DOCUMENT DETECTION OVERVIEW

Donna Harman

National Institute of Standards and Technology
Gaithersburg, MD. 20899

1. INTRODUCTION

The goal of the document detection half of the TIPSTER project was to significantly advance the state of the art in effective document detection from large, real-world document collections. This document detection needed to be used in both the routing environment (static queries against a constant stream of new data) and the adhoc environment (new queries against archival data). An additional requirement was that the algorithms for these tasks be as domain and language independent as possible. To demonstrate language independence, the project was done both in Japanese and English. To demonstrate domain independence, the test collection was selected to cover many different subject areas and different document structures.

The document detection task mirrors the general task known as information retrieval. This area of research has seen over 30 years of experimentation [1], leaving a legacy of proven evaluation methodologies. The most prominent of these methodologies is the use of a test collection. A test collection for information retrieval consists of a set of documents, a set of test queries or questions, and a set of relevance judgments that are considered to be the "right" answers to the questions. The first test collections, such as the Cranfield collection, were built in the early 1960's. The Cranfield collection contains 1400 documents (all abstracts), 225 queries (several sentence natural language statements), and 1827 relevance judgments, or an average of about 6 relevant documents per query. Since the early 1960's several other test collections have been built, but none contain the extremely large numbers of documents necessary to reflect the environment to be modeled in TIPSTER.

The first step of this project, therefore, was to create a very large test collection and to design the test methodology and evaluation measures needed for TIPSTER. The test design was based on traditional information retrieval models, and is detailed in the next section. Evaluation was done using recall, precision and fallout measures. These measures are discussed in the section on evaluation metrics.

The test design and test collection used for TIPSTER was also used for both the TREC conferences [2,3]. The only difference between the evaluation done for the TIPSTER contractors and the TREC participants was in the evaluation schedule and in the number of results submitted for evaluation. The first TREC conference took place 2 months after the 12-month TIPSTER evaluation and the second TREC conference coincided with the 24-month TIPSTER evaluation. The TIPSTER contractors had an additional evaluation at 18 months. TREC participants were limited to submitting only 2 sets of results for adhoc or routing evaluation, whereas the TIPSTER contractors were allowed to submit an unlimited number of runs for evaluation.

2. TEST DESIGN

The test design called for the creation of a set of training data and a set of test data. The training data consisted of large numbers of documents (between 1 and 2 gigabytes of text), 50 training topics, and lists of documents for each of the topics that were known to be relevant (the "right answers"). The test data consisted of 50 new topics and about a gigabyte of new documents.

A slight departure from traditional information retrieval methodology was needed to better handle the TIPSTER environment. All previous test collections have assumed that the test questions or topics are closely related to the actual queries submitted to the retrieval systems, as the test questions are generally transformed automatically into the structure of terms submitted to the retrieval systems as input. This input structure is called the query in the TIPSTER environment, with the test question itself referred to as the topic. Since most previous research has involved simple automatic generation of queries from topics, there was no need for a distinction to be made between topics and queries. In TIPSTER this distinction became important because the topics needed to carry a large amount of highly specific information, and the methods of query construction therefore became more complex.

