Overview of the Second Text REtrieval Conference (TREC-2)

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. Introduction

In November of 1992 the first Text REtrieval Conference (TREC-1) was held at NIST [Harman 1993]. The conference, co-sponsored by ARPA and NIST, brought together information retrieval researchers to discuss their system results on a new large test collection (the TIPSTER collection. This was the first time that such groups had ever compared results on the same data using the same evaluation methods and represented a breakthrough in cross-system evaluation in information retrieval. It was also the first time that most of these groups had used such a large test collection and therefore required a major effort by all groups to scale up their retrieval techniques.

The overall goal of the TREC initiative is to encourage research in information retrieval using large-scale test collections. It is hoped that by providing a very large test collection, and encouraging interaction with other groups in a friendly evaluation forum, new momentum in information retrieval will be generated. Because of the NIST involvement, groups with commercial retrieval products have participated in TREC, leading to increased technological transfer between the research labs and the commercial products. TREC has also provided a state-of-the-art showcase of retrieval methods for ARPA clients.

Whereas the TREC-1 conference demonstrated a wide range of different approaches to the retrieval of text from large document collections, the results should be viewed as very preliminary. Not only were the deadlines for results very tight, but the huge increase in the size of the document collection required significant system rebuilding by most groups. Much of this work was a system engineering task: finding reasonable data structures to use, getting indexing routines to be efficient enough to index all the data, finding enough storage to handle the large inverted files and other structures, etc. Still, the results showed that the systems did the task well, and that automatic construction of queries from the topics did as well as, or better than, manual construction of queries.

The second TREC conference (TREC-2) occurred in August of 1993, less than 10 months after the first conference. In addition to twenty-two of the TREC-1 groups, nine new groups took part, bringing the total number of participating groups to 31. Many of the original TREC-1 groups were able to "complete" their system rebuilding and tuning, and in general the TREC-2 results show significant improvements over the TREC-1 results.

This paper provides an overview of the TREC-2 conference, including a review of the TREC task, a brief description of the test collection being used, and an overview of the results. The papers from the individual groups should be referred to for more details on specific system approaches.

2. The TREC Task

2.1 Introduction

TREC is designed to encourage research in information retrieval using large data collections. Two types of retrieval are being examined -- retrieval using an "adhoc" query such as a researcher might use in a library environment, and retrieval using a "routing" query such as a profile to filter some incoming document stream. The TREC task is not tied to any given application, and is not primarily concerned with interfaces or optimized response time for searching. However it is helpful to have some potential user in mind when designing or testing a retrieval system. The model for a user in TREC is a dedicated searcher, not a novice searcher, and the model for the application is one needing monitoring of data streams for information on specific topics (routing), and the ability to do adhoc searches on archived data for new topics. It should be assumed that the users need the ability to do both high precision and high recall searches, and are willing to look at many documents and repeatedly modify queries in order to get high recall. Obviously they would like a system that makes this as easy as possible, but this ease should be reflected in TREC as added intelligence in the system rather than as special interfaces.

Since TREC has been designed to evaluate system performance both in a routing (filtering or profiling) mode, and in an adhoc mode, both functions need to be tested.

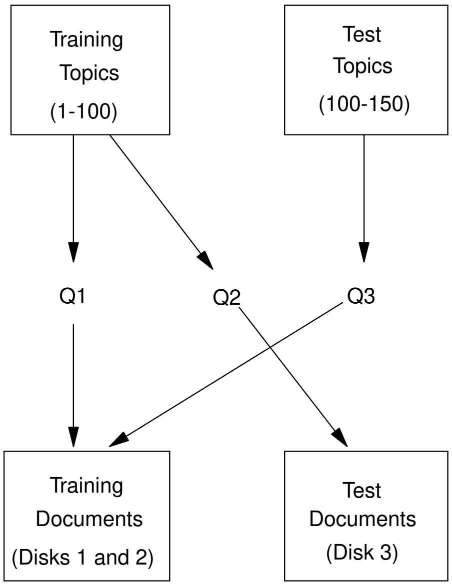


Figure 1 -- The TREC Task

The test design was based on traditional information retrieval models, and evaluation used traditional recall and precision measures. The above diagram of the test design shows the various components of TREC (fig. 1).

This diagram reflects the four data sets (2 sets of topics and 2 sets of documents) that were provided to participants. These data sets (along with a set of sample relevance judgments for the 100 training topics) were used to construct three sets of queries. Q1 is the set of queries (probably multiple sets) created to help in adjusting a system to this task, to create better weighting algorithms, and in general to train the system for testing. The results of this research were used to create Q2, the routing queries to be used against the test documents. Q3 is the set of queries created from the test topics as adhoc queries for searching against the training documents. The results from searches using Q2 and Q3 were the official test results sent to NIST.

2.1 Specific Task Guidelines

Because the TREC participants used a wide variety of indexing/knowledge base building techniques, and a wide variety of approaches to generate search queries, it was important to establish clear guidelines for the evaluation task. The guidelines deal with the methods of indexing/knowledge base construction, and with the methods of generating the queries from the supplied topics. In general, they were constructed to reflect an actual operational environment, and to allow as fair as possible a separation among the diverse query construction approaches.

There were guidelines for constructing and manipulating the system data structures. These structures were defined to consist of the original documents, any new structures built automatically from the documents (such as inverted files, thesauri, conceptual networks, etc.), and any new structures built manually from the documents (such as thesauri, synonym lists, knowledge bases, rules, etc.). The following guidelines were developed for the TREC task.

1. System data structures should be built using the initial training set (documents from disks 1 and 2, training topics 1-100, and the relevance judgments). They may be modified based on the test documents from disk 3, but not based on the test topics.

2. There are parts of the test collection, such as the Wall Street Journal and the Ziff material, that contain manually assigned controlled or uncontrolled index terms. These fields are delimited by SGML tags, as specified in the documentation files included with the data. Since the primary focus is on retrieval and routing of naturally occurring text, these manually indexed terms should not be used.

3. Special care should be used in handling the routing task. In a true routing situation, a single document would be indexed and compared against the routing topics. Since the test documents are generally indexed as a complete set, routing should be simulated by not using any information based on the full set of test documents (such as weighting based on the test collection, total frequency based on the test collection, etc.) in the searching. It is permissible to use training-set collection information however.

Additionally there were guidelines for constructing the queries from the provided topics. These guidelines were considered of great importance for fair system comparison and were therefore carefully constructed. Three generic categories were defined, based on the amount and kind of manual intervention used.

1. AUTOMATIC (completely automatic initial query construction)

   adhoc queries -- The system will automatically extract information from the topic to construct the query. The query will then be submitted to the system (with no manual modifications) and the results from the system will be the results submitted to NIST. There should be no manual intervention that would affect the results.

   routing queries -- The queries should be constructed automatically using the training topics, the training relevance judgments and the training documents. The queries should then be submitted to NIST before the test documents are released and should not be modified after that point. The unmodified queries should be run against the test documents and the results submitted to NIST.

2. MANUAL (manual initial query construction)

   adhoc queries -- The query is constructed in some manner from the topic, either manually or using machine assistance. Once the query has been constructed, it will be submitted to the system (with no manual intervention), and the results from the system will be the results submitted to NIST. There should be no manual intervention after initial query construction that would affect the results. (Manual intervention is covered by the category labelled FEEDBACK.)

   routing queries -- The queries should be constructed in the same manner as the adhoc queries for MANUAL, but using the training topics, relevance judgments, and training documents. They should then be submitted to NIST before the test documents are released and should not be modified after that point. The unmodified queries should be run against the test documents and the results submitted to NIST.

3. FEEDBACK (automatic or manual query construction with feedback)

   adhoc queries -- The initial query can be constructed using either AUTOMATIC or MANUAL methods. The query is submitted to the system, and a subset of the retrieved documents is used for manual feedback, i.e., a human makes judgments about the relevance of the documents in this subset. These judgments may be communicated to the system, which may automatically modify the query, or the human may simply choose to modify the query himself. At some point, feedback should end, and the query should be accepted as final. Systems that submit runs using this method must submit several different sets of results to allow tracking of the time/cost benefit of doing relevance feedback.

   routing queries -- FEEDBACK cannot be used for routing queries as routing systems have not supported feedback.

2.2 The Participants

There were 31 participating systems in TREC-2, using a wide range of retrieval techniques. The participants were able to choose from three levels of participation: Category A, full participation, Category B, full participation using a reduced dataset (1/4 of the full document set), and Category C for evaluation only (to allow commercial systems to protect proprietary algorithms). The program committee selected only 20 category A and B groups to present talks because of limited conference time, and requested that the rest of the groups present posters. All groups were asked to submit papers for the proceedings.

Each 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 (AUTOMATIC, MANUAL, or FEEDBACK), or by using different parameter settings for one query creation method. Groups could choose to do the routing task, the adhoc task, or both, and were requested to submit the top 1000 documents retrieved for each topic for evaluation.

