcliffe showing that there is no significant differences between the top 22 runs. It is, however, useful to look at the results on a per topic basis to find trends in performance across techniques.

The main issue for the TREC-3 routing runs is how to best select terms for topic expansion. Note that for the adhoc task the main issue was how to expand a topic beyond its original terms, hopefully with as little loss in precision as possible. For the routing task, however, the pool of terms for expansion is easily determined (i.e., the terms in the relevant documents), and the problem is how to select terms from this very large pool. Correspondingly, the major differences in results between the routing runs are not how many relevant documents were "missed" (as for the adhoc task), but how well the relevant documents were ranked.

An example of this is a comparison between the two City runs. The *cityr1* system used all relevant documents to select the top T terms, where T varied between 3 and 100 (average 47). Then they used the training material to optimize the queries, selecting only those terms that improved results. On average only about 17 terms were used in an optimized query. The unoptimized version of these queries was used at the *cityr2* run (not shown in Figure 10), which did not work as well. The difference in average precision between the two runs is only about 12%, but the optimized *cityr1* run had 14 superior topics (topics with a 20% or more improvement in average precision), all caused by better ranking (more relevant documents moved into the top 100 documents from the top 1000 documents). A similar comparison can be made between the *cityr1* run and the *pircs3* run. Even though there were more relevant documents found by the *pircs3* technique, the *cityr1* run had 15 superior topics (versus 7 superior for *pircs3*), all caused by better ranking.

The ability to assign better ranks to relevant documents is not strictly tied to being highly selective of terms. A comparison of the *cityr1*, *pircs3*, *INQ103* and *CrnlRR* runs shows that the INQUERY and PIRCS techniques both used an average of around 100 terms in their queries and retrieved the largest number of relevant documents in the top 1000 documents. The *cityr1* run, with only about 17 terms, missed a few relevant documents, but did a much better job of ranking the ones they found. However, even though the *CrnlRR* run used a massive expansion of greater than 300 terms, the *CrnlRR* runs were stronger in ranking than in finding relevant documents. A comparison of the *INQ103* run to that of Cornell shows that Cornell had 12 "inferior" topics, mostly due to missed relevant documents, and 9 superior topics, mostly due to better ranking. Clearly the appropriate number of terms to use in a routing query varies across retrieval techniques. This same result was seen in the adhoc task, where the appropriate number of expansion terms also varied across systems.

The top routing results tend to fall into three categories -- those groups that used minimal effort in selecting terms (*CrnlRR*, *lsir2*), those groups that selected terms based on using only a portion of the relevant material (*pircs3* and *westp2*), and those groups that used all the material, but carefully selected terms (*cityr1*, *INQ103*, *brkly8* and *losPA1*).

Both the Cornell runs and the LSI runs were repeats of their TREC-2 techniques. The LSI runs tested using only the topic to create a query (no expansion) versus using all the relevant documents (no topic) to create a centroid for use as the query (the *lsir2* run). There is a 30% improvement using the relevant documents only. The Cornell runs used both the topic and a massive Rocchio relevance feedback expansion (300+ terms). Both groups used techniques based on a vector-space model (loosely based for the LSI technique), and this model appears to be able to effectively rank documents despite very massive queries. The strength of the Cornell ranking was mentioned before, but the LSI ranking is comparable or even better (18 superior topics for LSI, 9 for Cornell, all caused by better ranking).

Two groups (the PIRCS system and the WIN system from West) experimented with using only portions of the training data. This is mostly an efficiency issue, but also serves as a term selection method. The *pircs4* run (not shown in Figure 10) used only short documents, where short is defined as not more than 160 unique nonstop stems. This run did somewhat worse than the *pircs3* run, where a combination of these short documents and the top 2400 subdocuments were used. In both runs many fewer documents were used (12% and 35% of the relevant material respectively), yet the results were excellent. The West group tried multiple experiments using various segments of the relevant documents (best documents only, best 200 paragraphs, and best top paragraph). Up to 50 terms were added using a combination of the various approaches, with selection of approaches done on a per topic basis. This selective use of material caused some relevant documents to be missed. A comparison of the *westp2* run and the

INQ103 run shows that the 12 topics in which the *INQ103* run was superior were mostly caused by new relevant documents being found, whereas the 7 topics in which the *westp2* run was superior were all caused by better ranking.

Four groups (*cityr1*, *INQ103*, *brkly8*, and *LosPA1*) used all the relevant documents, but made careful selection of the terms to use. The City results have already been discussed. The *INQ103* run used an adaptation of the Rocchio algorithm with their inference engine technique. A statistical formula was used to select the top 32 terms to use for expansion for each topic, and then 30 additional terms were selected based on their proximity to those terms already selected. This technique retrieved a large number of the relevant documents into the top 1000 slots, but had more difficulties doing the ranking within that set. The *brkly8* run selected an average of over 1000 terms by using a chi-square test to indicate which stems were statistically associated with document relevance to a topic. These terms were weighted and used as the query. The *losPA1* run used a similar technique, calculating a binomial probability to select the top 1000 terms, selecting a pool of documents using an OR of the top 10 terms, and then scoring the documents using a weighting algorithm based on occurrances of the 1000 terms in those documents. If results from these two systems are compared to the more traditional *INQ103* method, it seems that the strengths of these methods are in the ranking, with some problems in missing relevant documents.

