1 Common Evaluation Measures

- Recall
  A measure of the ability of a system to present all relevant items.

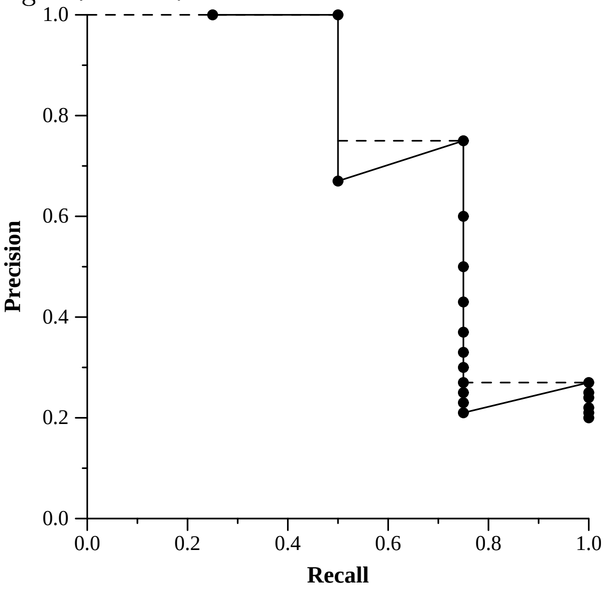
$$\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}}$$

- Precision.
  A measure of the ability of a system to present only relevant items.

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

Precision and recall are set-based measures. That is, they evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, precision can be plotted against recall after each retrieved document as shown in the example below. To facilitate computing average performance over a set of topics— each with a different number of relevant documents— individual topic precision values are interpolated to a set of standard recall levels (0 to 1 in increments of .1). The particular rule used to interpolate precision at standard recall level i is to use the maximum precision obtained for the topic for any actual recall level greater than or equal to i. Note that while precision is not defined at a recall of 0.0, this interpolation rule does define an interpolated value for recall level 0.0. In the example, the actual precision values are plotted with circles (and connected by a solid line) and the interpolated precision is shown with the dashed line.

> Example: Assume a document collection has 20 documents, four of which are relevant to topic t. Further assume a retrieval system ranks the relevant documents first, second, fourth, and fifteenth. The exact recall points are 0.25, 0.5, 0.75, and 1.0. Using the interpolation rule, the interpolated precision for all standard recall levels up to .5 is 1, the interpolated precision for recall levels .6 and .7 is .75, and the interpolated precision for recall levels .8 or greater is .27.



2 trec_eval Evaluation Report

The results from the cross-language track, the ad hoc task in the web track, and the routing task in the filtering track are ranked lists of documents. These lists are evaluated using `trec_eval`, a program written by Chris Buckley when he was at Cornell University that can be obtained by anonymous ftp from Cornell in the directory pub/smart at ftp.cs.cornell.edu. An evaluation report for a run evaluated by `trec_eval` is comprised of a header (containing the task and organization name), 3 tables, and 2 graphs as described below.

2.1 Tables

I. “Summary Statistics” Table
Table 1 is a sample “Summary Statistics” Table

Table 1: Sample “Summary Statistics” Table.

Summary Statistics	
Run	Cor7A1clt–automatic, title
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4674
Rel_ret:	2621

A. Run
A description of the run. It contains the run tag provided by the participant, and various details about the runs such as whether queries were constructed manually or automatically.

B. Number of Topics
Number of topics searched in this run (generally 50 topics are run for each task).

C. Total number of documents over all topics (the number of topics given in B).

i. Retrieved
Number of documents submitted to NIST. This is usually 50,000 (50 topics × 1000 documents), but is less when fewer than 1000 documents are retrieved per topic.

ii. Relevant
Total possible relevant documents within a given task and category.

iii. Rel_ret
Total number of relevant documents returned by a run over all the topics.

