SUMMARY PERFORMANCE COMPARISONS TREC-2, TREC-3, TREC-4, TREC-5, TREC-6

Karen Sparck Jones
Computer Laboratory, University of Cambridge

July 15, 1998

Data

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, TREC-4 and TREC-5 originally appeared as Appendix B to the TREC-5 Proceedings, i.e. to *The Fifth Text REtrieval Conference (TREC-5)*, Ed. E.M. Voorhees and D.K. Harman, NIST Special Publication 500-238, National Institute of Standards and Technology, 1997 (and see also the Editors' opening Overview).

The TREC-6 figures, like those for TREC-4 and TREC-5, are from the relevant conference working papers.

The tables cover Adhoc and Routing task results respectively. The Adhoc task, originally mandatory, has later still been strongly recommended, and most participants have therefore submitted runs for it. Routing has become more optional, but has still attracted submissions, and figures for it for TREC-6, its final year, are therefore included for completeness. Other tasks, i.e. tracks, are not covered here since, though they have become more important for TREC as a whole, they are too variable in definition and participation for systematic comparisons.

In the earlier comparisons for TREC-2 - 4, 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, there were further changes in TREC-5 and in TREC-6 in both the character of the supplied topic data and the precise specification of the options within the task. Trend comparisons over the whole series of TRECs can therefore only be at a general level, rather than in detail.

Thus for Adhoc up to TREC-4, the quantity and quality of content supplied per topic was reduced. But automatic and manual *modes* of query formation were accepted as legitimate alternatives. So in the comparisons for TREC-2 - 4 the best performing of the official runs submitted per team have been used, regardless of whether this was obtained by automatic or manual processing (details can be recovered from the Proceedings). However for TREC-5, the Adhoc topics were 'split' to give two *versions*, Short (S) and Long (L), with the latter adding content to and subsuming 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. For TREC-6, the topics were further separated into Very short (V), Short and Long, with automatic runs on S required and V and L runs optional. But it should be noted that it was found that in the original topic preparation for TREC-6 the S versions did not necessarily subsume the V ones: the S versions not infrequently covered only addditional, complementary topic content (see Voorhees and Harman's Overview). This probably accounts for the relatively good V performance and relatively poor S results. Manual searching was taken as using the L versions, as for TREC-5, and could be interactive. The comparisons for TREC-5 and TREC-6 Adhoc cover the best runs for each team for each topic version and query mode.

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

Adhoc

topic components TREC	1	2	3	4	5		6		
					S	L	V	S	L
title	x	x	x	x		x	x		x
description	x	x	x	x	x	x		x	x
narrative	x	x	x			x			x
concepts	x	x							
queries	a/m	a/m	a/m	a/m	a	m/(a)	(a)	a	m/(a)

Routing

topics for TREC-1 - 3 were in the same style as Adhoc TREC-1 - 2
topics for TREC-4 - 6 were subsets drawn from all previous topics,
so had variable composition.

Further comments and analysis follow the main tables.

Tables 1 and 2 present Precision performance at Document Cutoff 30. Table 1, for Adhoc, is divided into results for TREC-2 - 4 (Table 1a) and for TREC-5 - 6 (1b).

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 original run list 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. Simple, hopefully sufficiently identifiable, short names have been given to the teams.

TABLE 1a

ADHOC - DOCUMENT CUTOFF 30

	TREC-2	TREC-3	TREC-4	==> (TREC-5, -6 Table b)
	a/m	a/m	a/m	
>= 60		UMass City Berkeley		
>= 55	UMass HNC VT	Cornell Mead		
>= 50	Cornell Berkeley Dortmund CMU/Clarit Verity Siemens CUNY	Verity VT Westlaw ETH CUNY		
>= 45	City Bellcore ETH CITRI/RMIT Conquest	NYU CMU/Clarit RMIT RutgersK	Excalibur/ Conquest CUNY Waterloo	==> (best TREC-5 level)
>= 40	...	...	Berkeley Clarit/CMU Cornell GMU UMass InText ANU	
>= 35	...	...	City GE/NYU	
>= 30	...	...	...	
>= 25	...	...	...	
>= 20	...	...	...	==>