3. The Test Collection

3.1 Introduction

The creation of the test collection (called the TIPSTER collection) was critical to the success of TREC. Like most traditional retrieval collections, there are three distinct parts to this collection -- the documents, the queries or topics, and the relevance judgments or "right answers". These test collection components are discussed briefly in the rest of this section. For a more complete description of the collection, see Harman [1994].

3.2 The Documents

The documents needed to mirror the different types of documents used in the theoretical TREC application. Specifically they had to have a varied length, a varied writing style, a varied level of editing and a varied vocabulary. As a final requirement, the documents had to cover different timeframes to show the effects of document date on the routing task.

The documents were distributed as CD-ROMs with about 1 gigabyte of data each, compressed to fit. The following shows the actual contents of each disk.

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 (Ziff-Davis Publishing)
- FR -- *Federal Register* (1988)

Disk 3

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

The documents are uniformly formatted into an SGML-like structure, as can be seen in the following example.

<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 implications for computer and communications equipment markets.
AT&T said it is the first national long-distance carrier to announce prices for specific services under a world-wide standardization plan to upgrade phone networks. By announcing commercial services under the plan, which the industry calls the Integrated Services Digital Network, AT&T will influence evolving communications standards to its advantage, consultants said, just as International Business Machines Corp. has created de facto computer standards favoring its products.
.
.
.
</TEXT>
</DOC>

All documents have beginning and end markers, and a unique DOCNO id field. Additionally other fields taken from the initial data appear, but these vary widely across the different sources. The documents have differing amounts of errors, which were not checked or corrected. Not only would this have been an impossible task, but the errors in the data provide a better simulation of the TREC task. Errors in missing document separators or bad document numbers were screened out, although a few were missed and later reported as errors.

Table 1 shows some basic document collection statistics. Note that although the collection sizes are roughly equivalent in megabytes, there is a range of document lengths from very short documents (DOE) to very long (FR). Also the range of document lengths within a collection varies. For example, the documents from AP are similar in length (the median and the average length are very close), but the WSJ and ZIFF documents have a wider range of lengths. The documents from the Federal Register (FR) have a very wide range of lengths.

3.3 The Topics

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

The topics were designed to mimic a real user's need, and were written by people who are actual users of a retrieval system. Although the subject domain of the topics was diverse, some consideration was given to the documents to be searched. The topics were constructed by doing trial retrievals against a sample of the document set, and then those topics that had roughly 25 to 100 hits in that sample were used. This created a range of broader and narrower topics.

The following is one of the topics used in TREC.

<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.

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)	295	266	251	258	190
(disk 2)	255	248	188	211	
(disk 3)	315	248	358	251	
Number of records					
(disk 1)	98,736	84,930	75,180	26,207	226,087
(disk 2)	74,520	79,923	56,920	20,108	
(disk 3)	90,257	78,325	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	

Table 1: Document Statistics

<smry> Summary:
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, a one-sentence description, and a summary sentence or two that can be used as a surrogate for the full topic (often very similar to the 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 assorted concepts related to the topic. This section is designed to provide a mini-knowledge base 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-2 topics: a time factor (current, before a given date, etc.) and a nationality factor (either involving only certain countries or excluding certain countries).

While the TREC topics did not present a problem in scaling, the challenge of either automatically constructing a query, or manually constructing a query with little foreknowledge of its searching capability, was a major challenge for TREC participants. In addition to filtering the relatively large amount of information provided in the topics into queries, the sometimes narrow definition of relevance as stated in the narrative was difficult for most systems to handle.

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. For the TREC task, three possible methods for finding the relevant documents could have been used. In the first method, full relevance judgments could have been made on over one million documents, for each topic, resulting in over 100 million judgments. This was clearly impossible. As a second approach, a random sample of the documents could have been taken, with relevance judgments done on that sample only. The problem with this approach is that a random sample that is large enough to find on the order of 200 relevant documents per topic is a very large random sample, and is likely to result in insufficient relevance judgments. The third method, the one used in TREC, was to make relevance judgments on the sample of documents selected by the various participating systems. This method is known as the pooling method, and has been used successfully in creating other collections [Sparck Jones & van Rijsbergen 1975]. The sample was constructed by taking the top 100 documents retrieved by each system for a given topic and merging them into a pool for relevance assessment. This is a valid sampling method since all the systems used ranked retrieval methods, with those documents most likely to be relevant returned first.

Pooling proved to be an effective method. There was little overlap among the 31 systems in their retrieved documents, although considerably more overlap than in TREC-1.

	TREC-2		TREC-1	
	Max	Actual	Max	Actual
Unique Documents Per Topic (Adhoc, 40 runs 23 groups)	4000	1106.0	3300	1278.86
Unique Documents Per Topic (Routing, 40 runs 24 groups)	4000	1465.6	2200	1066.86

Table 2: Overlap of Submitted Results

Table 2 shows the overlap statistics. The first overlap statistics are for the adhoc topics (test topics against training documents disks 1 and 2), and the second statistics are for the routing topics (training topics against test documents disk 3 only). For example, out of a maximum of 4000 possible unique documents (40 runs times 100 documents), over one-fourth of the documents were actually unique. This means that the different systems were finding different documents as likely relevant documents for a topic. Whereas this might be expected (and indeed has been shown to occur, Katzer et. al. 1982) from widely differing systems, these overlaps were often between two runs for the same system. One reason for the lack of overlap is the very large number of documents that contain many of the same terms as the relevant documents, but the major reason is the very different sets of terms in the constructed queries. This lack of overlap should improve the coverage of the relevance set, and verifies the use of the pooling methodology to produce the sample.

The merged list of results was then shown to the human assessors. Each topic was judged by a single assessor to insure the best consistency of judgment. Varying numbers of documents were judged relevant to the topics. For the TREC-2 adhoc topics (topics 101-150), the median number of relevant documents per topic is 201, down from 277 for topics 51-100 (as used for adhoc topics in TREC-1). Only 11 topics have more than 300 relevant documents, with only 2 topics having more than 500 relevant documents. These topics were deliberately made narrower than topics 51-100 because of a concern that topics with more than 300 relevant documents are likely to have incomplete relevance assessments.

4. Evaluation

An important element of TREC was to provide a common evaluation forum. Standard recall/precision and recall/fallout figures were calculated for each TREC system and these are presented in Appendix A. A chart with additional data about each system is shown in Appendix B. This chart consolidates information provided by the systems that describe features and system timing, and allows some primitive comparison of the amount of effort needed to produce the results.

4.1 Definition of Recall/Precision and Recall/Fallout Curves

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 [Salton & McGill 1983] 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. Note that 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.

A second set of curves was calculated using the recall/fallout measures, where recall is defined as before and fallout is defined as

$$fallout = \frac{number\ of\ nonrelevant\ items\ retrieved}{total\ number\ of\ nonrelevant\ items\ in\ collection}$$

Note that recall has the same definition as the probability of detection and that fallout has the same definition as the probability of false alarm, so that the recall/fallout curves are also the ROC (Relative Operating Characteristic) curves used in signal processing. A sample set of curves corresponding to the recall/precision curves is shown in figure 3. These curves show the same order of performance as do the recall/precision curves and are provided as an alternative method of viewing the results. The present version of the curves is experimental as the curve creation is particularly sensitive to scaling (what range is used for calculating fallout). The high precision section of the curves does not show well in figure 3; the high recall area dominates the curves.

Whereas the recall/precision curves show the retrieval system results as they might be seen by a user (since precision measures the accuracy of each retrieved document as it is retrieved), the recall/fallout curves emphasize the ability of these systems to screen out non-relevant material. In particular the fallout measure shows the discrimation powers of these systems on a large document collection. For example, system A has a fallout of 0.02 at a recall of about 0.48; this means that this system has found almost 50% of the relevant documents, while only retrieving 2% of the non-relevant documents.

4.2 Single-Value Evaluation Measures

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

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.

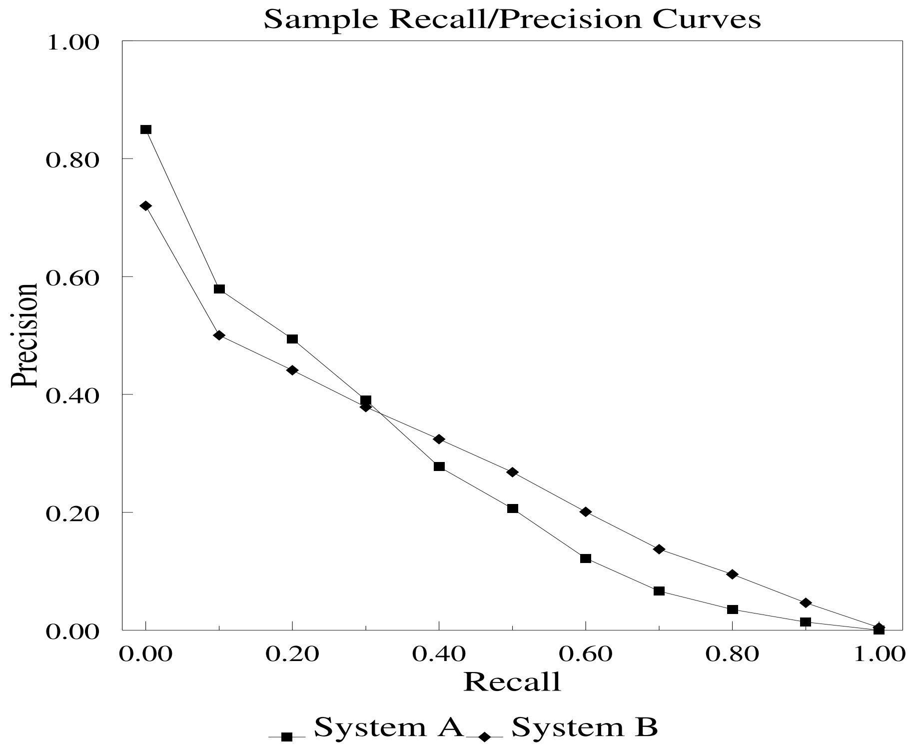
4.3 Problems with Evaluation

Since this was the first time that such a large collection of text has been used in open system evaluation, there were some problems with the existing methods of evaluation. The major problem concerned a thresholding effect caused by the inability to evaluate ALL documents retrieved by a given system.