As was the case in earlier TRECs, the manual construction of routing queries was not very competitive with automatic query construction. The manual *INQ104* run, consisting of a merge of the *INQ103* queries and a manually edited version of these queries was little different in results from the *INQ103* run. An exception to this was the reasonable results of the *UCF101* run. This run combined manually constructed detailed entity-relationship schema with manually constructed synonym lists. These were run against the documents, producing results that are comparable with the automatic results.

There is some improvement in overall routing results compared with those from TREC-2. This is mostly shown by the comparative position of the *CrnlRR* run, which was the "top-ranked" run in TREC-2, and now is more the "middle of the pack."

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 criteria. 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 11 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 the TREC-4 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). They 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 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. 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

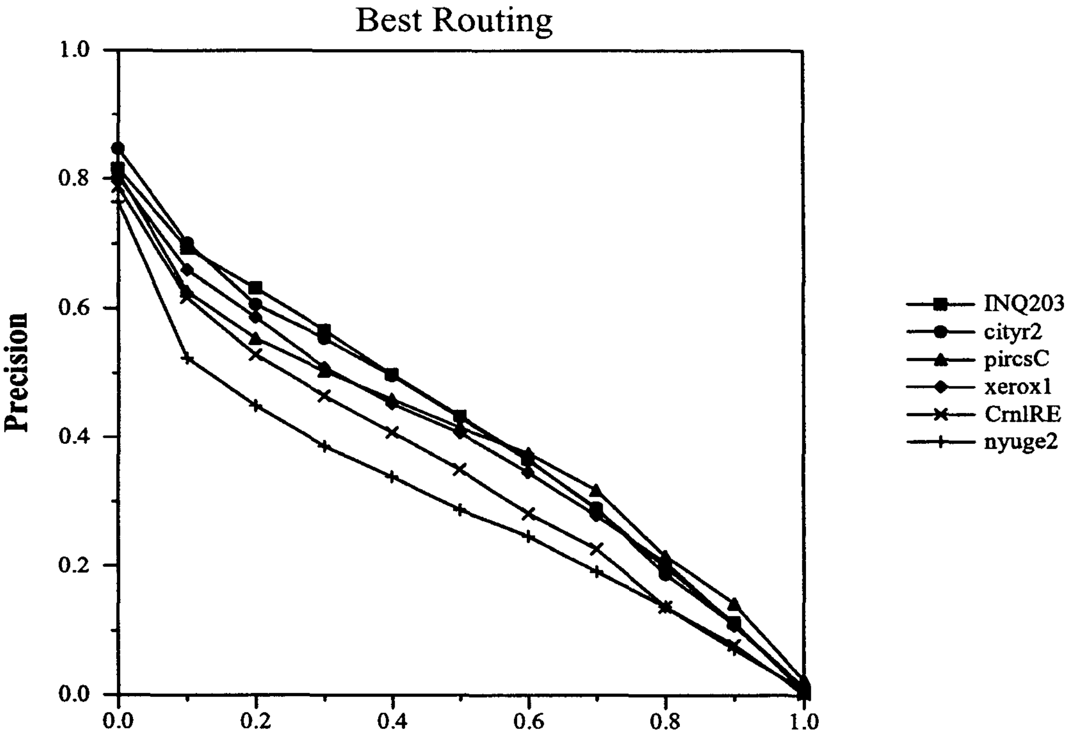

Figure 11. Best TREC-4 Routing Results.

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 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 is interesting that the six groups shown in Figure 11 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 cooccurance information for the whole collection, not just the relevant and nonrelevant documents. The GE/NYU group did a massive 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 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 11 when broken into the two subsets. This is shown in Figure 12. 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 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 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 13). However, the TREC-4 topics were more difficult, particularly the FR topics. Despite the harder topics, many of the systems achieved performance improvements, particularly 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 informal Spanish test was run in TREC-3, but the data arrived late and few groups were able to take part. A formal multilingual track was formed in TREC-4 and 10 groups took part. 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.

In TREC-3, four groups tried this task. Since there was no training data for testing (similar to the startup problems for TREC-1), the groups used simple techniques. No graphs are shown for the results since there were not enough groups to create a sufficient relevance pool. For more details on the individual experiments,

