APPENDIX C: SUMMARY PERFORMANCE COMPARISONS TREC-2, TREC-3, TREC-4, TREC-5

Karen Sparck Jones
Computer Laboratory, University of Cambridge

February 10, 1997

These comparisons are designed to allow trends over successive TRECs to emerge. But they are selective views, and do not summarise all there is to say about TREC overall. (TREC-1 is not included as it was a start-up, debugging, cycle.)

The performance comparisons between TREC-2, TREC-3 and TREC-4 originally appeared as Appendix C to the TREC-4 Proceedings, ie to 'The Fourth Text REtrieval Conference (TREC-4)' Ed D.K. Harman, NIST Special Publication 500-236, National Institute of Standards and Technology, 1996 (and see also the opening Overview paper by Donna Harman). The TREC-5 figures are from the TREC-5 conference working papers.

The tables cover Adhoc and Routing task results respectively. The former was required, the latter recommended, up to TREC-4. Adhoc runs were still required for TREC-5; and the TREC-5 Routing results are also given for completeness. However Interactive, as a separately defined task, though covered in the previous comparisons for TREC-3 and TREC-4, is not included here as the number of participants has remained small. Other TREC tracks are not covered either.

In the previous comparisons, changes in the nature of the topics used in successive TRECs were disregarded, and no distinctions were made between automatic and manual search queries. However for the Adhoc case in particular, further changes in the character of the supplied topic data, and in the precise specification of the options within the Adhoc task, make simple continuity of comparison through TREC-5 more problematic.

Thus for Adhoc up to TREC-4, the quantity and quality of content supplied per topic was reduced. But automatic and manual query formation were accepted as legitimate alternatives. So in the previous comparison tables the best performing of the official runs submitted per team was used, regardless of whether this was obtained by automatic or manual processing. However in TREC-5, the Adhoc topics were 'split' to give two forms. Short (S) and Long (L), with the latter adding content to the former. S was obligatory for automatic searching. L was supplied for manual searching, which could also be rather more interactive than in previous TRECs, and was also optional for automatic.

The data and relevant run sets used for the main comparison tables below are therefore as follows:

Adhoc

	TREC-1	TREC-2	TREC-3	TREC-4	TREC-5	
topic components					S	L
title	x	x	x			x
description	x	x	x	x	x	x
narrative	x	x	x			x
concepts	x	x				
queries	a/m	a/m	a/m	a/m	a	m/(a)

Routing

topics for TREC-1 – 3 were in the same style as Adhoc TREC-1 – 2
topics for TREC-4 – 5 were subsets drawn from all previous topics,
so had variable composition.

Tables 1 and 2 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 NOT in merit order, but in Workshop Notes or Proceedings order; the best of two official runs is taken, regardless of the particular strategy used, where there are two and these are deemed legitimate alternatives.

COMMENTS:

For both Adhoc and Routing, constant participants who initially performed well have continued to do so, though they have sometimes not profited from experiments, while others who started less well have successfully improved their performance. However many teams have not participated throughout the series, or have more recently concentrated on the tracks, so no inferences should be drawn where teams figure only occasionally in the tables.

Adhoc

1. Ad hoc best performance improved from TREC-2 to TREC-3, even though the TREC-3 topics were less rich. The sharp fall in performance for TREC-4 must reflect the minimal topics given for the tests. The further decline in TREC-5, even for the L topics which were fuller than the TREC-4 topics and like the TREC-3 ones, reflects the fact that the topics were deemed 'difficult' in relation to the definition of relevant documents.

2. The generally low levels of performance (even for the better-performing teams) in TREC-4 and TREC-5 must be taken as representing a more realistic retrieval situation than TREC-2 and TREC-3, as far as the S topics are concerned, though how far the TREC-5 topic 'difficulty' is representative is not clear.