For TREC-1 the groups were asked to send in only the top 200 documents retrieved by their systems. This artificial document cutoff is relatively low and systems did not retrieve all the relevant documents for most topics within the cutoff. All documents retrieved beyond the 200 mark were considered nonrelevant by default and therefore the recall/precision curves became inaccurate after about 40% recall on average. TREC-2 used the top 1000 documents for evaluation. Figure 4 shows the difference in the curves produced by various evaluation thresholds, including a curve for no threshold (similar to the way evaluation has been done on the smaller collections.). These curves show that the use of a 1000-document cutoff has solved most of the thresholding problem.

Two more issues in evaluation have become important. The first issue involves the need for more statistical evaluation. As will be seen in the results, the recall/precision curves are often close, and there is a need to check if there is truly any statistically significant differences between two systems' results or two sets of results from the same system. This problem is currently under investigation in collaboration with statistical groups experienced in the evaluation of information retrieval systems.

Another issue involves getting beyond the averages to better understand system performance. Because of the huge number of documents and the long topics, it is very



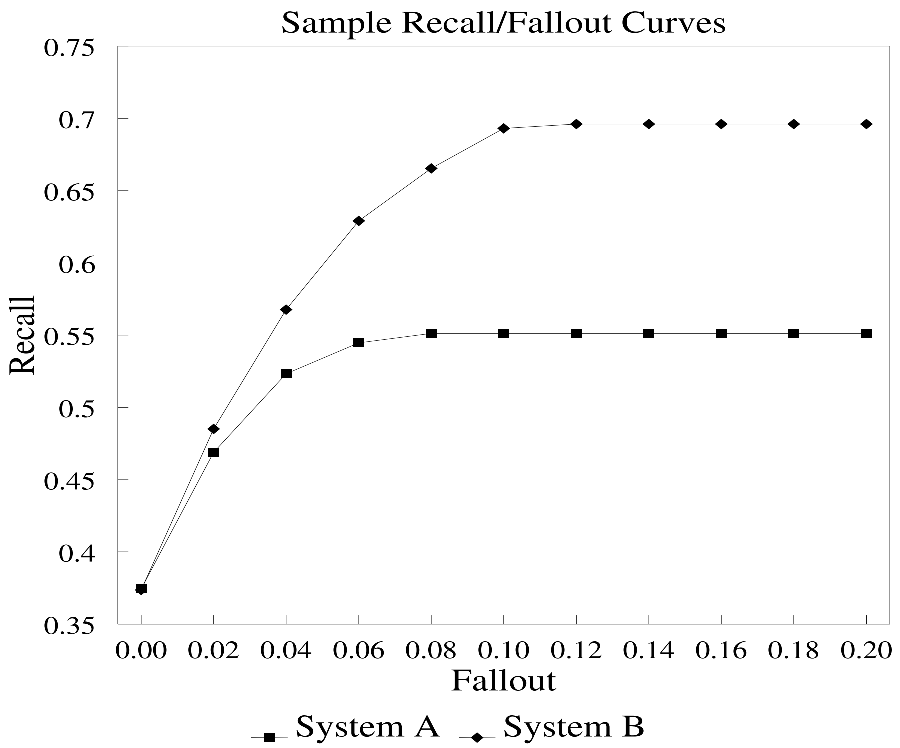


Figure 2 -- A Sample Recall/Precision Curve
Figure 3 -- A Sample Recall/Fallout Curve

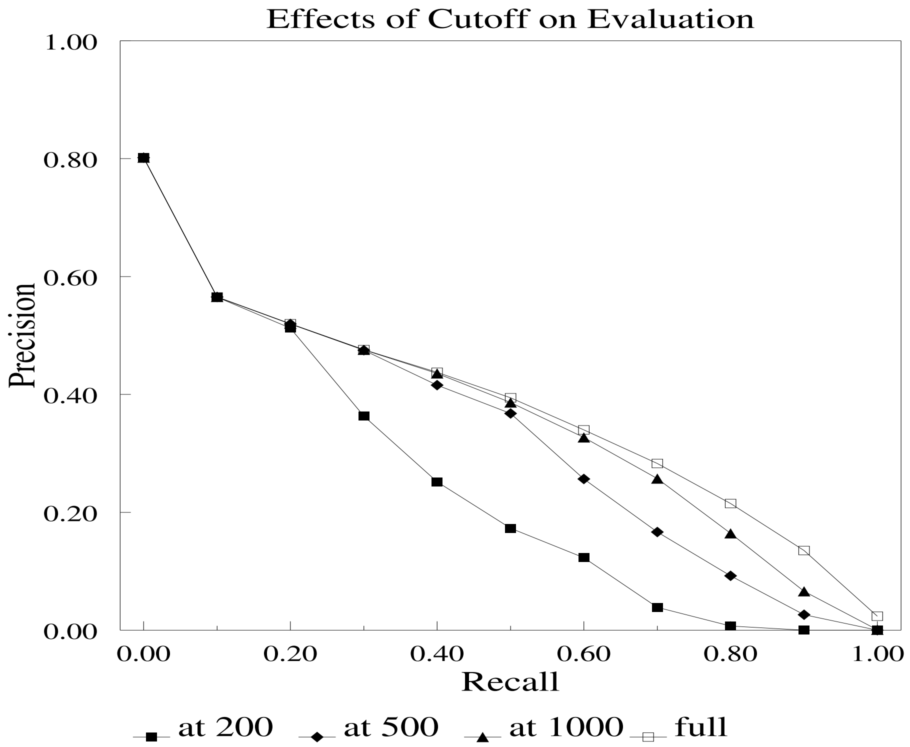


Figure 4: Effect of evaluation cutoffs on recall/precision curves

difficult to perform failure analysis on the results to better understand the retrieval processes being tested. Without better understanding of underlying system performance, it will be hard to consolidate research progress. Some preliminary analysis of per topic performance is provided in section 6, and and more attention will be given to this problem in the future.

5. Results

5.1 Introduction

In general the TREC-2 results showed significant improvements over the TREC-1 results. Many of the original TREC-1 groups were able to "complete" their system rebuilding and tuning tasks. The results for TREC-2 therefore can be viewed as the "best first-pass" that most groups can accomplish on this large amount of data. The adhoc results in particular represent baseline results from the scaling-up of current algorithms to large test collections. The better systems produced similar results, results that are comparable to those seen using these algorithms on smaller test collections.

The routing results showed even more improvement over TREC-1 routing results. Some of this improvement was due to the availability of large numbers of accurate relevance judgments for training (unlike TREC-1), but most of the improvements came from new research by participating groups into better ways of using the training data.

For full descriptions of each system discussed in this section, see the individual papers in this proceedings.

5.2 Adhoc Results

The adhoc evaluation used new topics (101-150) against the two disks of training documents (disks 1 and 2). There were 44 sets of results for adhoc evaluation in TREC-2, with 32 of them based on runs for the full data set. Of these, 23 used automatic construction of queries, 9 used manual construction, and 2 used feedback.

Figure 5 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using automatic construction of queries. The results marked "INQ001" are the INQUERY system from the University of Massachusetts (see Croft, Callan & Broglio paper). This system uses probabilistic term weighting and a probabilistic inference net to combine various topic and document features. The results marked "dortQ2", "Brkly3" and "crnlL2" are all based on the use of the Cornell SMART system, but with important variations. The "crnlL2" run is the basic SMART system from Cornell University (see Buckley, Allan & Salton paper), but using less than optimal term weightings (by mistake). The "dortQ2" results from the University of Dortmund come from using polynomial regression on the training data to find weights for various pre-set term features (see Fuhr, Pfeifer, Bremkamp, Pollmann & Buckley paper). The "Brkly3" results from the University of California at Berkeley come from performing logistic regression analysis to learn optimal weighting for various term frequency measures (see Cooper, Chen & Gey paper). The "CLARTA" system from the CLARIT Corporation expands each topic with noun phrases found in a thesaurus that is automatically generated for each topic (see Evans & Lefferts paper). The "lsiasm" results are from Bellcore (see Dumais paper). This group uses latent semantic indexing to create much larger vectors than the more traditional vector-space models such as SMART. The run marked "lsiasm" represents only the base SMART pre-processing results, however. Due to processing errors the "improved" LSI run produced unexpectedly poor results.