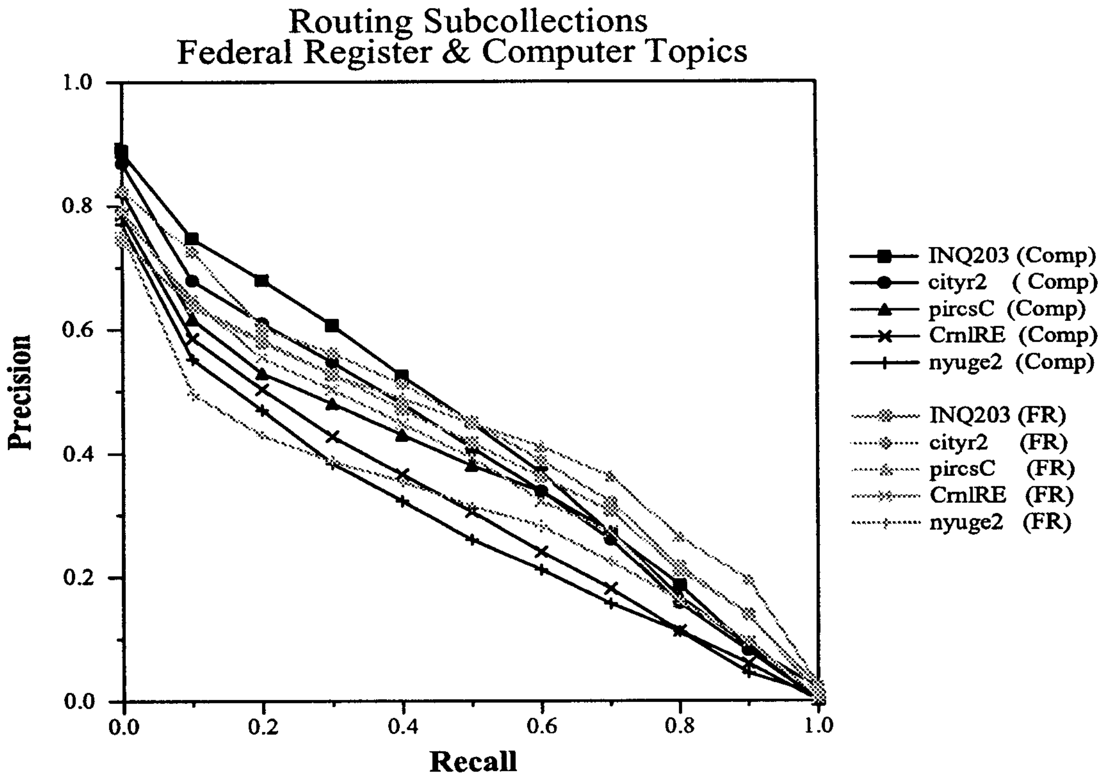

Figure 12. Comparison of Results for Federal Register Topics and Computer Topics

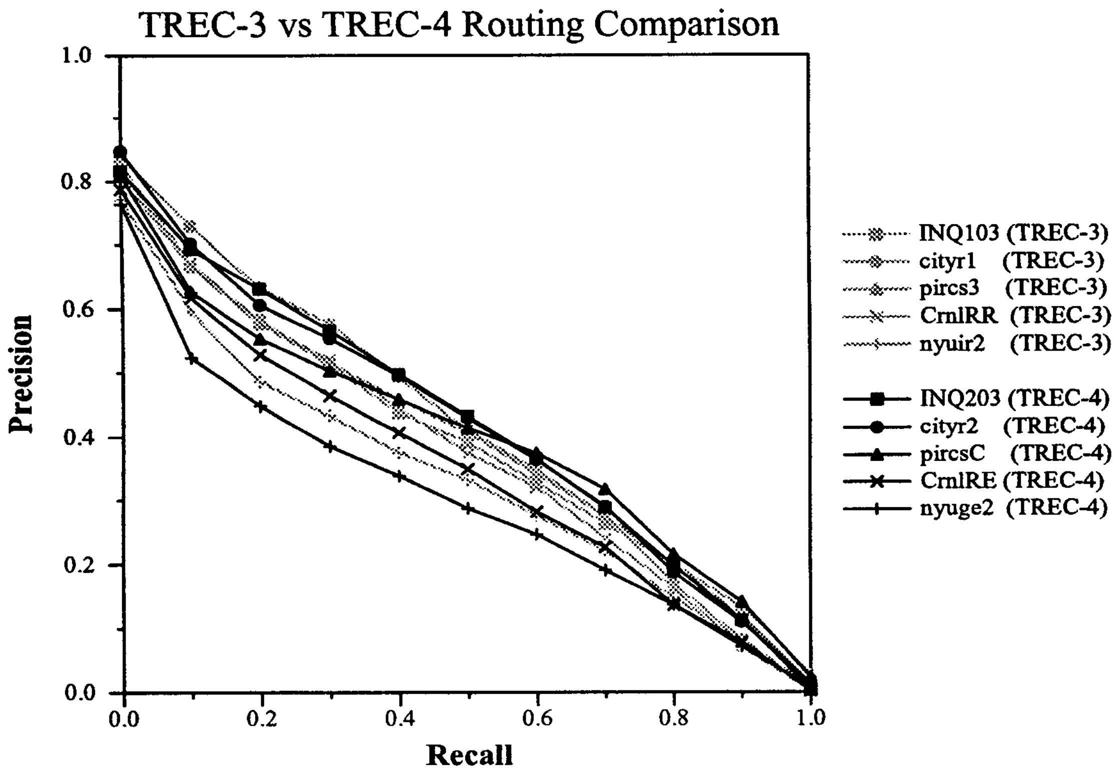

Figure 13. Comparison of Routing Results for TREC-3 and TREC-4

see the cited papers in the TREC-3 proceedings.

