# APPENDIX A

This appendix contains results for all TREC-3 participants. The first three pages list the system tags, organizations, and query construction methods for all runs. The next 9 pages describe the evaluation techniques and measures used. The rest of the appendix contains results.

**NOTE: Results are in landscape format. If you are using postscript, click on "orientation" and select the "swap landscape" button to view the results correctly.**

# Evaluation Techniques and Measures

## Categories

The results are separated into categories according to two tasks, adhoc or routing. Table 1 contains a further breakdown of runs depending on how the query was built (automatic, manual, or interactive) and the magnitude of the task (Category A or Category B).

TABLE 1

*Runs broken down by category and query construction method.*

| 48 Adhoc Runs | | | | |
|---|---|---|---|---|
| Category | Automatic | Manual | Interactive | Total |
| A | 28 | 12 | 2 | 42 |
| B | 4 | 2 | 0 | 6 |
| **49 Routing Runs** | | | | |
| Category | Automatic | Manual | Interactive | Total |
| A | 24 | 18 | 4 | 46 |
| B | 1 | 1 | 1 | 3 |

I. Adhoc.
   Retrieval using an "adhoc" 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 Disks 1 and 2 of the Tipster Collection.

   B. Category B.
      Systems running TREC topics against the Wall Street Journal (WSJ) on Disks 1 and 2 of the Tipster Collection. (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 all documents from Disk 3 of the Tipster Collection.

B. Category B.
Systems running TREC topics against only documents from the San Jose Mercury News (SJMN) on Disk 3 of the Tipster Collection. (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,

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

III. Fallout.
A measure of the ability of a system to filter out non-relevant items,

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

## System Results Description

Each page contains all the results for one system comprised of a header, 4 tables, and 3 graphs described as follows:

### Header

- The header contains the task and organization name, where task is either adhoc or routing and the organization is the organization name producing the run described on the page.

### Tables

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

I. "Summary Statistics" Table.
   Table 2 is a sample "Summary Statistics" Table.

TABLE 2

*Sample "Summary Statistics" Table.*

| Summary Statistics | |
|---|---|
| Run Number | CrnlEA–category A, automatic |
| Number of Topics | 50 |
| **Total number of documents over all topics** | |
| Retrieved: | 50000 |
| Relevant: | 9805 |
| Rel_ret: | 7267 |

A. Run Number.
   An identifier for a system of the form: *Tag* - *Category*, *Query Construction Method*, where *Tag* is the id of the run provided by the participant, *Category* is either Category A or Category B (full documents or a subset of full documents), and *Query Construction Method* is either automatic, manual, or interactive.

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

C. Total number of documents over all topics (the number of topics shown in B).
   i. Retrieved.
      Number of documents retrieved and sent in. This is normally 50,000 (50 topics $\times$ 1000 documents), but could be less.

ii. **Relevant.**
**Total possible relevant documents within a given task and category. Table 3 contains the possible relevant documents for TREC-3.**

TABLE 3
*Possible relevant documents within a given task and category for TREC-3.*

| Category | Adhoc | Routing |
|---|---|---|
| A | 9805 | 9353 |
| B | 3913 | 2559 |

iii. **Rel_ret.**
**Total number of relevant documents returned by a run for all the topics, i.e. the Number of Topics shown on the second line of the Summary Table.**

II. **"Recall Level Precision Averages" Table.**
**Table 4 is a sample "Recall Level Precision Averages" Table.**

TABLE 4
*Sample "Recall Level Precision Averages" Table.*

| Recall Level Precision Averages | |
|---|---|
| Recall | Precision |
| 0.00 | 0.7759 |
| 0.10 | 0.5910 |
| 0.20 | 0.5244 |
| 0.30 | 0.4735 |
| 0.40 | 0.4213 |
| 0.50 | 0.3595 |
| 0.60 | 0.3031 |
| 0.70 | 0.2365 |
| 0.80 | 0.1605 |
| 0.90 | 0.0664 |
| 1.00 | 0.0031 |
| Average precision over all relevant docs | |
| non-interpolated | 0.3419 |

A. **Precision at 11 recall cutoff values.**
**Each recall-precision average is computed by summing the precisions at the specified recall cutoff value (denoted by $\sum P_R$ where $P_R$ is the precision at recall cutoff value $R$) and then dividing by the number of topics, for TREC normally $NUM = 50$.**

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

- **Interpolating recall-precision.**
  **In order to graphically show precision at various recall averages, interpolation must be used. The interpolated precision at a recall cutoff $R$ is defined to be the maximum precision at all points $\leq$ R.**