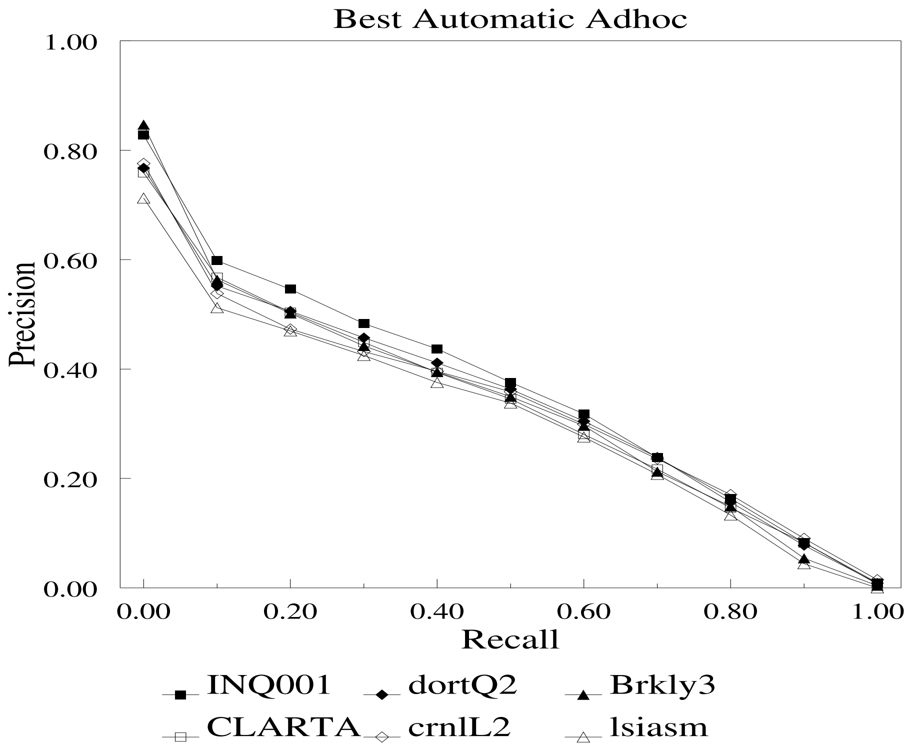
Figure 6 shows the recall/precision curve for the six TREC-2 groups with the highest non-interpolated average precision using manual construction of queries. It should be noted that varying amounts of manual intervention were used. The results marked "INQ002", "siems2", and "CLARTM" are automatically-generated queries with manual modifications. The "INQ002" results reflect various manual modifications made to the "INQ001" queries, with those modifications guided by strict rules. The "siems2" results from Siemens Corporate Research, Inc. (see Voorhees paper) are based on the use of the Cornell SMART system, but with the topics manually modified (the "not" phases removed). These results were meant to be the base run for improvements using WordNet, but the improvements did not materialize. The "CLARTM" results represent manual weighting of the query terms, as opposed to the automatic weighting of the terms that was used in "CLARTA". The results marked "Vtcms2", "CnQst2", and "TOPIC2" are produced from queries constructed completely manually. The "Vtcms2" results are from Virginia Tech (see Fox & Shaw paper) and show the effects of combining the results from SMART vector-space queries with the results from manually-constructed soft Boolean P-Norm type queries. The "CnQst2" results, from ConQuest Software (see Nelson paper), use a very large general-purpose semantic net to aid in constructing better queries from the topics, along with sophisticated morphological analysis of the topics. The results marked "TOPIC2" are from the TOPIC system by Verity Corp. (see Lehman paper) and reflect the use of an expert system working off specially-constructed knowledge bases to improve performance.

Several comments can be made with respect to these adhoc results. First, the better results (most of the automatic results and the three top manual results) are very similar and it is unlikely that there is any statistical differences between them. There is clearly no "best" method, and the fact that these systems have very different approaches to retrieval, including different term weighting schemes, different query construction methods, and different similarity match methods implies that there is much more to be learned about effective retrieval techniques. As will be seen in section 6, whereas the averages for the systems may be similar, the systems do better on different topics and retrieve different subsets of the relevant documents.

A second point that should be made is that the automatic query construction methods continue to perform as well as the manual construction methods. Two groups (the INQUERY system and the CLARIT system) did explicit comparision of manually-modified queries vs those that were not modified and concluded that manual modification provided no benefits. The three sets of results based on completely manually-generated queries had even poorer performance than the manually-modified queries. Note that this result is specific to the very rich TREC topics; it is not clear that this will hold for the short topics normally seen in other retrieval environments.

As a final point, it should be noted that these adhoc results represent significant improvements over the results from TREC-1. Figure 7 shows a comparison of results for a typical system in TREC-1 and TREC-2. Some of this improvement is due to improved evaluation, but the difference between the curve marked "TREC-1" and the curve marked "TREC-2 looking at top 200 only" shows significant performance improvement. Whereas this improvement could represent a difference in topics (the TREC-1 curve is for topics 51-100 and the TREC-2 curves are for topics 101-150), the TREC-2 topics are generally felt to be more difficult and therefore this improvement is likely to be an understatement of the actual improvements.

Only two groups worked with less than the full document collection. Figure 9 shows the results for the one group with official TREC-2 category B results (the results from UCLA were received after the deadline). This figure shows the best results from New York University (see Strzalkowski & Carballo paper), compared with a category B version of the Cornell SMART results. The "nyuir3" results reflect a very intensive use of natural language processing (NLP) techniques, including a parse of the documents to help locate syntactic phrases, context-sensitive expansion of the queries, and other NLP improvements on statistical techniques.



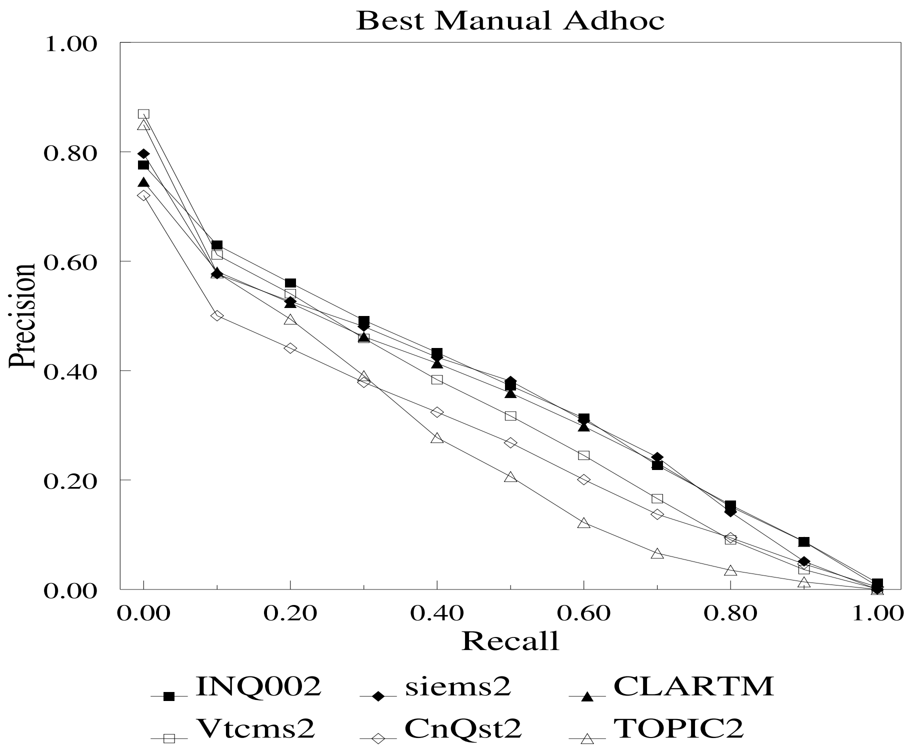
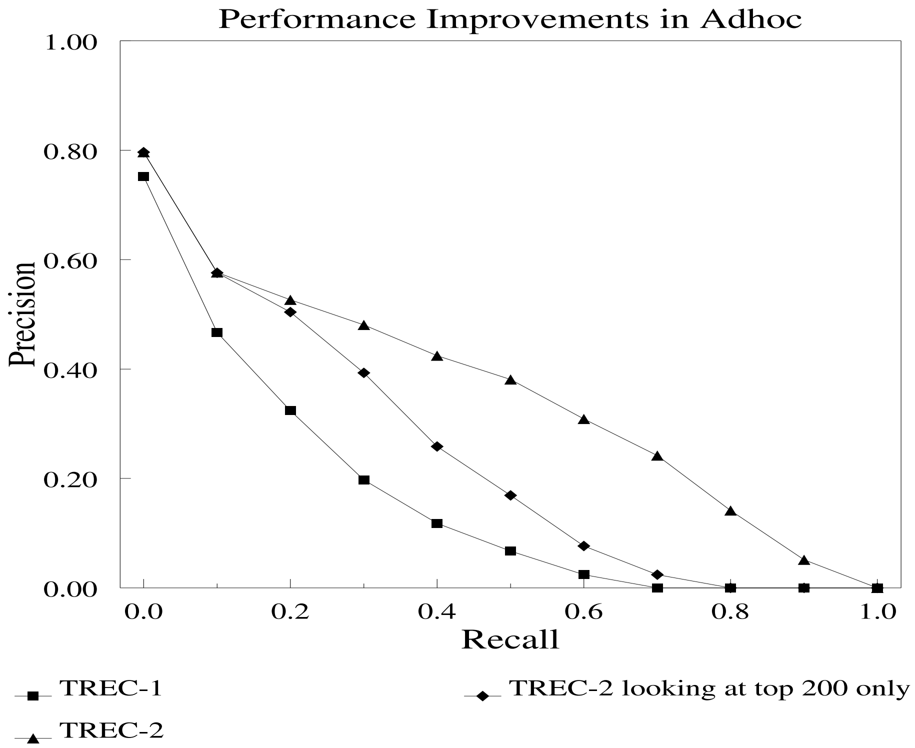


