APPENDIX C

SUMMARY

PERFORMANCE COMPARISONS

TREC-2, TREC-3, TREC-4

Karen Sparck Jones, November 7, 1995

The performance comparisons between TREC-2 and TREC-3 appeared as Appendix C to the TREC-3 Proceedings, i.e. to 'Overview of the Third Text REtrieval Conference (TREC-3)' Ed D.K. Harman, NIST Special Publication 500-225, National Institute of Standards and Technology, 1995 (and see also the opening Overview paper by Donna Harman, p 18).

The TREC-4 figures are from the TREC-4 conference working papers.

The tables present Precision performance at Document Cutoff 30.

The data are only for full Category A runs, not Category B, and cover only the higher levels of performance, not all the submitted runs.

The conventions are as follows: figures are not rounded; performance is assigned to 'blocks'; teams per block are in Proceedings order; the best of two official runs is taken where there are two, regardless of the retrieval method used (thus queries may be automatically or manually formed). The comparisons are for ad hoc, non-interactive searching, for routing, and for interactive searching (first tested for TREC-3), respectively: other TREC-4 test tracks had relatively few players and/or are not directly comparable and so are omitted here.

DOCUMENTS 30

	ADHOC			**ROUTING**			**INTERACTIVE ('Routing'-style TREC-3,** Task 2 TREC-4)	
	TREC2	TREC3	TREC4	TREC2	TREC3	TREC4	TREC3	TREC4
>=60		UMass City Berkeley		Cornell Dortmund	City			
>=55	UMass HNC VT	Cornell Mead		City Berkeley UMass Bellcore CMU CUNY	UMass Cornell Berkeley Dortmund Bellcore	City UMass Xerox		
>=50	Cornell Berkeley Dortmund CMU Verity Siemens CUNY	VT Westlaw ETH CUNY		Rutgers HNC GE TRW Verity Siemens	CMU Westlaw Logicon TRW Florida	Cornell CUNY		Cornell
>=45	City Bellcore ETH CITRI Conquest	NYU CMU RMIT RutgersK	Excalibur CUNY Waterloo	VT	Xerox NYU Verity ETH NSA NEC	Logicon GE/NYU	RutgersB Verity	Rutgers City
>=40			Berkeley Clarit/CMU Cornell GMU UMass InText ANU				City	
>=35			City GE/NYU					

NOTES:

a) UMass was formerly labelled Amherst.
b) RMIT was joined with CITRI in TREC-2.
c) RutgersK was one of two Rutgers groups in TREC-3.
d) Excalibur merged with Conquest during TREC-4.
e) While some groups recur in the tables, others have varied in participation in TREC or have not always done well enough (whether through goofs or deliberate (but unprofitable) experiment) to figure in the upper performance levels selected for the tables. Some of the TREC-4 participants also concentrated on tracks not included here.

COMMENTS:

1) Ad hoc best performance improved from TREC-2 to TREC-3, even though the TREC-3 topics (source requests) were less rich. The sharp fall in performance for TREC-4 must reflect the very minimal requests supplied for the tests.

2) Routing performance also fell slightly for TREC-4, apparently reflecting a tough set of test topics (though these were still rich). The smaller number of groups must be attributed to participants' diversion to other test tracks (compared with ad hoc which, though not compulsory for all players, was strongly recommended).

3) While upper-level ad hoc performance in TREC-3 was as good as routing, showing that rich starting topics can be almost as useful as extensive training data, the difference between the two in TREC-4 is, not surprisingly, quite marked.

4) In general, constant participants who initially performed well have continued to do so, though some others who started less well have successfully improved their performance.

5) Interactive performance remains disconcertingly poor.

OVERALL REMARKS:

1) The General Findings for TREC-2 given in my 'Reflections on TREC' (Information Processing and Management 31, 1995, 291-314), and noted as continuing to apply in TREC-3 in my previous comparison tables, remain true for TREC-4.

2) Many (very) different approaches give similar performance.

3) Good performance is obtained with variants of the statistical approach souped up with phrases etc: there indeed seems to be an emerging consensus on the right weighting concepts, and on the value of phrases (even statistically defined ones).

4) Routing performance seems to have reached a plateau, in that provided there is a good supply of training data and fairly well-developed starting queries, a very respectable performance level can be achieved.
NOTE ESPECIALLY, ALL of the TREC-4 routing results above were obtained with automatically developed queries.

5) The TREC-4 ad hoc performance, on the other hand, is clearly more representative of normal retrieval situations than the earlier results. MORE SPECIFICALLY, there was a clear shift between TREC-3 and TREC-4 from automatic to manual query as the normal route to query development, doubtless reflecting a perceived need to beef up the search input. However Cornell, the only high performer with automatic query in TREC-3, remained in the top groups for TREC-4.
NOTE ALSO that the definition 'manual' covers a wide range of human effort from the fairly minimal to the very intensive, though it seems that relatively modest effort can deliver as well as much more intensive work.

6) There are too few relevant interactive results for TREC-4 (i.e. ones comparable with those for TREC-3) to draw any useful conclusions about why interactive searching does not upgrade performance in a more striking way.

7) OVERALL, the difference between the best performance blocks for ad hoc >= 45 % P ; routing >= 55 ; interactive >= 50 % is not very large (equivalent to 13.5, 16.5 or 15 relevant documents respectively), and best automatic performance for ad hoc delivers >= 40 against manual >= 45 (equivalent to 12 as opposed to 13.5 relevant documents).

AS BEFORE, I EMPHASISE that

8) ALL of these points are broad brush ones: there may be significant differences

within performance blocks, and also none between members of adjoining blocks. However even without accepting the degree of similarity between runs that Jean Tague's TREC-3 Scheffe tests show, obliterating many of my block separations, and while also recognising that a Precision difference between 45 and 60 at Document cutoff may be not only statistically significant but meaningful to a user, I think a conservative view of performance differences is in order. Thus the really important point is (2) above, which applies both to performance in general and to best performance as exhibited in the tables.