CrnlVS, CrnlES -- Cornell University ("Automatic Query Expansion Using SMART: TREC-3 by Chris Buckley, Gerard Salton, James Allan and Amit Singhal) used a baseline SMART run (*CrnlVS*) and a SMART run with massive topic expansion (*CrnlES*) similar to their TREC-3 English adhoc run. A simple stemmer and a stoplist of 342 terms were used.

SIN002, SIN001 -- University of Massachusetts at Amherst ("Document Retrieval and Routing Using the INQUERY System" by John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar) used the INQUERY system, with *SIN001* being a manually modified version of a basic TREC-3 automatic INQUERY run (*SIN002*). There was no topic expansion. A Spanish stemmer produced a 12% improvement in later experiments.

DCUSP1 -- Dublin City University ("Indexing Structures Derived from Syntax in TREC-3: System Description" by Alan Smeaton, Ruairi O'Donnell and Fergus Kelledy) used a trigram retrieval model, with weighting of the trigrams from traditional frequency weighting. A Spanish stemmer based on the Porter algorithms was also used.

erims1 -- Environmental Research Institute of Michigan ("Using an N-Gram-Based Document Representation with a Vector Processing Retrieval Model" by William Cavnar) used a quad-gram retrieval model, also with weighting using some of the traditional weighting mechanisms.

The major result from this very preliminary experiment in a second language was the ease of porting the retrieval techniques across languages. Cornell reported that only 5 to 6 hours of system changes were necessary (beyond creation of any stemmers or stopword lists).

Three of these four groups also did the Spanish task in TREC-4, along with 7 new groups. This time there was training data (the results of TREC-3), and groups were able to do more elaborate testing. Figure 14 shows the recall/precision curves for these 10 TREC-4 groups, ordered by non-interpolated average precision. The cited papers are in the TREC-4 proceedings.

UCFSP1 -- University of Central Florida ("Multilingual 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.

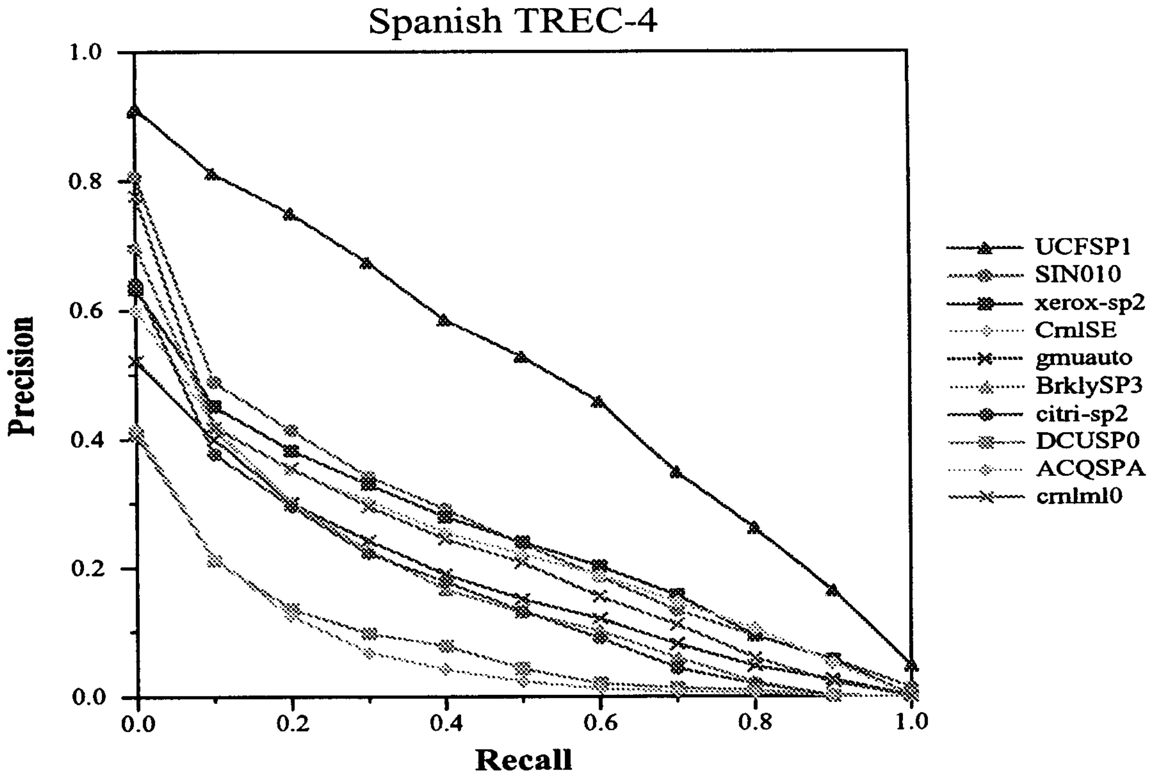

Figure 14. Results of TREC-4 Spanish Track

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 a 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. This was a new track proposed during the TREC-3 conference. 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 15 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 the TREC-4 proceedings.

