APPENDIX A

This appendix contains the evaluation results for all TREC-5 runs. The initial pages list each of the runs (identified by the run tags) that were included in the different tasks/tracks. Associated with each tag is the organization that produced the run and additional information such as whether the queries were produced manually or automatically as appropriate. Following the run list is a description of the evaluation measures used for the main tasks and most of the tracks. The confusion, filtering, and interactive tracks use different evaluation measures as described in the respective track reports. The remainder of the appendix contains the evaluation results themselves, in the order given in the run list.

Evaluation Techniques and Measures

Categories

The results following this section are organized according to the task accomplished by the run: ad hoc, routing, or a track task.

I. Ad hoc
Retrieval using an "ad hoc" topic such as a researcher might use in a library environment. In TREC this implies that the input topic has no training material such as relevance judgments to aid in the construction of the input query.

A. Category A
Systems running TREC topics against all documents from TREC Disks 2 and 4.

B. Category B
Systems running TREC topics against the Wall Street Journal (WSJ) on TREC Disk 2. (Intended for new groups, allowing them to scale their systems to handle large collections.)

II. Routing
Retrieval using a "routing" query such as a profile to filter some incoming document stream. In TREC this implies that the input topic has training material, including relevance judgments against the training documents, to use in constructing the input query or profile. This query is then used against new documents (the test documents).

A. Category A
Systems running TREC topics against a set of Foreign Broadcast Information Service (FBIS) documents.

B. Category B
Systems running TREC topics against FBIS documents contained in files beginning with 'FB3'. (Intended for new groups, allowing them to scale their systems to handle large collections.)

Evaluation Measures

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

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

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

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

System Results Description

Each of the following pages contains the evaluation results for one run. A page is comprised of a header (containing the task and organization name), 3 tables, and 2 graphs.

Tables

Tables are generated by *trec_eval* courtesy of Chris Buckley using the SMART methodology as defined in Salton and McGill [1].

I. "Summary Statistics" Table
Table 1 is a sample "Summary Statistics" Table

Table 1: Sample "Summary Statistics" Table.

Summary Statistics	
Run	Cor5A2cr–category A, automatic, short topic
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	5524
Rel_ret:	2848

A. Run
A description of the run. It contains the run tag provided by the participant, and as applicable, whether the run is Category A or B, whether queries were constructed manually or automatically, and whether long or short topic descriptions were used.

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 $\times$ 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.

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

Recall Level Precision Averages	
Recall	Precision
0.00	0.5857
0.10	0.3927
0.20	0.3252
0.30	0.2799
0.40	0.2521
0.50	0.2131
0.60	0.1776
0.70	0.1395
0.80	0.0885
0.90	0.0415
1.00	0.0118
Average precision over all relevant docs	
non-interpolated	0.2109

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 precisions at the specified recall cutoff value (denoted by $\sum P_\lambda$ where P_λ is the precision at recall level λ) and then dividing by the number of topics.

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

- Interpolating recall-precision
Standard recall levels facilitate averaging and plotting retrieval results. Since the actual recall levels of individual queries are seldom equal to the standard levels, interpolation is used to define the precision at those points. The interpolated precision at the ith recall level (R^i) is defined to be the maximum precision at all points p such that $R^{i-1} \leq p \leq R^i$.

For example, if there are only 3 relevant documents, and these are retrieved at ranks 4, 9, and 20, then the exact recall points are 0.33, 0.67, and 1.0. Interpolated precisions are computed using the true recall values (precision 0.25 at recall 0.33, precision 0.22 at recall 0.67, and precision 0.15 at recall 1.0, respectively) and mapping them to the 11 recall cutoff values using the above rule. Therefore, the precisions at recall points 0.0, 0.1, 0.2, 0.3 are 0.25, the precision at recall points 0.4, 0.5 0.6, are 0.22 and precision at recall points 0.7, 0.8, 0.9, 1.0 are 0.15. Note that theoretically precision is not defined at a recall of 0.0, however this interpolation rule allows values to be determined.

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. 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.

Table 3: Sample "Document Level Averages" Table.

Document Level Averages	
	Precision
At 5 docs	0.4240
At 10 docs	0.3800
At 15 docs	0.3453
At 20 docs	0.3270
At 30 docs	0.2913
At 100 docs	0.2018
At 200 docs	0.1544
At 500 docs	0.0933
At 1000 docs	0.0570
R−Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.2404

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 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.

Graphs

I. Recall-Precision Graph

Figure 1 is a sample Recall-Precision Graph.

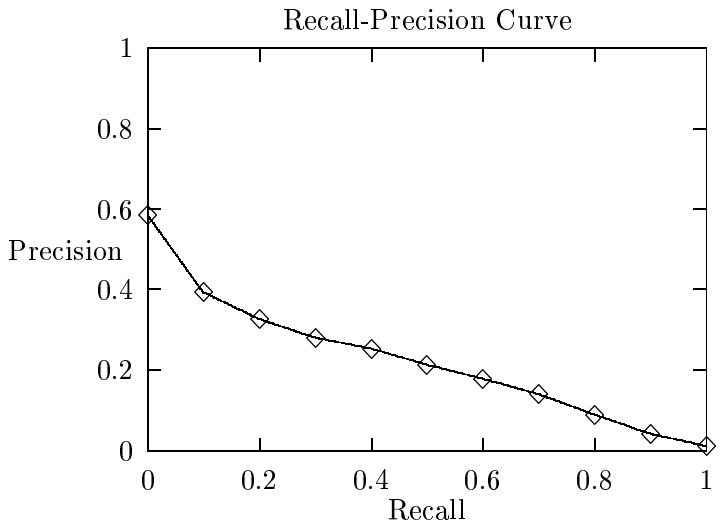

Figure 1: Sample Recall-Precision Graph.

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.

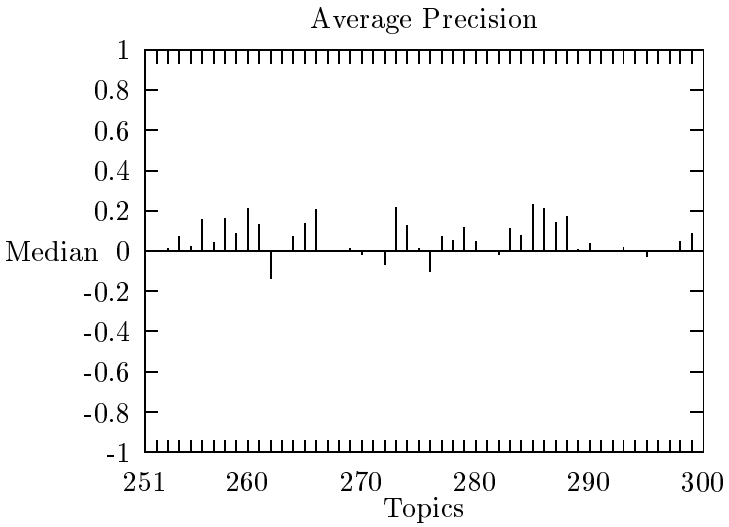

Figure 2: Sample Average Precision Histogram.

The Average Precision Histogram measures the precision of a run on each topic against the median 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.

References

[1] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, first ed., 1983.