  **For example, if there are only 3 relevant documents retrieved, 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 measure is not an average of the above cutoff values, but an average calculated after each relevant returned document.**

   **Consider a system returning 10 documents and four of the 10 documents are relevant. The rankings of the four documents are 1, 2, 4, 7 giving precisions of 1, 1, 0.75, and 0.57, respectively. By averaging the 4 precisions the single value measure of average precision over all relevant documents is 0.83.**

III. **"Document Level Averages" Table.**
   **Table 5 is a sample "Document Level Averages" Table.**

A. **Precision at 9 document cutoff values.**
   **Each document precision average is computed by summing the (50) precisions at the specified document cutoff value divided by the number of topics (50).**

B. **R-Precision is the precision after R documents (whether relevant or non-relevant) have been retrieved, where R is the number of relevant documents for a topic.**
   **R-Precisions are computed, one for each query (50), and then they are averaged.**

   **Suppose a topic with 50 relevant documents was run returning 200 documents. In the top 50 documents returned, 17 of them are relevant. Then the R-Precision is $\frac{17}{50}$ or 0.34.**

TABLE 5

*Sample "Document Level Averages" Table.*

| Document Level Averages | |
|---|---|
| | Precision |
| At 5 docs | 0.5760 |
| At 10 docs | 0.5540 |
| At 15 docs | 0.5480 |
| At 20 docs | 0.5370 |
| At 30 docs | 0.5187 |
| At 100 docs | 0.4168 |
| At 200 docs | 0.3413 |
| At 500 docs | 0.2254 |
| At 1000 docs | 0.1453 |
| R−Precision (precision after R docs retrieved (where R is the number of relevant documents)) | |
| Exact | 0.3890 |

**IV. "Recall Fallout Averages" Table.**

**Table 6 is a sample "Recall Fallout Averages" Table.**

TABLE 6

*Sample "Recall Fallout Averages" Table.*

| Recall Fallout Averages | |
|---|---|
| Recall | Fallout |
| 0.00 | 0.00000 |
| 0.10 | 0.00003 |
| 0.20 | 0.00008 |
| 0.30 | 0.00014 |
| 0.40 | 0.00023 |
| 0.50 | 0.00037 |
| 0.60 | 0.00058 |
| 0.70 | 0.00094 |
| 0.80 | 0.00175 |
| 0.90 | 0.00528 |
| 1.00 | 0.13424 |

**A. Fallout at 11 recall cutoff values.**

**Each recall-fallout average is computed by summing the fallouts at the specified recall cutoff value (denoted by $\sum F_R$ where $F_R$ is the fallout at recall cutoff value $R$) and then divided by the number of**

topics ($NUM = 50$).

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

- Interpolating recall-fallout.
  The recall at $X$ non relevant documents is used for the recall at $X-1$ non relevant documents. When fallout is not exactly defined at $R$, the fallout of $X-1$ is used.

## Tabular Interpretation

I. Recall Level Precision Averages.

A. Precision at 11 cutoff values.
This table allows comparisons of systems.

B. Average precision over all relevant documents.
This is a single valued number which reflects the performance over all relevant documents. It is intended to reward those systems retrieving relevant documents quickly (highly ranked).

II. Document Level Averages.

A. Precision at 9 document cutoff values.
Document level reflects actual measured system performance as a user might see it. However, the averages computed using document-level measures are difficult to compare.

B. R-Precision.
It is a measure that is intended to de-emphasize the exact ranking of the documents. As such, it is especially valuable for examining routing systems, where the real-life criteria is whether a document is given to a user, rather than at what rank the document would be. R-Precision is particularly useful in TREC where there are large numbers of relevant documents.

III. Recall Fallout Averages.
Fallout is a parallel measure to recall for measuring the nonrelevant documents. An effective retrieval system will exhibit maximum recall and minimal fallout.

## 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. This graph is useful for comparing systems. The graphs of different systems can be superimposed on the same graph to determine which system is superior. Curves closest to the upper right-hand corner of the graph (where recall and precision