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 final ranking of 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 percent 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 addressed in TREC-3 by the Siemens group (see paper "The Collection Fusion Problem" by Ellen Voorhees, Narendra Gupta and Ben Johnson-Laird in the TREC-3 proceedings). This group examined two different collection fusion techniques and was able to obtain results within 10% of the average precision of a run using a merged collection index. This type of investigation is important for real-world collections, and also to allow researchers to take advantage of possible variations in retrieval techniques for heterogeneous collections.

The general interest in this area led to the formation of a formal track in TREC-4. 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

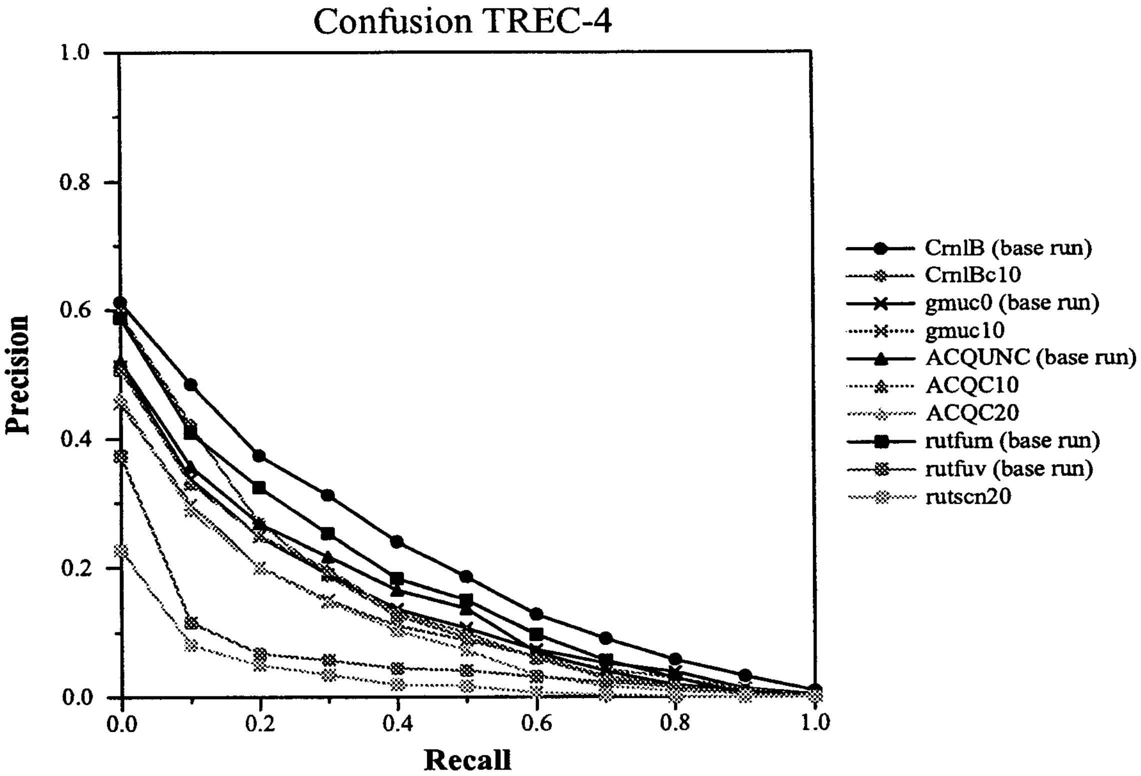

Figure 15. Results of TREC-4 Confusion Track

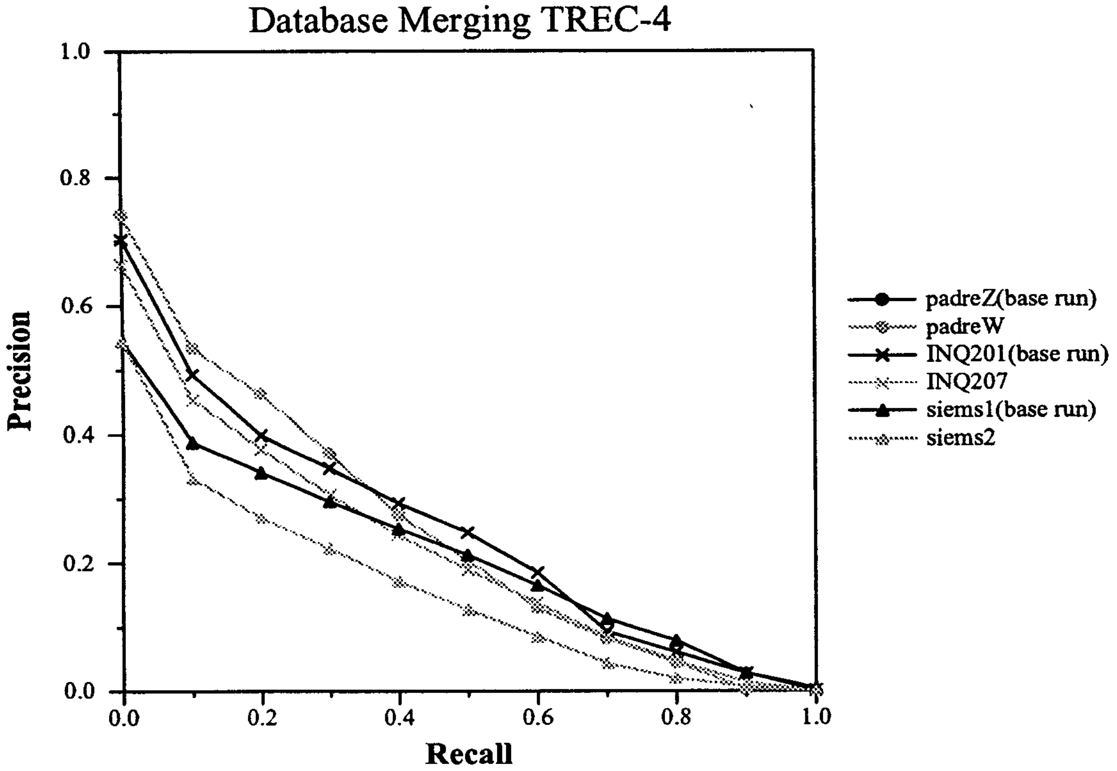

Figure 16. Results of TREC-4 Database Merging Track

these results, along with a baseline run treating the subcollections as a single collection.

Figure 16 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 the TREC-4 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 [8]. 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).

More work needs to be done in this area, and hopefully more groups will try this track in TREC-5.

6.4 The Filtering Track

The filtering track represents a variation of the current routing task. For several years some participants have been concerned about the definition of the routing task, and a few groups experimented in TREC-3 with an alternative method of evaluating routing. For details on one of these experiments, see the paper "TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS" by K.L. Kwok, L. Grunfeld and D.D. Lewis in the TREC-3 proceedings.

In TREC-4 the track was formalized and 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, see the paper "The TREC-4 Filtering Track" by David Lewis (in the TREC-4 proceedings).

Figure 17 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 the TREC-4 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

The largest area of focussed experimentation in TREC-3 was in interactive query construction, with four groups participating. One of the questions addressed by these groups was how well humans could perform the routing task, given a "rules-free" environment and access to the training material. The larger issue addressed by these experiments, however, was the entire interaction process in retrieval systems, since the "batch