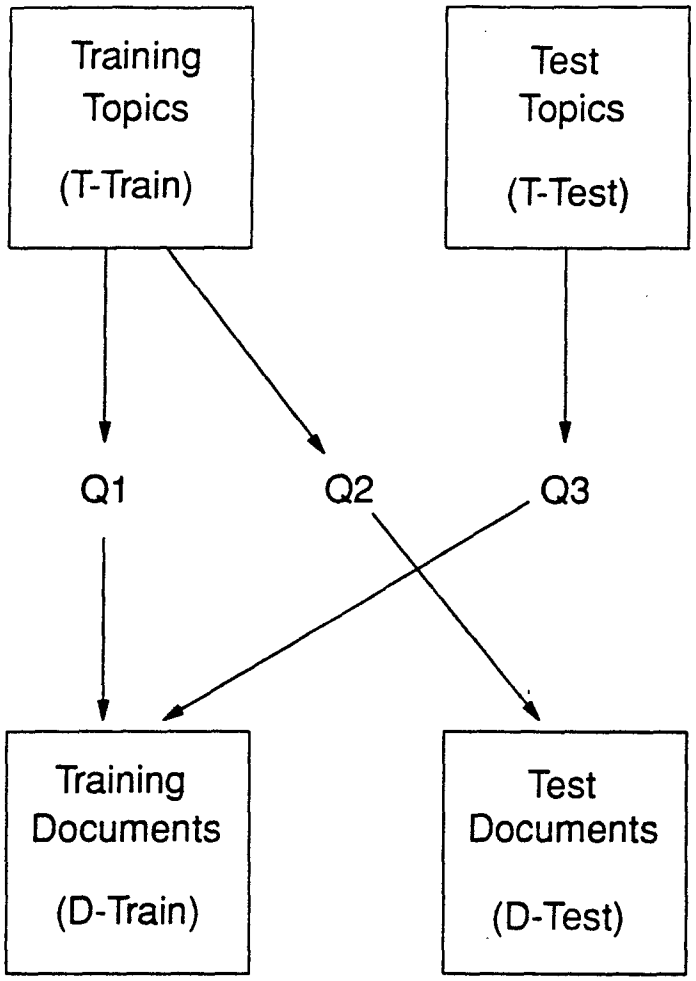


Figure 1 -- The TIPSTER Document Detection Task

Figure 1 shows a schematic of the test design, including the various components of the test methodology. The diagram reflects the four data sets (2 sets of topics and 2 sets of documents) that were provided to contractors. The first set of topics and documents (T-Train and D-Train) were provided to allow system training and to serve as the base for routing and adhoc experiments. The routing task assumes a static set of topics (T-Train), with evaluation of routing done by providing new test documents (D-Test). The adhoc task assumes a static set of documents (D-Train), with evaluation of adhoc retrieval done by providing new topics (T-Test).

Three different sets of queries were generated from the data sets. Q1 is the set of queries (probably multiple sets) created to help in adjusting a retrieval system to this task. The results of this research were used to create Q2, the routing queries to be used against the new test documents (D-Test). Q3 is the set of queries created from the new test topics (T-Test) as adhoc queries for searching against the old documents (D-Train). The results from searches using Q2 and Q3 were the official evaluation results sent to NIST for both TIPSTER and TREC.

The Japanese language test design paralleled exactly the English language test design.

3. EVALUATION SCHEDULE

For the English language document detection task there were three evaluations conducted during the 2-year phase I program.

12-month evaluation

- D-Train -- disk 1 (about 1 gigabyte of documents)
- T-Train -- topics 1-50
- D-Test -- disk 2 (about 1 gigabyte of documents)
- T-Test -- topics 51-100

- routing test -- topics 1-50 against disk 2

Because of the lateness of data availability, and the scarcity of sample relevance assessments for training, the emphasis was put on doing adhoc evaluation and only half of the routing test was done.

18-month evaluation

- D-Train -- disks 1 & 2 (about 2 gigabytes of documents)
- T-Train -- topics 51-100
- D-Test -- subset of future disk 3 (about 500 megabytes of documents)
- T-Test -- revised topics 1-50
- adhoc test -- topics 1-50 against disks 1 & 2
- routing test -- topics 51-100 against subset of disk 3

By the 18-month evaluation point, large numbers of relevance judgments were available for training (due to the many TREC-1 participants). This second evaluation therefore concentrated on the routing task, although adhoc evaluation was also done.

24-month evaluation

- D-Train -- disks 1 & 2 (about 2 gigabytes of documents)
- T-Train -- topics 1-100
- D-Test -- disk 3 (about 1 gigabyte of documents)
- T-Test -- topics 101-150
- adhoc test -- topics 101-150 against disks 1 & 2
- routing test -- topics 51-100 against all of disk 3

This data point corresponded directly to the TREC-2 data and therefore allows comparison between the 24-month TIPSTER results and the TREC-2 results.

4. SPECIFIC TASK GUIDELINES

Because the TIPSTER contractors and 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 TIPSTER task.

1. System data structures should be built using the initial training set (documents D-Train, training topics T-Train, and the relevance judgments). They may be modified based on the test documents D-Test, but not based on the test topics. In particular, the processing of one test topic should not affect the processing of another test topic. For example, it is not allowed to update a system knowledge base based on the analysis of one test topic in such a way that the interpretation of subsequent test topics was changed in any fashion.
2. There are several parts of 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 topics. 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 test document information (such as IDF 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. Method 1 -- completely automatic initial query construction.

   adhoc queries -- The system will automatically extract information from the topic (the topic fields used should be identified) 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. Method 2 -- manual initial query construction.

adhoc queries -- The query is constructed in some manner from the topic, either manually or using machine assistance. The methods used should be identified, along with the human expertise (both domain expertise and computer expertise) needed to construct a query. 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 Method 3.)

routing queries -- The queries should be constructed in the same manner as the adhoc queries for method 2, 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. Method 3 -- automatic or manual query construction with feedback.

adhoc queries -- The initial query can be constructed using either Method 1 or Method 2. 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 -- Method 3 cannot be used for routing queries as routing systems have typically not supported feedback.