Figure 5 -- Best Automatic Adhoc Results
Figure 6 -- Best Manual Adhoc Results



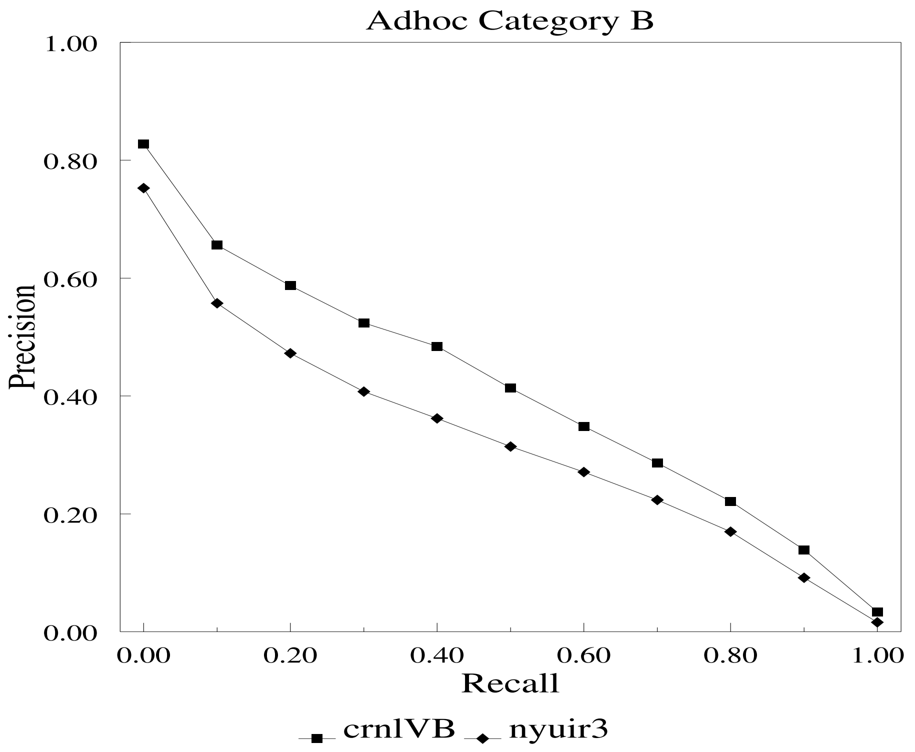


Figure 7 -- Typical Improvements in Adhoc Results
Figure 8 -- Category B Adhoc Results

5.3 Routing Results

The routing evaluation used a subset of the training topics (topics 51-100 were used) against the new disk of test documents (disk 3). There were 40 sets of results for routing evaluation, with 32 of them based on runs for the full data set. Of the 32 systems using the full data set, 23 used automatic construction of queries, and 9 used manual construction.

Figure 9 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using automatic construction of the routing queries. Again three systems are based on the Cornell SMART system. The plot marked "crnlC1" is the actual SMART system, using the basic Rocchio relevance feedback algorithms, and adding many terms (up to 500) from the relevant training documents to the terms in the topic. The "dortP1" results come from using a probabilistically-based relevance feedback instead of the vector-space algorithm, and adding only 20 terms from the relevant documents to each query. These two systems have the best routing results. The "Brkly5" system uses logistic regression on both the general frequency variables used in their adhoc approach and on the query-specific relevance data available for training with the routing topics. The results marked "cityr2" are from City University, London (see Robertson, Walker, Jones, Hancock-Beaulieu & Gafford paper). This group automatically selected variable numbers of terms (10-25) from the training documents for each topic (the topics themselves were not used as term sources), and then used traditional probabilistic reweighting to weight these terms. The "INQ003" results also use probabilistic reweighting, but use the topic terms, expanded by 30 new terms per topic from the training documents. The results marked "lsir2" are more latent semantic indexing results from Bellcore. This run was made by creating a filter of the singular-value decomposition vector sum or centroid of all relevant documents for a topic (and ignoring the topic itself).

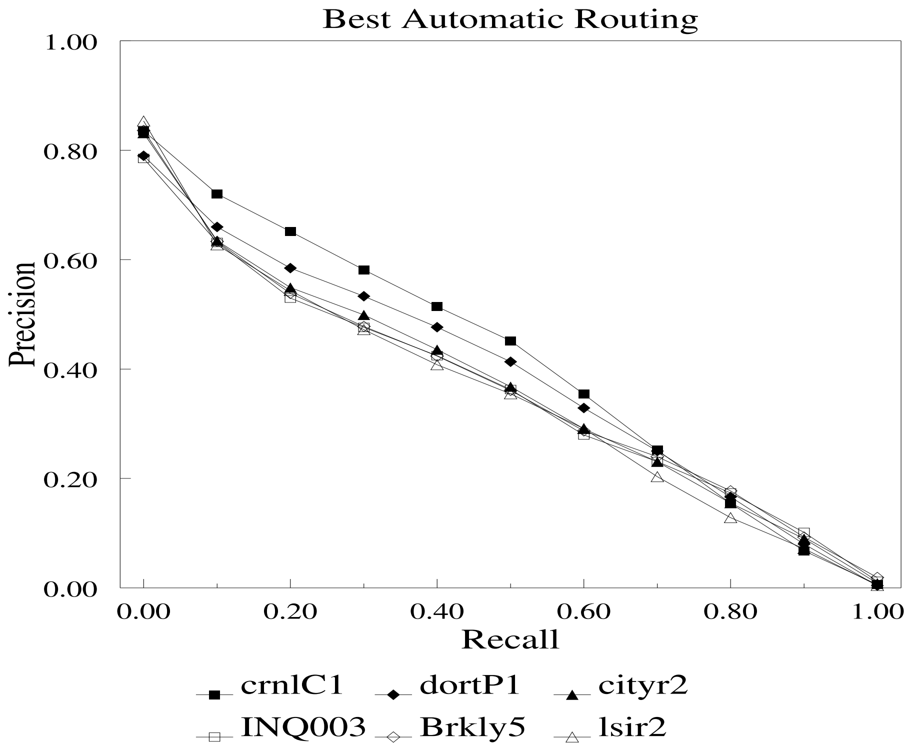
Figure 10 shows the recall/precision curves for the six TREC-2 groups with the highest non-interpolated average precision using manual construction of the routing queries. The results marked "INQ004" are from the INQRY system using an inferential combination of the "INQ003" queries and manually modified queries created from the topic. The "trw2" results represent an adaptation of the TRW Fast Data Finder pattern matching system to allow use of term weighting (see Mettler paper). The queries were manually constructed and the term weighting was learned from the training data. The "gecrd1" results from GE Research and Development Center (see Jacobs paper) also come from manually-constructed queries, but using a general-purpose lexicon and the training data to suggest input to the Boolean pattern matcher. The results marked "CLARTM" are similar to the "CLARTM" adhoc results except that the training documents were used as the source for thesaurus building, as opposed to using the top set of retrieved documents. The "rutcombx" results from Rutgers University (see Belkin, Kantor, Cool & Quatrain paper) come from combining 5 sets of manually-generated Boolean queries to optimize performance for each topic. The results marked "TOPIC2" are from the TOPIC system and reflect the use of an expert system working off specially-constructed knowledge bases to improve performance.

As was the case with the adhoc topics, the automatic query construction methods continue to perform as well as, or in this case, better than the manual construction methods. A comparision of the two INQRY runs illustrates this point and shows that all six results with manually-generated queries perform worse than the six runs with automatically-generated queries. The availability of the training data allows an automatic tuning of the queries that would be difficult to duplicate manually without extensive analysis.