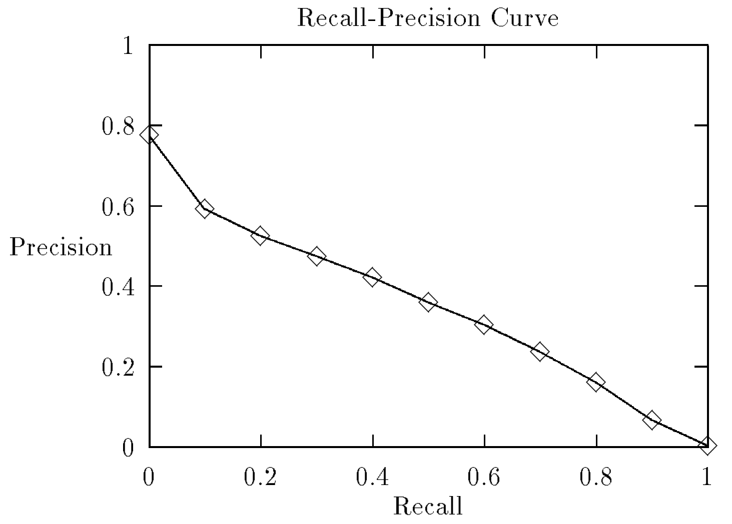


FIG. 1. *Sample Recall-Precision Graph.*

**are maximized) indicates the best performance. Comparisons are best made in three different recall ranges: 0 to 0.2, 0.2 - 0.8, and 0.8 to 1. These ranges characterize systems as high precision, middle recall, and high recall, respectively.**

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

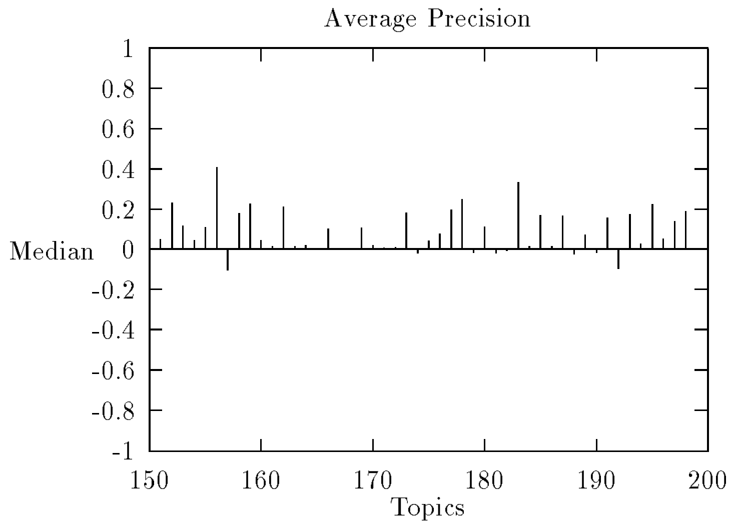


FIG. 2. *Sample Average Precision Histogram.*

**The Average Precision Histogram measures the precision of a system on each topic against the median precision of all systems on that topic.**

**III. Recall-Log Fallout Curve.**
**Figure 3 is a sample Recall-Log Fallout Curve.**

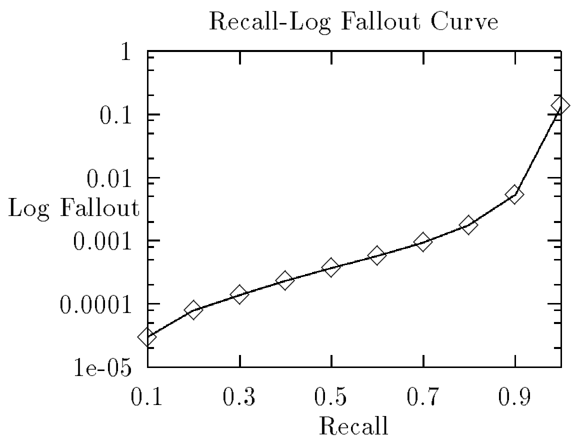


FIG. 3. *Sample Recall-Log Fallout Curve.*

**The Recall-Log Fallout Graph is generated using the 11 cutoff values from the Recall Fallout Averages (exactly like the Recall-Precision Graph). Curves closest to the lower left-hand corner of the graph (where recall is maximized and fallout is minimized) indicates the best performance. Fallout is graphed using a logarithmic scale for viewing purposes.**

## Graphical Interpretation

**I. Recall-Precision Graph.**
**This graph is the most common used to compare information retrieval systems. Typically these graphs slope downward from left to right, enforcing the notion that as more relevant documents are retrieved (recall increases) then the more non relevant documents are retrieved (precision decreases).**

**II. Average Precision Histogram.**
**This graph is intended to give insight into the performance of individual systems and the types of topics they handle well.**

**III. Recall-Log Fallout Curve.**
**This graph illustrates the same concept as the recall-precision graph, except the fallout measure is used. As systems retrieve all relevant documents (recall increases) the number of non-relevant documents returned increases (fallout increases).**