5. EVALUATION METRICS

5.1 Recall/Precision Curves

Standard recall/precision figures were calculated for each TIPSTER and TREC system and the tables and graphs for the results were provided. 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

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

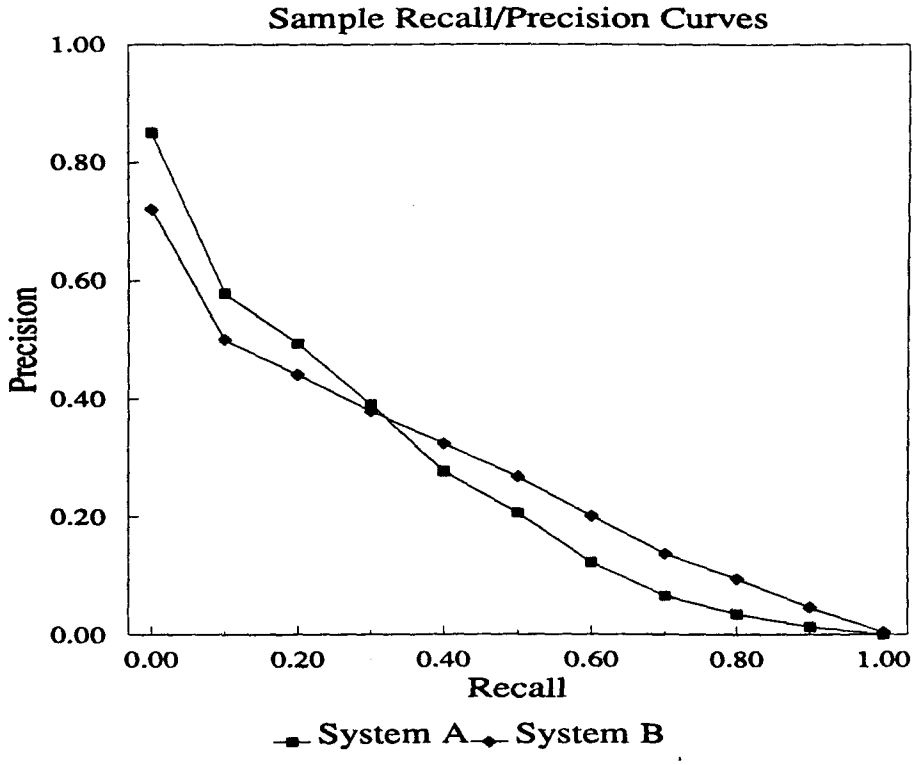
These curves represent averages over the 50 topics. The averaging method was developed many years ago [2] and is well accepted by the information retrieval community. It was therefore used unchanged for the TIPSTER evaluation. The curves show system performance across the full range of retrieval, i.e. 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 TIPSTER/TREC programs.

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 higher 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.

5.2 Recall/Fallout Curves

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

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



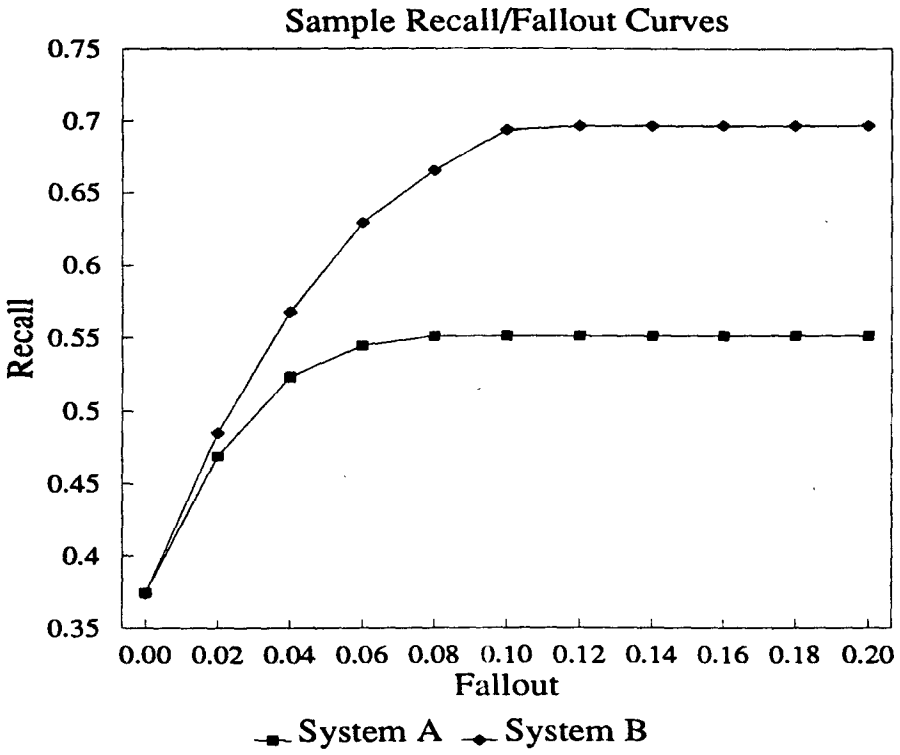


Figure 2 -- A Sample Recall/Precision Curve
Figure 3 -- A Sample Recall/Fallout Curve

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 are 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 performance does not show well in figure 3; the high recall performance 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 discrimination powers of these systems on a large document collection. Since recall/precision measures do not include any indication of the collection size, the recall and precision of a system based on a 1400 document collection could be the same as that of a system based on a million document collection, but obviously the discrimination powers on a million document collection would be much greater. This was not have been a problem on the smaller collections, but the discrimination power of systems on TIPSTER-sized collections is very important.