Unlike the adhoc results, there are two runs ("crnlC1" and "dortP1") that are clearly better than the others, with a significant difference between the "crnlC1" results and the "dortP1" results and also significant differences between these results and the rest of the automatically-generated query results. In particular the Cornell group's ability to effectively use many terms (up to 500) for query expansion was one of the most interesting findings in TREC-2 and represents a departure from past results (see Buckley, Allan, & Salton paper for more on this).

As a final point, it should be noted that the routing results also represent significant improvements over the results from TREC-1. Figure 11 shows a comparison of results for a typical system in TREC-1 and TREC-2. Some of this improvement is due to the improved evaluation techniques, but the difference between the curve marked "TREC-1" and the curve marked "TREC-2 looking at top 200 only" shows significant performance improvement. There is even more improvement for the routing results than for the adhoc results, due to better training data (mostly non-existent for TREC-1) and to major efforts by many groups in new routing algorithm experiments.

Only four groups worked with less than the full document collection. Figure 12 shows the results for two of the groups in category B compared with a category B version of the Cornell SMART results. These curves show the results of runs from New York University (that were done in a similar method as that used for the adhoc results) and results from Dalhousie University.



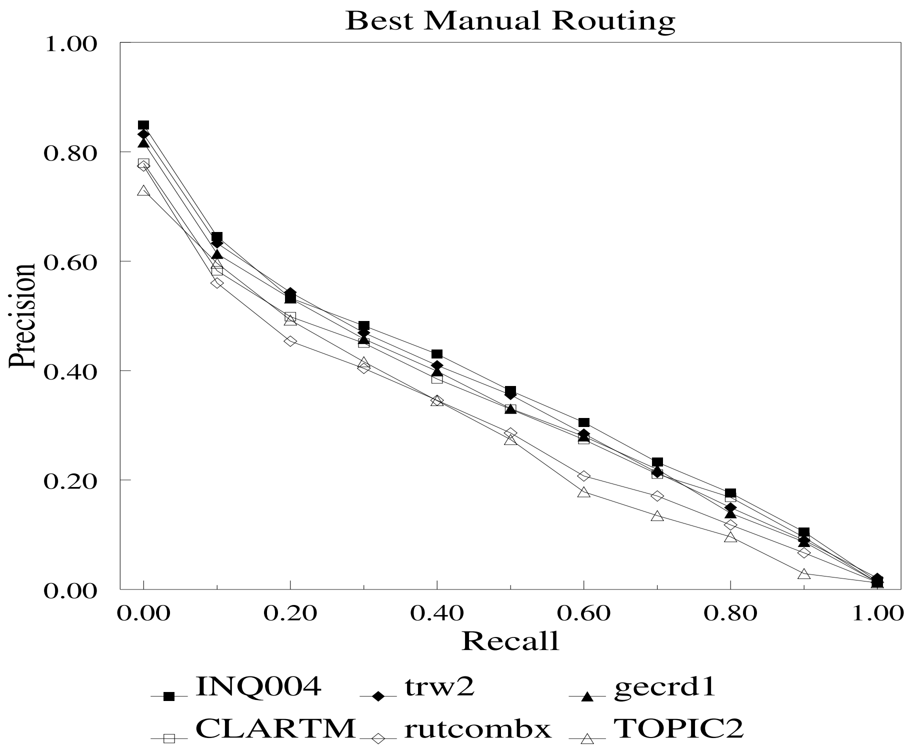
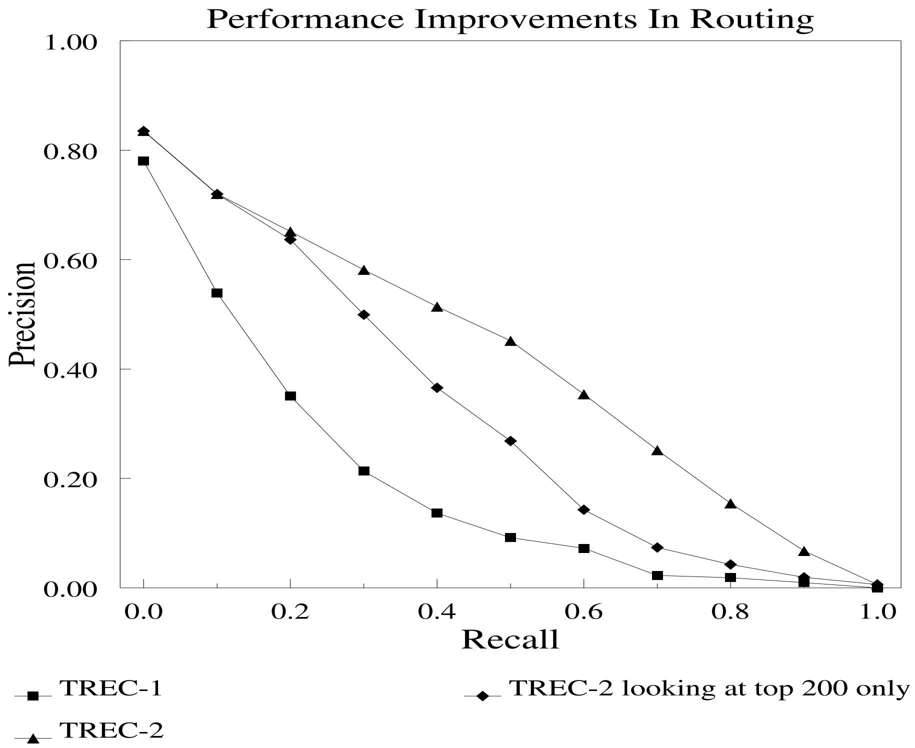


Figure 9 -- Best Automatic Routing Results
Figure 10 -- Best Manual Routing Results



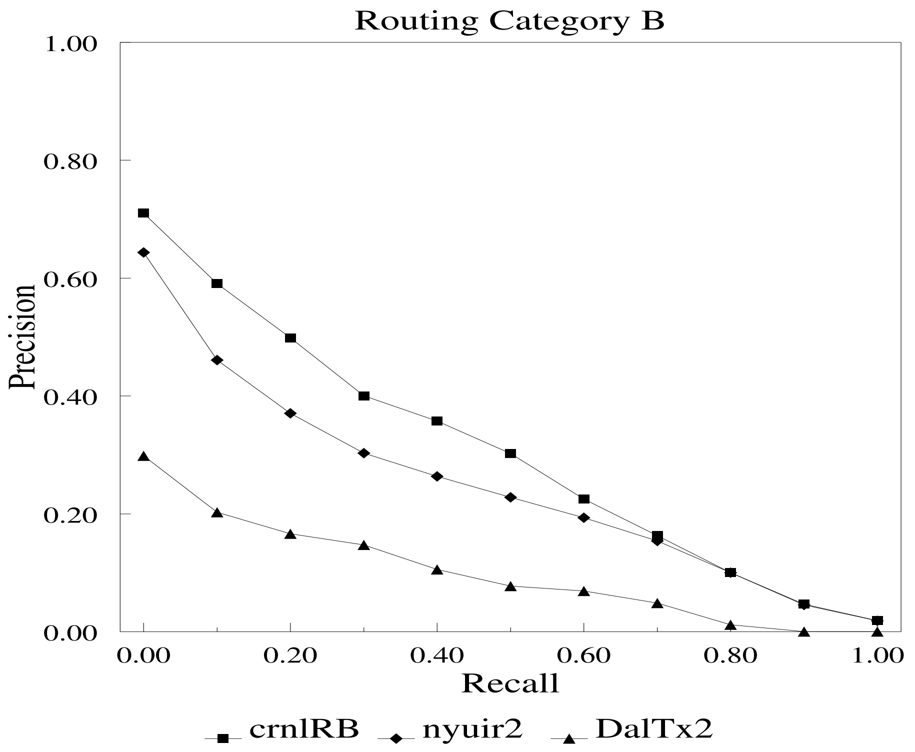


Figure 11 -- Typical Improvements in Routing Results
Figure 12 -- Category B Routing Results

6. Some Preliminary Analysis

6.1 Introduction

The recall/precision curves shown in section 5 represent the average performance of the various systems on the full sets of topics. It is important to look beyond these averages in order to learn more about how a given system is performing and to discover some generalizable principles of retrieval.

Individual systems are able to do this by performing failure analysis (see Dumais paper in this proceedings for a good example) and by running specific experiments to test hypotheses on retrieval behavior within a given system. However, additional information can be gained by doing some cross-system comparison: information about specific system behavior and information about generalized information retrieval principles. One way to do this is to examine system behavior with respect to test collection characteristics. A second method is to compare system behavior on a topic by topic basis.

6.2 The Effects of Test Collection Characteristics

One particular test collection characteristic is the length of documents, both the average length of documents in a collection, and the variation in document length across a collection. Document length has significant effect on system performance. A term that appears 10 times in a "short" document is likely to be more important to that document than if the same term appeared 10 times in a "long" document. Table 3 shows system performance across the different document subcollections for each of the adhoc topics, listing the total number of documents that were retrieved by the system as well as the number of relevant documents that were retrieved.

