APPENDIX A

Results of the Monolingual, Bilingual, Multilingual and Domain-Specific Tracks

Prepared by:

Martin Braschler

Eurospider Information Technology, Zurich, Switzerland

Results for CLEF 2004 Multilingual, Bilingual, Monolingual and Domain-Specific Tracks

The following pages contain the results for all valid experiments for the multilingual, bilingual, monolingual and domain-specific (GIRT) tasks that were officially submitted to the organizers of the CLEF 2004 campaign.

The first 4 pages give a listing of all runs and their characteristics:

Institution: the name of the organization responsible for run.
Country: country of participating organization.
Run Tag: unique identifier for each experiment. Can be used to match individual result pages with this list.
Task: track/task to which the experiment belongs.
Topic language: language of the topics used to create the experiment (ISO identifiers for language).
Topic fields: identifies the parts of the topics used to create the experiment (T=title, D=description, and N=narrative).
Run Type: type of experiment (automatic/manual).
Judged: specifies if experiment was used for relevance assessment pooling (Y=Yes)

The following eight pages compare the top entries for each track/task. The recall/precision curves for at most five groups are shown. The best entries for the mandatory title+description runs are shown – and only automatic experiments are used[1].

The rest of the pages provide the individual results for each official experiment. Each experiment is presented on one page containing a set of tables and two graphs:

1. The tables provide the following information:
   - Average precision figures for every individual query. This allows comparison of system performance for single queries, which is important since variation of performance across queries is often very high and can be significant.
   - Overall statistics, giving:
     - the total number of documents retrieved by the system
     - the total number of overall relevant documents in the collection, and
     - the total number of relevant documents actually found by the system
     - interpolated precision averages at specific recall levels (see above)
     - non-interpolated average precision over all queries (see above)
     - precision numbers after inspecting a specific number of documents (see above)
     - R-precision: precision after the last relevant document was retrieved.
2. The graphs consist of:
   - a recall/precision graph, providing a plot of the precision values for various recall levels. This is the standard statistic and is the one most commonly reported in the literature.
   - a comparison to median performance. For each query, the difference in average precision, when compared to the median performance for the given task, is plotted. This graph gives valuable insight into which type of queries is handled well by different systems.

The results page for a specific experiment can be most quickly located by using the table of contents at the beginning of the Appendix. This table is sorted by track/task, and topic language. The individual results pages are sorted by track/task and run tag.

More information on the interpretation of the standard measures used for scoring experiments (average precision, recall levels, precision/recall graphs, etc.) can be found in, e.g.: Martin Braschler, Carol Peters: CLEF 2002 Methodology and Metrics, Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. Lecture Notes in Computer Science, Vol. 2785, Springer 2003.

[1] **Errata:** Please note that just before printing we were informed that the run tagged UBmulti03, shown in the table for the best five performing groups for the multilingual task, was labelled incorrectly and is, in fact, a manual run.