5.3 Single-Value Evaluation Measures

In addition to these recall/precision and recall/fallout curves, there were 3 single-value measures often used in TIPSTER. The first two measures are precision averages across the curves, and the third measure is precision at a particular cutoff of documents retrieved.

One of the averages, the non-interpolated average precision, combines the average precision for each topic, with that topic average computed by taking the precision after every retrieved relevant document. The final average corresponds to the area under an ideal (non-interpolated) recall/precision curve.

The second precision average (the 11-point precision average) averages across interpolated precision values (which makes it somewhat less accurate). It is calculated by averaging the precision at each of the 11 standard recall points on the curve (0.0, 0.1, ... ,1.0) for each topic. Often this average is stated as an improvement over some baseline average 11-point precision.

The third measure used is an average of the precision at each topic after 100 documents have been retrieved for that topic. This measure is useful because it contains no interpolation, and reflects a clearly comprehended retrieval point. It took on added importance in the TIPSTER 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.

5.4 Problems with Evaluation

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

For the TIPSTER 12-month evaluation and 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 were considered non-relevant by default and therefore the recall/precision curves became inaccurate after about 40% recall on average. The 18-month TIPSTER evaluation used a cutoff of 500 documents, and the TIPSTER 24-month and TREC-2 used the top 1000 documents. Figure 4 shows the difference in the curves produced by these 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 mostly resolved 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.

The second 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 difficult to perform failure analysis, or any type of 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 was provided for the TIPSTER 24-month evaluation and TREC-2, and more attention will be given to this problem in the future.

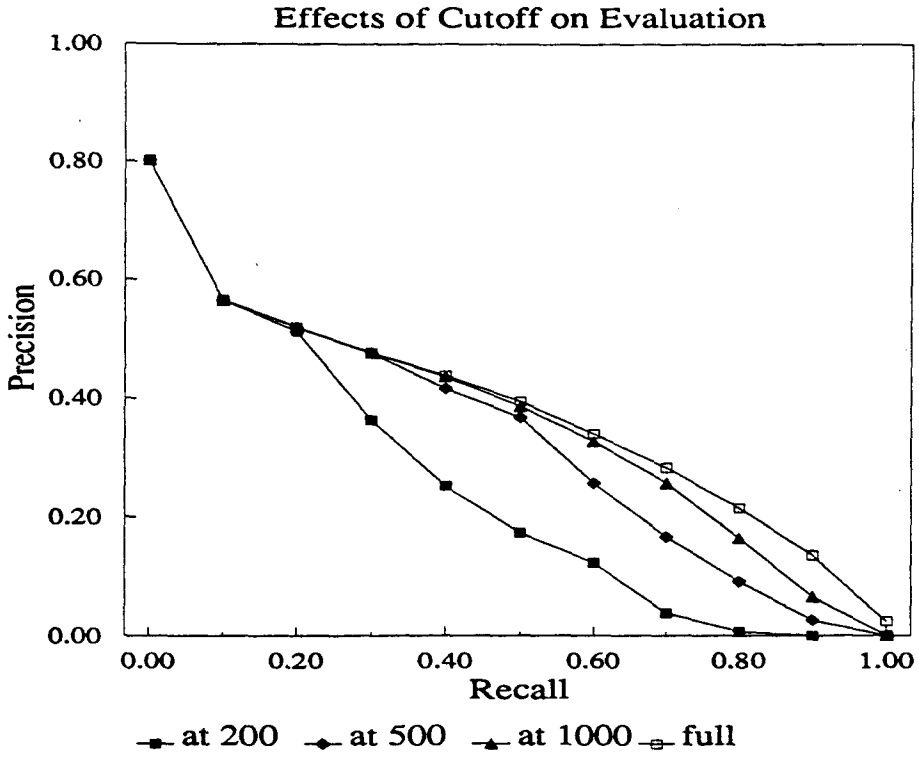


Figure 4: Effect of evaluation cutoffs on recall/precision curves

6. REFERENCES

[1] Belkin N.J. and Croft W.B. Retrieval Techniques. In Williams, M. (Ed.), *Annual Review of Information Science and Technology* (pp. 109-145). New York, NY: Elsevier Science Publishers, 1987.

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

[3] Harman D. (Ed.).*The Second Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology Special Publication 500-215, in press.

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