Two particular points can be seen from table 3. First, the better systems retrieve about 50% relevant documents from all the subcollections except the *Federal Register* (FR). For this subcollection the retrieval rates are in the 25% range because the varied length of these documents makes retrieval difficult.

The second point concerning table 3 is that the retrieval rate across the subcollections is highly varied among the systems. For example the "Brkly3" results show that many fewer *Federal Register* documents and more *AP* were retrieved than for the INQUERY system, whereas the "CLARTA" results show more *DOE* abstracts and fewer *Wall Street Journal* being retrieved. These "biases" towards particular subcollections reflect the methods used by systems such as the length normalization issues, domain concentrations of terminology, and methods used to "merge" results across subcollections (often implicit merges during indexing).

A second test collection characteristic worth examining is the varied broadness and varied difficulty of the topics. An analysis was done [Harman 1994] to find the topics for which the systems retrieved the lowest percentage of the relevant documents on average. These topics are 61, 67, 76, 77, 81, 85, 90, 91, 93, and 98 for the routing topics and 101, 114, 120, 121, 124, 131, 139, 140, 141, and 149 for the adhoc topics. Tables 4 and 5 show the top 8 system runs for the individual topics based in the average precision (non-interpolated). These tables mix automatic, manual, and feedback results for category A, and also category B results, so they should be interpreted carefully. However they do demonstrate that no consistent patterns appear for the "hard" topics. The two best routing runs ("crnlC1" and "dortP1") only do well on about half of these topics, and the adhoc results are even more varied. Often systems that do not perform well on average are the top performing system for a given topic. This verifies that, as usual, the variation across the topics is greater than the variation across systems.

6.3 Cross-System Analysis

Tables 4 and 5 not only show the wide variation in system performance, but also raise several questions about system performance in general.

1. Does better average performance for a system result from better performance on most topics or from comparable performance on most topics and significantly better performance on other topics?
2. If two systems perform similarly on a given topic, does that mean that they have retrieved a large proportion of the same relevant documents?
3. Do systems that use "similar" approaches have a high overlap in the particular relevant documents they retrieve?
4. And, if number 3 is not true, what are the issues that affect high overlap of relevant documents?

Work is ongoing at NIST on these questions and other related issues.

Run Tag	AP	DOE	FR	WSJ	ZIFF
Brkly3	2414/1155	293/155	97/22	1847/831	349/121
citri1	2152/970	474/179	129/16	1865/770	380/177
citri2	2206/1019	348/156	250/67	1814/756	382/179
cityau	1578/794	93/46	1359/147	1603/661	367/108
citymf	1939/1226	24/21	584/146	2026/812	427/173
CLARTA	2034/1048	403/170	412/95	1795/819	356/190
CLARTM	2131/1087	388/208	315/87	1769/820	397/221
CnQst1	2272/921	254/112	311/74	1763/689	400/181
CnQst2	2214/980	191/94	453/93	1787/738	355/184
crnlL2	1914/944	648/240	155/33	1774/773	509/213
crnlV2	2164/1083	687/236	79/22	1600/682	470/194
dortL2	2081/1053	305/169	473/78	1818/815	323/166
dortQ2	2205/1053	357/171	186/44	1924/874	328/171
erima1	1077/501	85/51	1364/110	2251/752	223/94
erima2	1267/614	122/68	1219/125	2124/773	268/133
gecrd2	2250/852	294/91	319/68	1952/743	185/77
HNCad1	2140/1042	409/145	164/53	1875/839	412/181
HNCad2	2163/1286	306/159	171/67	1974/1005	386/237
INQ001	2031/1071	206/107	297/115	2184/1023	282/151
INQ002	2087/1111	201/120	276/111	2141/1010	295/177
lsia1	2278/771	587/124	124/0	1448/376	563/61
lsiasm	2168/1052	711/211	70/17	1607/690	444/183
nyuir1	0/0	0/0	0/0	5000/1360	0/0
nyuir2	0/0	0/0	0/0	5000/1547	0/0
nyuir3	0/0	0/0	0/0	5000/1547	0/0
pircs3	2109/1021	358/152	246/86	1999/835	288/139
pircs4	2108/1014	342/148	254/85	2012/863	284/137
prceo1	1099/1024	315/83	1178/205	1377/980	1031/277
prceo1	1667/1024	695/83	381/205	1350/980	907/277
rutcomb1	1029/368	181/72	112/18	963/312	215/79
rutfmed	945/309	131/46	161/9	963/282	200/77
schau1	2038/901	534/189	173/18	1778/706	477/186
siems2	2225/1147	631/218	62/8	1655/770	427/202
siems3	2238/1173	654/208	53/7	1619/764	436/194
TMC8	2054/859	146/44	763/59	1472/526	565/183
TMC9	1923/802	77/29	975/63	1401/507	624/171
TOPIC2	2292/996	152/98	344/100	1762/889	384/229
UREKA2	385/215	0/0	4003/87	354/144	258/10
UREKA3	755/405	5/2	2654/67	1045/348	441/22
uicah	1612/628	234/104	797/137	1846/356	511/167
VTcms2	2110/1130	232/107	444/95	1859/894	355/169
totals	71354/4630	12073/669	21407/396	79396/3929	15504/1154

Table 3: Number of Documents Retrieved/Relevant by Document Subcollection

51	nyuir2	nyuir1	gecrd1	TOPIC2	ADS2	cityr2	INQ004	INQ003
52	INQ004	INQ003	Brkly4	pircs2	VTcms2	gecrd1	pircs1	trw1
53	gecrd1	nyuir2	trw2	nyuir1	CLARTM	CLARTA	dortP1	INQ003
54	siems1	crnlR1	schau1	Brkly4	INQ003	crnlC1	lsir1	CLARTM
55	dortP1	crnlR1	crnlC1	lsir2	dortV1	CLARTM	cityr1	CLARTA
56	trw2	dortP1	dortV1	INQ003	INQ004	HNCrt1	crnlR1	crnlC1
57	INQ003	lsir2	INQ004	trw2	crnlC1	TMC6	VTcms2	crnlR1
58	nyuir2	nyuir1	rutcombx	INQ003	lsir2	INQ004	gecrd1	Brkly5
59	trw2	Brkly5	gecrd1	lsir1	HNCrt1	VTcms2	HNCrt2	lsir2
60	dortP1	dortV1	rutcombx	crnlR1	INQ004	crnlC1	INQ003	TOPIC2
61	TOPIC2	rutcombx	Brkly4	idsra2	cityr2	lsir1	INQ004	Brkly5
62	crnlR1	crnlC1	dortP1	lsir1	CLARTA	Brkly4	CLARTM	Brkly5
63	dortV1	crnlC1	pircs2	crnlR1	pircs1	siems1	HNCrt1	dortP1
64	nyuir2	lsir2	INQ004	INQ003	Brkly5	crnlC1	crnlR1	cityr2
65	crnlC1	dortV1	dortP1	HNCrt1	crnlR1	trw2	HNCrt2	lsir1
66	pircs2	pircs1	dortP1	dortV1	crnlR1	crnlC1	siems1	INQ004
67	crnlR1	crnlC1	INQ004	nyuir2	dortP1	cityr2	lsir2	INQ003
68	Brkly5	crnlC1	cityr1	cityr2	trw2	INQ003	lsir2	CLARTA
69	erimr1	Brkly5	dortV1	cityr2	cityr1	erimr2	lsir1	Brkly4
70	TMC6	TMC7	rutcombx	VTcms2	HNCrt2	Brkly5	INQ004	cityr2
71	crnlR1	crnlC1	HNCrt2	siems1	CLARTM	HNCrt1	CLARTA	lsir2
72	crnlR1	crnlC1	dortP1	siems1	INQ003	Brkly5	INQ004	cityr1
73	INQ003	crnlR1	cityr2	INQ004	crnlC1	trw2	dortP1	dortV1
74	crnlR1	rutcombx	crnlC1	CLARTA	Brkly5	dortP1	siems1	dortV1
75	crnlC1	ADS2	crnlR1	trw1	lsir2	dortP1	cityr2	nyuir2
76	trw2	cityr2	TOPIC2	TMC6	TMC7	crnlC1	crnlR1	INQ003
77	crnlR1	crnlC1	INQ003	CLARTM	dortV1	dortP1	INQ004	CLARTA
78	rutcombx	TOPIC2	INQ004	CLARTM	INQ003	dortV1	pircs2	CLARTA
79	cityr2	crnlR1	crnlC1	INQ004	dortP1	gecrd1	lsir2	INQ003
80	trw1	crnlC1	crnlR1	cityr1	Brkly5	INQ003	INQ004	cityr2
81	gecrd1	TMC7	TMC6	cityr2	trw2	VTcms2	HNCrt2	cityr1
82	CLARTM	CLARTA	trw2	Brkly5	pircs1	pircs2	dortV1	dortP1
83	TOPIC2	gecrd1	trw1	crnlC1	HNCrt1	crnlR1	cityr2	cityr1
84	dortP1	crnlC1	lsir2	gecrd1	crnlR1	dortV1	trw1	VTcms2
85	crnlR1	crnlC1	dortP1	Brkly5	trw2	nyuir2	dortV1	siems1
86	gecrd1	VTcms2	lsir2	lsir1	cityr1	crnlR1	cityr2	crnlC1
87	lsir2	gecrd1	cityr1	cityr2	HNCrt1	Brkly5	crnlC1	HNCrt2
88	crnlC1	cityr2	crnlR1	Brkly4	dortP1	lsir2	dortV1	Brkly5
89	trw2	nyuir1	TOPIC2	TMC6	HNCrt1	uicr1	HNCrt2	gecrd1
90	gecrd1	trw1	crnlC1	crnlR1	schau1	VTcms2	Brkly5	dortP1
91	trw1	INQ004	schau1	Brkly5	trw2	TOPIC2	HNCrt2	HNCrt1
92	gecrd1	crnlR1	lsir2	crnlC1	CLARTM	CLARTA	nyuir1	INQ003
93	INQ004	rutcombx	INQ003	TMC6	trw1	Brkly5	TMC7	gecrd1
94	lsir2	crnlC1	cityr2	gecrd1	INQ004	CLARTM	trw2	cityr1
95	VTcms2	gecrd1	crnlC1	Brkly5	crnlR1	Brkly4	trw1	siems1
96	dortP1	TOPIC2	cityr1	dortV1	cityr2	lsir2	crnlC1	rutcombx
97	idsra2	HNCrt1	nyuir2	dortP1	HNCrt2	lsir2	crnlC1	TOPIC2
98	HNCrt1	HNCrt2	crnlC1	trw2	DalTx2	INQ004	crnlR1	dortP1
99	lsir2	crnlR1	dortP1	CLARTA	crnlC1	CLARTM	dortV1	cityr2
100	crnlC1	crnlR1	dortP1	lsir2	dortV1	CLARTA	CLARTM	lsir1