3. In TREC-2 and TREC-3 automatic query formation was more common than manual, and often performed well, appearing even in the top blocks. Indeed there was relatively more use of automatic query in TREC-3 than TREC-2. But in TREC-4 there was a clear shift towards manual, doubtless in response to the perceived need to beef up the initial minimal

Precision	TREC-2	TREC-3	TREC-4	TREC-5		
	a/m	a/m	a/m	Short a	Long m	Long a
≥ 60		UMass City Berkeley				
≥ 55	UMass HNC VT	Cornell Mead				
≥ 50	Cornell Berkeley Dortmund CMU/Clarit Verity Siemens CUNY	VT Westlaw ETH CUNY				
≥ 45	City Bellcore ETH RMIT Conquest	NYU CMU/Clarit RMIT RutgersK	Excalibur/ Conquest CUNY Waterloo		ETH	
≥ 40	...	...	Berkeley Clarit/CMU Cornell GMU UMass InText ANU		Waterloo	
≥ 35	...	...	City GE/NYU	**Lexis	ANU Clarit Cornell GE/NYU GMU Lexis	
≥ 30	...	...	...		OpenText CUNY Berkeley	City CUNY ETH
≥ 25	...	...	...	Apple City Cornell IBMTJW ETH UMass	DCU IBM	Apple GE/NYU RMIT Berkeley

Table 1: Average precision after 30 documents retrieved for the Adhoc task

Precision	TREC-2	TREC-3	TREC-4	TREC-5
≥ 60	Cornell Dortmund	City		
≥ 55	City Berkeley UMass Bellcore CMU/Clarit CUNY	UMass Cornell Berkeley Dortmund Bellcore	City UMass Xerox	
≥ 50	Rutgers HNC GE TRW Verity Siemens	CMU/Clarit Westlaw Logicon TRW Florida	Cornell CUNY	City Cornell UMass
≥ 45	VT	Xerox NYU Verity ETH NSA NEC	Logicon GE/NYU	CUNY
≥ 40	...	...	...	GE/NYU ETH Berkeley

Table 2: Average precision after 30 documents retrieved for the Routing task

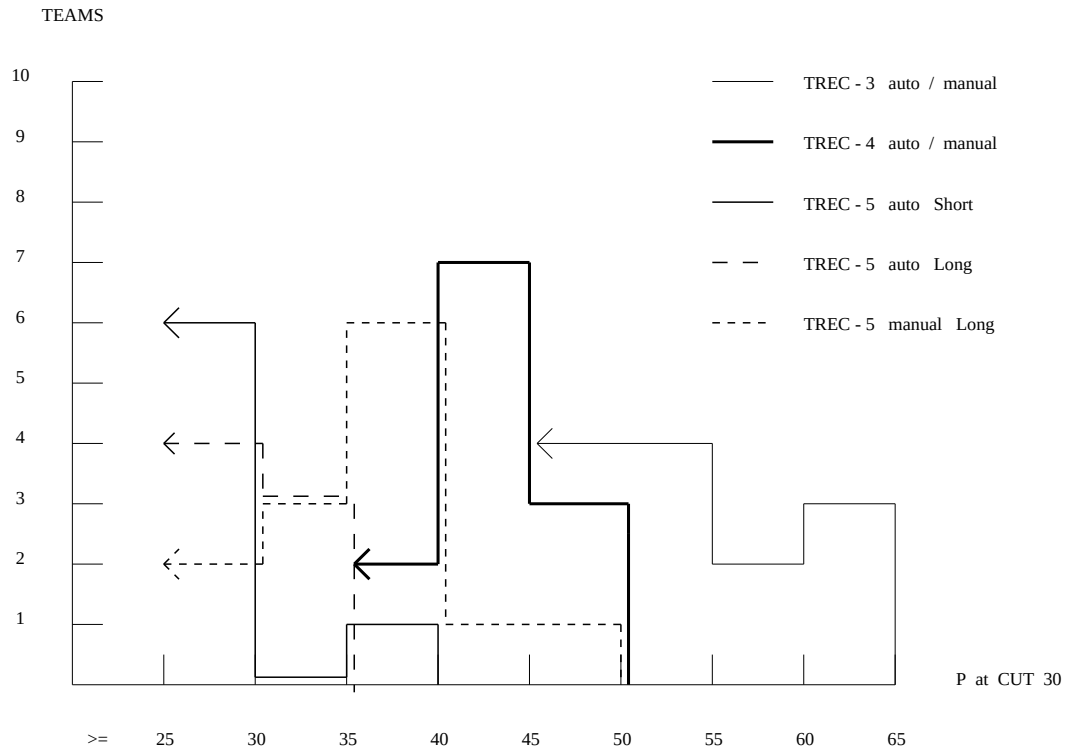

Figure 1: Teams performance distribution (upper levels) TREC-3 – 5 for the Adhoc task

