Notes on Evaluation for CLIR Task

"Evaluation Measures" and "System Results Description" in the followings are based on the TREC's, i.e. the description, "Evaluation Techniques and Measures" in Proc. of TREC-8, 2000[1],[2].

Relevance Judgment

The relevance judgement is done using four grades: highly relevant, relevant, partially relevant, and non-relevant. The TREC's evaluation program was run against two different lists of relevant documents produced two different thresholds of relevance, i.e., Level 1, in which "highly relevant (S)" and "relevant (A)" are rated as "relevant", and Level 2, in which S, A and "partically relevant (B)" were rated as "relevant".

Evaluation Measures

I. Recall

A measure of the ability of a system to present all relevant documents.

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

II. Precision

A measure of the ability of a system to present only relevant documents.

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

System Results Description

Each of the following pages contains the evaluation results for one run. A page in comprised of a header, i.e. Run-ID, 6 tables, and 8 graphs. The tables on the left are for Relevant Level1, and ones on the right are for Relevant Level2.

Tables

Tables are generated by the TREC's evaluation program *trec_eval* courtesy of Chris Buckley using the SMART methodology.

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

Table 1: Sample "Summary Statistics" Table.

Summary Statistics	
Run ID	NTCIR4
Task	J-J
Relevant Level	Level1 (S or A)
Method	automatic
Query Field	D
Query Type	D only
Number of Topics	60
Total number of documents over all topics	
Retrieved:	60000
Relevant:	3259
Rel-ret:	2379

A. Run-ID
Run-ID provided by the participant. It contains Group's ID of the participant.

B. Task
Sub-Task in which the run carried out.

C. Relevant Level
Relevant level of relevance judgments used for the evaluation of the run. There are two relevant levels, Level1 (S or A) and Level2 (S, A, or B). In the relevance judgement with "Relevance Level1" "highly relevant (S)" and "relevant (A)" are rated as "relevant", and in the one with "Relevance Level2" S, A and "partically relevant (B)" were rated as "relevant".

D. Method
Method used in building queries.

E. Query Type
Categories of field(s) of search topics used in the run. There are three categories, "D only", "T only", and "Others" The correspondences between the categories and the field(s) are shown below.

Table 2: Correspondences between the Categories and the Field(s).

Category	Field(s)
D only	DESC (Description)
T only	TITLE (Title)
Others	Other(s)

F. Number of Topics
Number of topics searched in this run.

G. Total number of documents over all topics (the number of topics given in F).

i. Retrieved
Number of documents submitted to the organizers of each sub-task. This is usually N $\times$ 1000 (N topics $\times$ 1000 documents), but less when fewer than 1000 documents are retrieved per topic.

ii. Relevant
Total possible relevant documents within a given sub-task.

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

II. "Recall Level Precision Averages" Table
Table 3 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 graphs (see below). Each recall-precision average is computed by summing the interpolated precision at the specified recall cutoff value (denoted by $\sum P_\lambda$ where P_λ is the interpolated precision at recall level λ and the dividing by the number of topics.

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

- 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 system that retrieve relevant documents quickly (highly ranked). The measure is not an average of the precision at standard recall levels. Rather, ir is the average of the precision value obtained after each relevant documents is retrieved.

Table 3: Sample "Recall Level Precision Averages" Table.

Recall Level Precision Averages	
Recall	Precision
0.00	0.8870
0.10	0.6912
0.20	0.5836
0.30	0.4819
0.40	0.4142
0.50	0.3701
0.60	0.2916
0.70	0.2224
0.80	0.1447
0.90	0.0809
1.00	0.0201
Average precision over all relevant documents	
non-interpolated	0.3602

Table 4: Sample "Document Level Averages" Table.

Document Level Averages	
	Precision
At 5 docs	0.6571
At 10 docs	0.5857
At 15 docs	0.5497
At 20 docs	0.4969
At 30 docs	0.4395
At 100 docs	0.2667
At 200 docs	0.1719
At 500 docs	0.0860
At 1000 docs	0.0486
R-Precision (precision after R docs retreived (where R is the number of relevant documents))	
Exact	0.3780

III. "Documents Level Averages" Table
Table 4 is a sample "Documents Level Averages" Table.

A. Precision at 9 documents 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 (49).

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-emphasized the exact ranking of the retrieved relevant documents, which can be particularly useful 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.

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 graphs is the most commonly used method for comparing systems. The plots of different runs can be superimposed on the same graphs to determine which run is superior. Curves closest to the upper right-hand corner of the graph (where recall and precision are maximaized) indicate the best performance. Comparisons are best made in three different ranges: 0.0 to 0.2, 0.2 to 0.8, 0.8 to 1. These ranges characterize high precision, middle precision, and high recall performance, respectively.

II. Difference from Median in Average Precision Graph

Figures 2, 3, and 4 are samples Difference from Median in Average Precision Graphs.

The Difference from Median in Average Precision Graph measures the average precision of a run on each topic against the median average precision of all corresponding runs in a sub-task, using a method, in a category of used filed(s) on that topic. This graph is intended to give insight into the performance of individual systems and the types of topics that thet handle well.

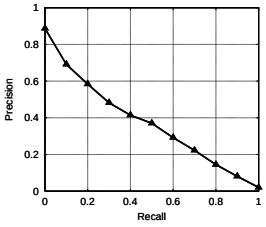

Figure 1: Sample Recall-Precision Graph

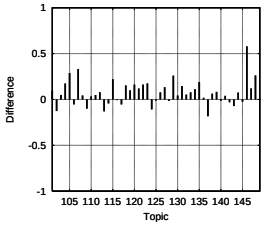

Figure 2: Sample Difference from Median in Average Precision per Topic (all)

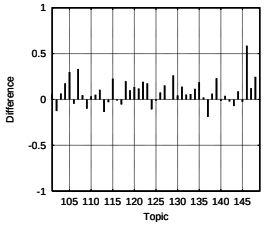

Figure 3: Sample Difference from Median in Average Precision per Topic (automatic)

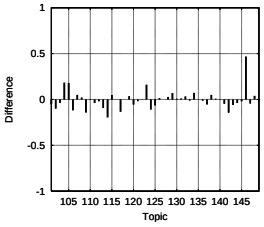

Figure 4: Sample Difference from Median in Average Precision per Topic (D)

References

[1] Voorhees E., Harman D. eds. The Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246, Maryland, U.S.A., 2000.

[2] Text REtrieval Conference (TREC). http://trec.nist.gov/ (visited January 12th, 2001).