II. “Recall Level Precision Averages” Table.
Table 2 is a sample “Recall Level Precision Averages” Table.

A. Precision at 11 standard recall levels
The precision averages at 11 standard recall levels are used to compare the performance of different systems and as the input for plotting the recall-precision graph (see below). Each recall-precision average is computed by summing the interpolated precisions at the specified recall cutoff value (denoted by $\sum P_\lambda$ where P_λ is the interpolated precision at

Table 2: Sample "Recall Level Precision Averages" Table.

Recall Level Precision Averages	
Recall	Precision
0.00	0.6169
0.10	0.4517
0.20	0.3938
0.30	0.3243
0.40	0.2715
0.50	0.2224
0.60	0.1642
0.70	0.1342
0.80	0.0904
0.90	0.0472
1.00	0.0031
Average precision over all relevant docs	
non-interpolated	0.2329

recall level λ) and then dividing by the number of topics.

$$\frac{\sum\limits_{i=1}^{NUM} P_\lambda}{NUM} \qquad \lambda = \{0.0, 0.1, 0.2, 0.3, \ldots, 1.0\}$$

- Interpolating recall-precision
  Standard recall levels facilitate averaging and plotting retrieval results.

B. Average precision over all relevant documents, non-interpolated
This is a single-valued measure that reflects the performance over all relevant documents. It rewards systems that retrieve relevant documents quickly (highly ranked).

The measure is not an average of the precision at standard recall levels. Rather, it is the average of the precision value obtained after each relevant document is retrieved. (When a relevant document is not retrieved at all, its precision is assumed to be 0.) As an example, consider a query that has four relevant documents which are retrieved at ranks 1, 2, 4, and 7. The actual precision obtained when each relevant document is retrieved is 1, 1, 0.75, and 0.57, respectively, the mean of which is 0.83. Thus, the average precision over all relevant documents for this query is 0.83.

III. "Document Level Averages" Table
Table 3 is a sample "Document Level Averages" Table.

A. Precision at 9 document cutoff values
The precision computed after a given number of documents have been retrieved reflects the actual measured system performance as a user might see it. Each document precision average is computed by summing the precisions at the specified document cutoff value and dividing by the number of topics (50).

B. R-Precision
R-Precision is the precision after R documents have been retrieved, where R is the

Table 3: Sample "Document Level Averages" Table.

Document Level Averages	
	Precision
At 5 docs	0.4280
At 10 docs	0.3960
At 15 docs	0.3493
At 20 docs	0.3370
At 30 docs	0.3100
At 100 docs	0.2106
At 200 docs	0.1544
At 500 docs	0.0875
At 1000 docs	0.0524
R−Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.2564

number of relevant documents for the topic. It de-emphasizes the exact ranking of the retrieved relevant documents, which can be particularly useful in TREC where there are large numbers of relevant documents.

The average R-Precision for a run is computed by taking the mean of the R-Precisions of the individual topics in the run. For example, assume a run consists of two topics, one with 50 relevant documents and another with 10 relevant documents. If the retrieval system returns 17 relevant documents in the top 50 documents for the first topic, and 7 relevant documents in the top 10 for the second topic, then the run's R-Precision would be $\frac{\frac{17}{50}+\frac{7}{10}}{2}$ or 0.52.

2.2 Graphs

I. Recall-Precision Graph
Figure 1 is a sample Recall-Precision Graph.

The Recall-Precision Graph is created using the 11 cutoff values from the Recall Level Precision Averages. Typically these graphs slope downward from left to right, enforcing the notion that as more relevant documents are retrieved (recall increases), the more nonrelevant documents are retrieved (precision decreases).

This graph is the most commonly used method for comparing systems. The plots of different runs can be superimposed on the same graph to determine which run is superior. Curves closest to the upper right-hand corner of the graph (where recall and precision are maximized) indicate the best performance. Comparisons are best made in three different recall ranges: 0 to 0.2, 0.2 to 0.8, and 0.8 to 1. These ranges characterize high precision, middle recall, and high recall performance, respectively.