topics, with almost all the teams covered by the table using manual queries. However at least one of the top-level teams using automatic query (Cornell) continued to do comparatively well in TREC-4. It is evident that manual query formation was advantageous for TREC-5 even when the same initial topic information (L) was used. But the L topics supplied more information for query formation than the S ones, so the comparison between S automatic queries and L manual ones for TREC-5 is not strictly fair. It is not completely evident what inferences are to be drawn from the outlying lead results in TREC-5. (Late Note: After the TREC-5 conference, the Lexis result marked ** in Table 1 was reclassified from automatic to manual.)

The histogram in Figure 1 shows the table data for TREC-3 – 5 in a more direct form (TREC-2 is omitted because of the 'over-good' topics.) The bars in the histogram show the number of teams that achieved a precision value (for Document Cutoff 30) in each given range. The backward-pointing arrows simply indicate that there were other teams with lower performance that are not included in any of the analysis in this Appendix.

It is important to note that the definition "manual" covers a wide range of human effort from the fairly minimal to the very intensive, and was also explicitly widened in TREC-5 to allow 'feedback' strategies. It is nevertheless not clear, in general, what forms of manual effort are especially profitable, or how far intensive effort (and hence time) pays off, or how manual input and automatic devices are best combined. In earlier TRECs it appeared that relatively modest effort could deliver as well as much more intensive work. Detailed analysis is needed for TREC-4 and TREC-5.

Routing

1. Routing has shown a slow decline in overall performance, again reflecting less good topic starting information: thus both TREC-4 and TREC-5 had 'tough' topics.

2. But when topics are less problematic, and there is rich training data (as with TREC-2 and TREC-3) a good level of performance can be obtained.

3. In TREC-2 automatic query formation was only slightly more common than manual, but by TREC-5 manual formation (for the teams shown in the table) had disappeared, reflecting the value of the large training data availability and its utility in compensating for any weakness in automatic query management.

Comparing Adhoc and Routing results, performance in TREC-2 and TREC-3 reached similar levels, attributable to the good topic specifications for the former even though Routing also benefitted in the same way and from the training data. However Adhoc performance has fallen much below that for Routing in TREC-4 and TREC-5, since Routing has been able to benefit from training data.

OVERALL REMARKS:

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

2. The overall results, and the general findings, for early TRECs reported in 'Reflections on TREC', Information Processing and Management 31 (3), 1995, p 309 and p 311 respectively, still essentially hold.

3. However while for earlier TRECs some convergence could be detected among practitioners of the statistical approach, the changing data conditions mean that strategy details are open to investigation.

4. More generally the range of devices, and of combinations of devices, in TREC remains very wide. So while one natural conclusion is that relatively simple and wholly or largely automated approaches can perform reasonably, the effect of the data changes over successive TRECs show that more understanding of the effects of environment variables on system parameters, for large text files, is required. As already noted, a detailed comparative analysis of what manual query formation involves and achieves, and how it related to independent rather than query-dependent document indexing, would be very useful.

NB: as in the previous comparisons, all the points made here are broad brush ones; and the nature of the tables must always be borne in mind. Thus there may be real differences within performance blocks, and also none between members of adjoining blocks. However as it is not obvious what performance differences are statistically significant, what differences are meaningful to users, and how significance and meaning are related, it is only proper to take a generally rather conservative view of apparent performance differences in the tables. More concretely, Precision of 55% and 45% are respectively equivalent to 16.5 and 13.5 relevant documents retrieved, a difference which may not matter much to a user.