TABLE 1b

ADHOC - DOCUMENT CUTOFF 30

	TREC-5 S a	TREC-5 L a	TREC-5 L m	TREC-6 V a	TREC-6 S a	TREC-6 L a	TREC-6 L m
>= 60							
>= 55							
>= 50							Waterloo
>= 45			ETH				Clarit
>= 40			Waterloo				ANU
>= 35	Lexis *		ANU Clarit Cornell GE/NYU GMUetc Lexis				GEetc Lexis
>= 30		City CUNY ETH	OpenText CUNY Berkeley	Apple ATT City IRIT Lexis CUNY Waterloo		ANU Cornell IRIT CUNY Berkeley	ISS Berkeley
>= 25	Apple City Cornell IBMTJW ETH UMass	Apple GE/NYU RMIT Berkeley	DCU IBM	DCU ISS	ATT ANU City Cornell GMUetc IBMTJWs IRIT Lexis Waterloo	City IBMTJWg MDS/RMIT UMass GMUetc	FS GMUetc
>= 20	...	...	...	MDS/RMIT Glasgow	Apple GEetc IBMTJWg MDS/RMIT CUNY Berkeley Maryland UMass Verity	Verity	Glasgow

* reclassified to manual after the conference

TABLE 2

ROUTING - DOCUMENT CUTOFF 30

	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
>= 60	Cornell Dortmund	City			
>= 55	City Berkeley UMass Bellcore CMU/Clarit CUNY	UMass Cornell Berkeley Dortmund Bellcore	City UMass Xerox		ATT
>= 50	Rutgers HNC GE TRW Verity Siemens	CMU/Clarit Westlaw Logicon TRW Florida	Cornell CUNY	City Cornell UMass	City Cornell CUNY
>= 45	VT	Xerox NYU Verity ETH NSA NEC	Logicon GE/NYU	CUNY	Clarit IRIT ETH Waterloo
>= 40	...	...	...	GE/NYU ETH Berkeley	SRI Berkeley UCSD UMass Verity

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. Performance for TREC-5 and TREC-6 is generally similar, presumably reflecting a data 'plateau', even for the L versions of the topics.

2. The lower levels of performance (even for the better-performing teams) in TREC-4 - TREC-6 must be taken as representing a more realistic retrieval situation than TREC-2 and TREC-3. This statement has, however, to be heavily qualified. The L versions of the topics used in TREC-5 and TREC-6, though less elaborate than the very full earlier ones, are still more elaborate than are typically encountered in Adhoc retrieval practice, especially as end-user input to an automatic system. The defects of the TREC-6 S versions, already noted, probably depressed performance, and the S results for TREC-4 and TREC-5 may have been similarly affected. It is thus unfortunately not possible to draw any grounded inferences, based on systematic comparisons, about the effects of increasing topic fullness on performance. The only runs with any fairly direct bearing on practical situations, where end-users approach automatic search systems in a simple-minded way, are therefore only the (optional) TREC-6 V version ones. These suggest that where at least some attention is paid to the choice of the few initial search terms, adequate, though not high, performance can be obtained with automatic techniques, even without explicit relevance feedback. The V versions averaged only 2.6 words per topic.

   (It would be very useful to enhance the TREC-6 results with new 'proper' short version runs: these experiments would use topic titles as well as descriptions, and thus be true intermediates between V and L: to avoid confusion with the older Short versions, they might be labelled "Medium".)

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 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, quite full, initial topic information (L) was used, and the same applies to TREC-6.

But 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 device or 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 human effort could deliver as well as much more intensive work, but this was from good bases. Detailed analysis is needed for TREC-4 - 6

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. Performance in TREC-6 is slightly better, reflecting data with better fitting between topics and document file.

2. In general, when topics are not problematic, and there is rich training data (as with TREC-2 and TREC-3 and also TREC-6) 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 general findings about retrieval strategies for early TRECs reported in 'Reflections on TREC', *Information Processing and Management* 31 (3), 1995, p 309 and p 311 respectively, still essentially hold. Thus term weighting, query expansion and so forth are valuable, and in automatic searching quite simple strategies can be as effective as more elaborated ones, so e.g. sophisticated natural language processing is not especially helpful. This has led to some convergence on what may be called the *generic* $tf * idf$ paradigm with relevance feedback refinement. But even with good data (as illustrated by the TREC-6 L version topics, Precision at Document Cutoff 30 is more often than not below 30 %. For the collection data this corresponds roughly to Recall of 30 %.

3. Moreover the range of specific devices, and of combinations of devices, in TREC remains very wide, so more understanding of the effects of environment variables on system parameters for large text files is required, while, as already noted, a detailed comparative analysis of what manual query formation contributes would be very useful.

4. 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 whether significant differences are also meaningful, it is only proper to take a generally rather conservative view of apparent performance differences in the tables. More concretely, Precision of 45 % and 35 % are respectively equivalent to 13.5 and 10.5 relevant documents retrieved, a difference which may not matter much to a user; and even if the difference between 45 % and 40 % was statistically significant, the corresponding difference between 13.5 and 12 relevant documents retrieved would almost certainly not matter.