II. Average Precision Histogram.
Figure 2 is a sample Average Precision Histogram.

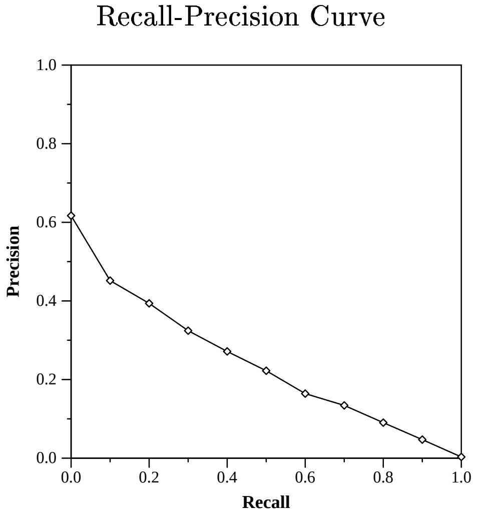


Figure 1: Sample Recall-Precision Graph.

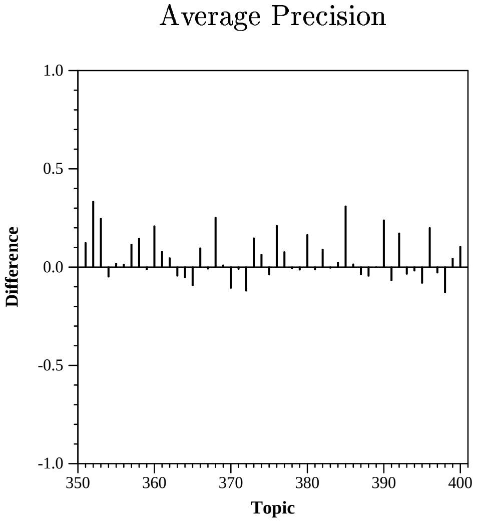


Figure 2: Sample Average Precision Histogram.

The Average Precision Histogram measures the average precision of a run on each topic against the median average precision of all corresponding runs on that topic. This graph is intended to give insight into the performance of individual systems and the types of topics that they handle well.

3 Question Answering Evaluation Report

The different tasks in the question answering track each used different evaluation metrics and have different evaluation reports.

3.1 Main task

The basic evaluation measure used in the main task is the reciprocal rank: the score for an individual question is the reciprocal of the rank at which the first correct response was found, or zero if no correct response was found in the first five responses. The score for a run as a whole is the mean of the reciprocal rank over the test set of questions. The judging for the question answering track distinguished between correct answers that were supported by the document returned and correct answers that were not supported. In strict evaluation, unsupported responses were counted as incorrect; in lenient evaluation unsupported answers were counted as correct.

The evaluation report for the main task consists of a table giving detailed evaluation scores for the run and a graph that compares the run to a hypothetical median run. An example of the table is shown in Table 4

Table 4: Sample QA Main Task Table.

Summary Statistics	
Run ID	insight
Num questions	492
Mean reciprocal rank (strict)	0.676
Mean reciprocal rank (lenient)	0.686
Num answers not found (strict)	152 (30.9%)
Num answers not found (lenient)	147 (29.9%)
Number of times NIL returned	120
Number of times NIL correctly returned	38
Percentage of answers system confident about	75%
Percentage of confident answers that were correct	77%

The scores given include:

- The mean reciprocal rank for both strict and lenient evaluation.
- The number and percentage of questions for which the correct response was not returned in the top five responses for both strict and lenient evaluation.
- The number of questions for which 'NIL' was returned as a a response. (NIL indicates the system's belief that no correct response is contained in the document collection.)
- The number of questions for which NIL was returned as a response and it was the correct answer.
- The percentage of questions for which the system was confident it had correctly determined the answer.

- The percentage of questions the system was confident about that were actually correct. (For this computation, the system was judged on its selection of one final answer, not on the list of five responses.)

A sample median graph is shown in Figure 3. The graph is a histogram of the number of questions for which the correct response was returned at a given rank. Plotted is both the run's results and the results for a hypothetical run that retrieved the correct response at the median rank for each question. The median is computed over the entire set of runs submitted to the main task.