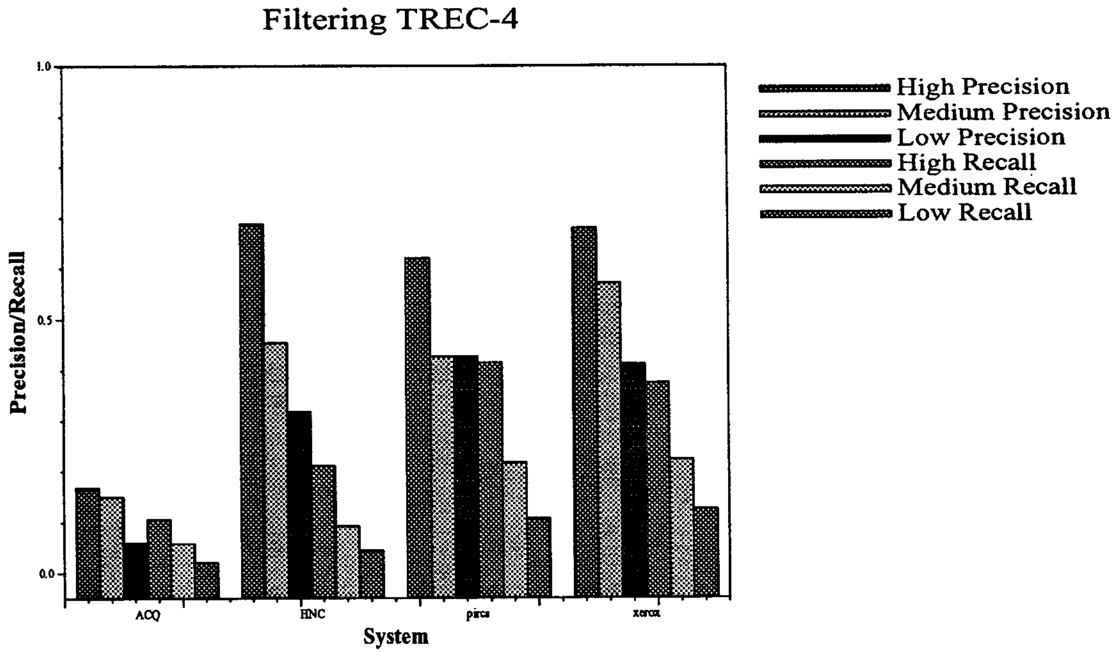

Figure 17. Results of TREC-4 Filtering Track

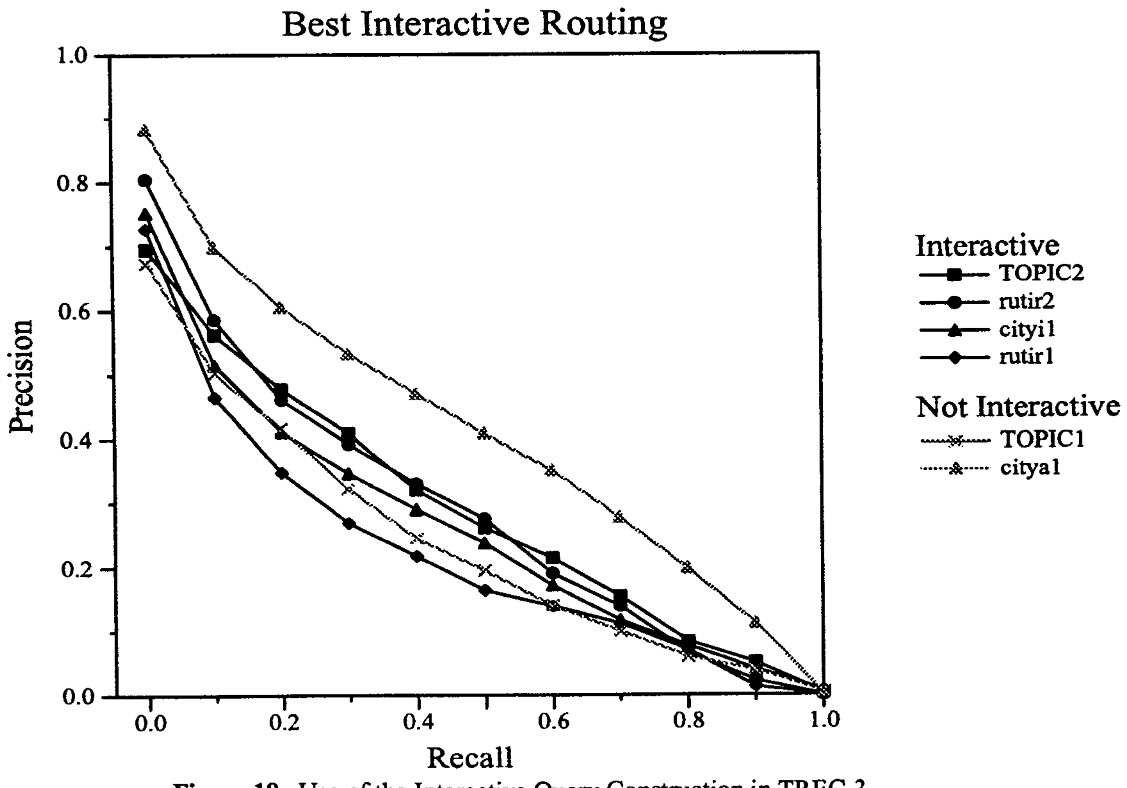

Figure 18. Use of the Interactive Query Construction in TREC-3

mode" evaluation of TREC does not reflect the way that most systems are used.

Figure 18 shows the three sets of results for the category A interactive runs in TREC-3, plus several baseline runs for comparison. A short summary of the systems follows, and readers are referred to the individual papers in the TREC-3 proceedings for more details.

TOPIC2 -- Verity, Inc. ("Interactive Document Retrieval Using TOPIC (A report on the TREC-3 experiment)" by Richard Tong) used 12 Verity staff members ranging in search experience using TOPIC from novice to expert to build their queries. The initial queries were the manual-constructed queries used by Verity in TREC-2, and the results from these queries are shown in Figure 18 as *TOPIC1*. The searchers then improved the initial queries by periodically evaluating their "improved" queries against the training data. When sufficiently improved scores were achieved, the queries were declared final and used for TREC-3.

rutir1, *rutir2* -- Rutgers University ("New Tools and Old Habits: The Interactive Searching Behavior of Expert Online Searches using INQUERY" by Jurgen Koenemann, Richard Quatrain, Colleen Cool and Nicholas Belkin) used the INQUERY system and had 10 experienced online searchers with no prior experience using that system build their queries. The entire query building process was restricted to 20 minutes per topic, and used the training data both for automatic relevance feedback (if desired) and for the searchers to check if a given retrieved document was relevant (as opposed to periodically evaluating their results). At some point during the 20 minute limit the queries were declared finished by the searchers and the results from these queries are shown in Figure 18 as *rutir1*. As a comparison, the experimenters also did the task themselves (*rutir2*).

cityi1 -- City University, London ("Okapi at TREC-3" by S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford) used the OKAPI team as searchers. The initial query was manually generated using traditional operations. The retrieved documents (or a brief summary of them) were then displayed, and searchers checked the relevance judgments (generally viewing 10 or 12 relevant documents). Automatic relevance feedback was then applied and the searchers could choose to modify the resulting query or not (35 of the 50 topics were modified). Multiple iterations could be done before a decision was made on the final query.