Table 4: System Rankings (using Average Precision) on Individual Topics

101	rutcomb1	VTcms2	crnlV2	INQ002	dortQ2	pircs3	Brkly3	CLARTM
102	crnlL2	crnlV2	VTcms2	siems3	dortL2	INQ002	siems2	CLARTM
103	siems3	siems2	schau1	citri1	crnlV2	lsiasm	HNCad2	HNCad1
104	dortQ2	CLARTM	CLARTA	pircs4	pircs3	dortL2	HNCad2	lsiasm
105	citri2	lsiasm	citri1	siems2	siems3	crnlV2	schau1	crnlL2
106	VTcms2	INQ002	INQ001	TOPIC2	pircs4	pircs3	CLARTM	dortL2
107	CnQst1	CnQst2	rutcomb1	TOPIC2	VTcms2	INQ002	rutfmed	CLARTM
108	citri1	dortQ2	siems3	VTcms2	siems2	HNCad2	schau1	dortL2
109	dortL2	crnlL2	dortQ2	CLARTA	CLARTM	pircs3	crnlV2	pircs4
110	INQ002	INQ001	Brkly3	dortQ2	nyuir3	nyuir2	cityau	siems2
111	CLARTA	CLARTM	INQ001	dortQ2	Brkly3	siems2	siems3	pircs4
112	INQ002	INQ001	VTcms2	nyuir2	nyuir3	HNCad1	HNCad2	CnQst2
113	VTcms2	crnlL2	dortL2	crnlV2	nyuir1	siems2	CLARTM	INQ002
114	INQ002	cityau	VTcms2	INQ001	siems3	siems2	lsia1	TOPIC2
115	nyuir2	nyuir3	nyuir1	siems2	dortL2	crnlV2	siems3	crnlL2
116	VTcms2	CLARTA	HNCad2	HNCad1	siems3	siems2	CLARTM	Brkly3
117	citri2	citri1	dortQ2	INQ001	TMC8	lsiasm	gecrd2	schau1
118	nyuir2	nyuir3	nyuir1	TOPIC2	citymf	dortQ2	CLARTA	INQ001
119	nyuir1	nyuir2	nyuir3	INQ002	INQ001	dortQ2	citymf	VTcms2
120	citymf	nyuir2	nyuir3	nyuir1	CnQst2	CnQst1	VTcms2	erima2
121	TOPIC2	CLARTM	VTcms2	Brkly3	nyuir1	prceo1	INQ002	rutfmed
122	siems2	siems3	INQ002	INQ001	dortQ2	Brkly3	CLARTM	crnlV2
123	nyuir1	nyuir2	nyuir3	CLARTA	INQ001	INQ002	CLARTM	pircs4
124	nyuir2	nyuir3	nyuir1	dortL2	dortQ2	INQ001	Brkly3	TMC9
125	crnlV2	Brkly3	crnlL2	CLARTM	siems3	CLARTA	pircs4	pircs3
126	siems3	crnlL2	siems2	Brkly3	crnlV2	INQ002	CLARTM	INQ001
127	cityau	Brkly3	CLARTA	HNCad2	INQ001	INQ002	siems2	siems3
128	VTcms2	CLARTA	siems3	siems2	CLARTM	TOPIC2	citri1	lsiasm
129	INQ001	INQ002	cityau	CLARTM	siems2	Brkly3	crnlL2	CLARTA
130	INQ002	INQ001	dortQ2	crnlL2	pircs4	CLARTM	dortL2	pircs3
131	TOPIC2	VTcms2	HNCad1	HNCad2	siems3	Brkly3	siems2	INQ002
132	dortL2	INQ001	INQ002	citri1	citri2	dortQ2	HNCad2	crnlL2
133	CnQst2	CnQst1	rutcomb1	pircs4	INQ002	pircs3	cityau	INQ001
134	crnlL2	dortL2	nyuir1	nyuir2	nyuir3	INQ002	INQ001	dortQ2
135	nyuir2	nyuir3	nyuir1	Brkly3	INQ001	INQ002	siems3	siems2
136	VTcms2	CnQst1	CnQst2	CLARTM	pircs4	CLARTA	dortQ2	TOPIC2
137	CLARTA	nyuir2	nyuir3	Brkly3	siems2	siems3	CLARTM	nyuir1
138	nyuir2	nyuir3	rutfmed	rutcomb1	nyuir1	schau1	gecrd2	citri1
139	nyuir2	nyuir3	nyuir1	VTcms2	dortL2	HNCad2	dortQ2	HNCad1
140	nyuir2	nyuir3	nyuir1	dortQ2	dortL2	INQ002	siems3	siems2
141	VTcms2	INQ002	CnQst2	INQ001	Brkly3	dortL2	dortQ2	CnQst1
142	dortQ2	siems2	crnlL2	VTcms2	siems3	CLARTM	crnlV2	Brkly3
143	INQ002	INQ001	siems2	siems3	crnlL2	crnlV2	nyuir2	nyuir3
144	VTcms2	Brkly3	citymf	crnlV2	siems3	lsiasm	siems2	HNCad2
145	crnlL2	crnlV2	dortL2	CLARTM	nyuir1	siems3	siems2	dortQ2
146	Brkly3	siems3	siems2	lsiasm	crnlV2	schau1	CLARTM	citri1
147	HNCad2	HNCad1	VTcms2	citri1	INQ002	INQ001	citymf	CLARTA
148	lsiasm	crnlL2	crnlV2	siems2	siems3	Brkly3	dortL2	dortQ2
149	nyuir1	CnQst2	TOPIC2	CnQst1	CLARTA	rutfmed	Brkly3	rutcomb1
150	crnlL2	dortQ2	CLARTM	siems3	INQ002	INQ001	crnlV2	siems2

Table 5: System Rankings (using Average Precision) on Individual Topics

6.4 Summary

The TREC-2 conference demonstrated a wide range of different approaches to the retrieval of text from large document collections. There was significant improvement in retrieval performance over that seen in TREC-1, especially in the routing task. The availability of large amounts of training data for routing allowed extensive experimentation in the best use of that data, and many different approaches were tried in TREC-2. The automatic construction of queries from the topics continued to do as well as, or better than, manual construction of queries, and this is encouraging for groups supporting the use of simple natural language interfaces for retrieval systems.

How well is the TREC initiative meeting its goals? There is certainly increased research using a much larger collection than had previously been tested. This leads not only to discovering interesting research problems, but also to developing algorithms that are ripe for transfer into commercial systems. The conference itself provided the opportunity for this; there was open exchange between the research groups in universities and the research groups in commercial organizations and this is a very critical part of technology transfer.

There will be a third TREC conference in 1994, and all the systems that participated in TREC-2 will be back, along with additional groups.

7. References

Harman D. (1993) (Ed.).*The First Text REtrieval Conference (TREC-1).* National Institute of Standards and Technology Special Publication 500-207,

Harman D. (1994). Data Preparation. In: Merchant R. (Ed.).*The Proceedings of the TIPSTER Text Program - Phase I.* San Mateo, California: Morgan Kaufmann Publishing Co., 1994.

Katzer J., McGill M.J., Tessier J.A., Frakes W., and Das-Gupta P. (1982). A Study of the Overlap among Document Representations. *Information Technology: Research and Development*, 1(2), 261-274.

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

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. 1993.