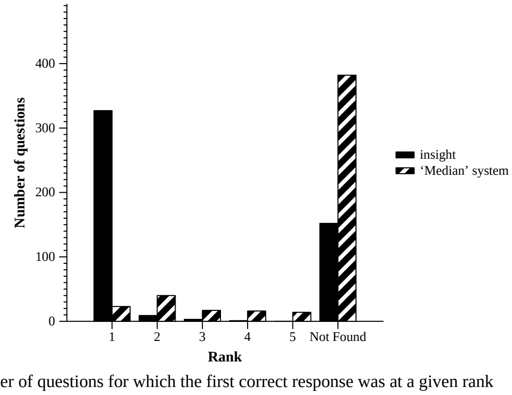


Figure 3: Sample QA Main Task Median Graph.

3.2 List task

The evaluation metric used for the list task is mean accuracy, where the accuracy of a single question is the number of distinct instances retrieved divided by the target number of instances (i.e., the number of instances the question specified should be retrieved). The evaluation report gives the run's mean accuracy computed over the 25 question in the test set. Also included is a histogram that shows the difference between the system's accuracy score and the median accuracy score for each question.

3.3 Context task

The context task was a pilot study to investigate how well systems can track discourse objects through a series of questions. Since answering later questions in a series requires correctly answering earlier questions in the series, a mean reciprocal rank score over all questions confounds important variables. Since there were only seven runs submitted to the task, median results are also uninformative. The evaluation report for the context task therefore consists simply of the rank at which the first correct response was returned for each of the 42 questions in the test set. Questions are numbered by series and then given a letter for the individual questions within the series. Thus question 3b is the second question of the third series.

4 Filtering Evaluation Report

The result of a filtering run is an unordered set of documents, so it cannot be evaluated using `trec_eval`. (Routing runs do produce a ranked list of documents and are thus evaluated using `trec_eval`.) The evaluation measures used in the TREC 2001 filtering track were a linear utility function (scaled when averaged) and a variant of F-beta. If R^+ is the number of relevant documents a run retrieved, R^- the number of relevant documents that were not retrieved, and N^+ the number of non-relevant documents that were retrieved, the F-beta score used in the track is defined as

$$\mathrm{T10F} = \begin{cases} 0 & \text{if } \mathrm{R}^+ = \mathrm{N}+ = 0 \\ \dfrac{1.25\mathrm{R}^+}{.25\mathrm{R}^- + \mathrm{N}^+ + 1.25\mathrm{R}^+} & \text{otherwise} \end{cases}$$

and the utility function as

$$\mathrm{T10U} = 2\mathrm{R}^+ - \mathrm{N}^+.$$

To compute the average utility over a set of topics, the T10U score for the individual topics was scaled between a maximum score of twice the number of relevant documents and a minimum score of -100.

The evaluation report for an adaptive filtering run consists of a table giving run characteristics and summary measures, a table and plot of average utility scores over different time periods, and a median graph. The batch filtering report contains just the characteristics table and median graph.

A sample characteristics table is given in Table 5. The characteristics of the run include whether the run was an adaptive or batch run, whether external resources were used in the run, and the measure the run was optimized for (F-beta, T10U, or neither). The scores reported are the recall of the retrieved sets averaged over all topics, the precision of the retrieved sets averaged over all topics, the mean utility, the mean F-beta score, and the number of topics for which no documents were retrieved.

Table 5: Sample Filtering Table.