Not shown in Figure 18 is a category B interactive result from the University of Toronto ("Interactive Exploration as a Formal Text Retrieval Method: How Well can Interactivity Compensate for Unsophisticated Retrieval Algorithms" by Nipon Charoenkitkarn, Mark Chignell and Gene Golovchinsky). This group developed their TREC experiments from what was initially a browsing system. Boolean operators and promixity operators were used to construct the initial query. The queries were then "loosened" until around 1000 documents were retrieved. Then the results of these queries were run against the training data and reviewed, with changes possibly made to the query based on retrieval results.

As a group, the interactive results were considerably worse than the automatic routing results. This was somewhat unexpected since in all four cases the queries could be classified as the best manual queries possible. Although no definite reasons have been cited for this, the likely cause is the very strong performance of the automatic systems given the large amounts of training data.

A comparison of the City interactive run (*cityi1*) and the City automatic run (*citya1*) illustrates the problems. For BOTH runs, the query lengths were short, an average of around 17 terms. Only about 20% of these terms were in common, i.e., the searchers (*cityi1*) and the "computer" (*citya1*) picked different sets of terms. The difference in the results from these queries, however, is very large, as shown in Figure 18. The automatic run has a 63% improvement in average precision, and 33 topics with superior results (a 20% or more improvement in average precision) versus one topic with inferior results.

Regardless of the poorer performance, all four groups were able to draw interesting conclusions about their own interactive experiments. The Verity group found a 24% improvement in results (*TOPIC1* to *TOPIC2*) that can be obtained by humans using the training material over the (manually created) initial query. Other groups were able to gain insight into better tools needed by their system or insight into how online searchers handle the new techniques available. Of particular interest are the reports in these papers about the detailed human/computer interactions, as this provides insight on how systems might work in an operational setting.

A formal 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.

Three of the four groups that did interactive query construction in TREC-3 also participated in TREC-4. Seven new groups also tried this track. The cited papers are in the TREC-4 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-3" 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 od 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.

The various results presented from this track were very interesting and useful. However, all participants were concerned about the difficulties of comparing results. One of the major outcomes of this track in TREC-4 was the 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 in 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 TREC-3 and TREC-4 evaluations have produced many important experiments for all the participating groups. Some general conclusions can be drawn from each evaluation effort.

The main conclusions that can be drawn from TREC-3 are as follows:

- Automatic construction of routers or filters from training data was very effective, much more effective than manual construction of these types of queries. This held even if the manual construction was based on unrestricted use of the training data.
- Expansion of the "less-rich" TREC-3 topics was highly successful, using either automatic topic expansion, manual topic expansion, or manually modified versions of automatically expanded topics. Many different techniques were effective, with research just beginning in this new area.
- The use of passage retrieval, subdocuments, and local weighting brought consistent performance improvements, especially in the adhoc task. Experiments in TREC-3 showed continued improvement coming from various methods of using these techniques to improve ranking.
- Preliminary results suggested that the extension of basic English retrieval techniques into another language (in particular Spanish) did not appear difficult. TREC-3 represented the first large-scale test of this portability issue.

Do these conclusions hold in the real world of text retrieval? Certainly the use of automatic construction of routers will work in any environment having reasonable amounts of training material. Of greater question is the transferability of the adhoc results. Two particular issues need to be addressed here. First, even though the topics in TREC-3 were 'less-rich", they were still considerably longer than most queries used in operational settings. A couple of sentences is likely to be the maximum a user is willing to type into a computer, and it is unclear if the TREC topic expansion methods would work on these shorter input strings. Shorter topics may also need different techniques of passage retrieval and local weighting. TREC-4 addressed this issue by using appropriately shorter topics.

The second mismatch of the TREC-3 (and TREC-4) results to the real-world is the emphasis on high recall in TREC. Requesting 1000 ranked documents and calculating the results on these goes well beyond average user needs. Karen Sparck Jones addressed this issue by looking at retrieval performance based only on the top 30 documents retrieved [9], and has updated her conclusions for TREC-3 and TREC-4 in appendices to the appropriate proceedings. An improvement of 20% in precision at this cutoff means that six additional relevant documents will be returned to the user, and this is likely to be noticeable by many users. Many of the techniques used in TREC produced this difference; additionally some of the tools being investigated in TREC, such as the topic expansion tools, will make query modification much easier for the average user.

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-build 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 -- 4 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

- 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 Software and Intelligent Systems Technology 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

[1] 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.

[2] Harman D. (Ed.). (1996). *The Fourth Text REtrieval Conference (TREC-4).* National Institute of Standards and Technology Special Publication, in press.

[3]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.

[4] Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval.* New York, NY.: McGraw-Hill.

[5] Singhal A., Buckley C., and Mitra M. (1996). Pivoted Document Length Normalization. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, in press.

[6] Kwok K.L. (1996). A New Method of Weighting Query Terms. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, in press.

[7] Xu J. and Croft W.B. (1996). Query Expansion using Local and Global Document Analysis. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, in press.

[8] 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.

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