Summary Statistics	
Run ID	CMUCATsr10
Subtask	adaptive
TREC data used in training?	yes
Reuters data used in training?	no
Other data used in training?	no
Optimized for	T10U
Number of Topics	84
Total retrieved	342552
Relevant retrieved	245386
Macro average recall	0.248
Macro average precision	0.603
Mean T10SU	0.228
Mean F-Beta	0.415
Zero returns	10

A sample median graph is shown in Figure 4. The graph shows the difference between the run's

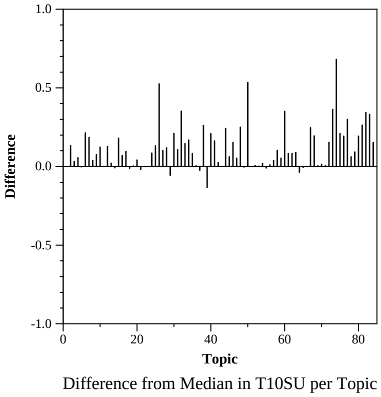


Figure 4: A Sample Filtering Median Graph.

evaluation score and the median score for each topic. The evaluation score is either the F-beta score or the utility score, depending on what the run was optimized for.

In adaptive filtering, systems can modify profiles based on relevance information of retrieved documents. One strategy is to have a "liberal" retrieval policy early in the process to gain more information and then become more stringent as more is learned. The time graph for adaptive runs plots average utility for four different time periods where time periods are labeled by the document identifiers that exist in the time period. A sample time graph is shown in Figure 5.

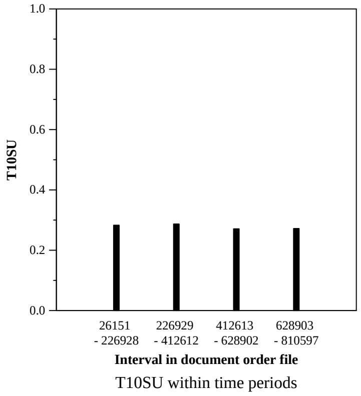


Figure 5: A Sample Filtering Median Graph.

5 Homepage Finding Evaluation Report

The result of a homepage finding task is a ranked list of documents, but the homepage finding task is a known-item search and thus us not evaluated using `trec_eval`. Instead, the runs were evaluated using the rank at which the first correct homepage was retrieved. The evaluation report consists of a table of evaluation scores and a median graph.

An example table of evaluation scores is given in Table 6. The table contains a description of the run that specifies whether document structure was exploited in the run (docstruct-used or docstruct-notused), whether URL text was exploited in the run (urltext-used or urltext-notused), and whether link structure was exploited in the run (links-used or links-notused). The evaluation scores reported include:

Table 6: Sample Homepage Finding Task Table.

Summary Statistics	
Run ID	tnout10epCAU
Run Description	docstruct-notused, urltext-used, links-used
Num topics	145
Mean reciprocal rank	0.774
Num found at rank 1	102 (70.3%)
Num found in top 10	128 (88.3%)
Num not found in top 100	7 (4.8%)

- The mean reciprocal rank for the run (see the question answering track description for a definition of mean reciprocal rank).

- The number and percentage of topics for which a correct homepage was retrieved in the first rank.

- The number and percentage of topics for which a correct homepage was retrieved in the top ten ranks (includes those topics for which the homepage was returned at rank one).

- The number and percentage of topics for which no correct homepage was returned in the top 100 ranks.

A sample median graph is shown in Figure 6. The graph plots the cumulative percentage of topics for which a correct homepage was retrieved by a given rank. Two lines are plotted, the results for the run, and the results for a hypothetical median run that retrieves the homepage at the median rank for each topic.

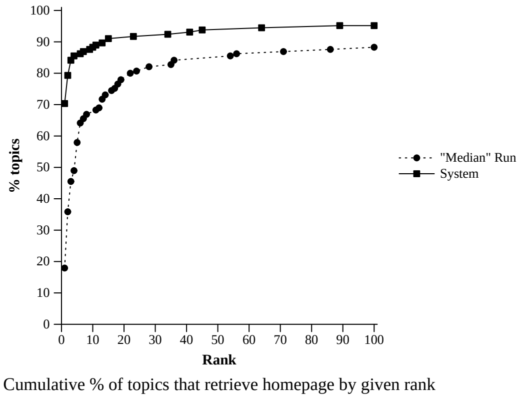


Figure 6: Sample Median Graph for the Homepage Finding Task.