SUMMARY PERFORMANCE COMPARISONS TREC-2 THROUGH TREC-8

Karen Sparck Jones
Computer Laboratory, University of Cambridge

December 8, 1999

The context

This comparison series has attempted to illustrate long-term TREC trends, as embodied in the results for the baseline Adhoc task. As in last year's comparisons, covering TREC-2 - TREC-7, from TREC-5 onwards there has been a more careful separation of different *versions* of the topics, ranging from Very short (titles only) to Long (titles, descriptions and narratives), and between automatic and manual *modes* of query formulation: see the detail given in Table 1.

While last year's comparisons (Appendix B, TREC-7 Proceedings) gave performance details for the whole series from TREC-2 onwards, this year's detail is restricted to TREC-7 and TREC-8 only. First, the way the TREC-6 topics were formed could lead to titles and descriptions that were viewed as complementary rather than as less or more inclusive: this meant that controlled study of the effects of increasing topic length and detail was impossible. In TREC-7 and TREC-8 title terms are included in descriptions (so the difference between descriptions and titles+descriptions is in term frequency for the queries): TREC-7 and TREC-8 therefore supply two cycles of testing on the same topic basis. At the same time, it is evident from the detailed results for these two cycles in Table 2 that there is little difference in performance, whether of best levels or (to a considerable extent) by hardy perennial teams. The TREC-8 results can therefore be seen as a 'wind-up' on the long programme of Adhoc evaluations with the 'traditional' TREC data, and the end of a phase that is also signalled by the fact that evaluation with this type of data is being mothballed for TREC-9.

Table entries

Table 2 follows the same conventions as in previous summaries. Thus the detailed figures are taken from the Working Notes, and cover only the better performing, not all, the teams.

The conventions are as follows: figures are not rounded; performance is assigned to 'blocks'; teams per block are NOT in merit order, but in in Working Notes results order; where there is more than one run per team the best is taken, regardless of the particular strategy used. Simple, hopefully sufficiently identifiable, short names have been given to the teams (with some streamlining where teams have changed name or composition over the years).

TABLE 1 : TOPIC DETAILS

Topic fields available as base for queries, TREC-2 - TREC-7 :

	(TREC-1	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7	TREC-8
T= title	x	x	x		x	x	x	x
D= description	x	x	x	x	x	x	x	x
N= narrative	x	x	x		x	x	x	x
C= concepts	x	x						

Average topic and field length :

Total	107.4	130.8	103.4	16.3	82.7	88.4	57.6	51.8
T	3.8	4.9	6.5	-	3.8	2.7	2.5	2.5
D	17.9	18.7	22.3	16.3	15.7	20.4	14.3	13.8
N	64.5	78.8	74.6	-	63.2	65.3	40.8	35.5
C	21.2	28.5	-	-	-	-	-	-

TABLE 2 : RETRIEVAL PERFORMANCE, TREC-7, TREC-8

TREC ADHOC SEARCH RESULTS FOR PRECISION AT DOCUMENT CUTOFF 30

KEY TO TABLE NOTATIONS :

a = fully automatic searches
m = manual searches

V	= very short queries, i.e.	title only from topics, aka	T
S	= short queries	description only	D
M	= medium queries	title+description	T+D
L	= long queries	title+description+narrative	T+D+N

/contd

	TREC-7 a V	TREC-7 a S	TREC-7 a M	TREC-7 a L	TREC-7 m L	TREC-8 a V	TREC-8 a S	TREC-8 a M	TREC-8 a L	TREC-8 m L
>=60										ManInst
>=55					Clarit					IITetc
>=50					ManInst Waterlo					Oracle
>=45					GMUetc					Clarit GEetc
>=40		NEC	ATT Cityetc UMass	BBN Cityetc NEC UMass	ANU Harris Berkely Toronto	CUNY		ATT FUB IBMTJWs Msoft MIT CUNY	FUB Fujitsu Msoft MIT CUNY Neuchat	
>=35	Cityetc	Cornell CUNY Fujitsu	Lexis RMIT	ANU Cornell CUNY IRIT TwentyO Iowa	GEetc Lexis	ATT Fujitsu IBMTJWs Msoft MultTxt RICOH Sab/Crn	UMass	Fujitsu GEetc IBMTJWg IRIT MultTxt NTT Sab/Crn TwentyO Neuchat UMass Twente	ACSys GEetc IRIT JHopk NTT Sab/Crn UMass	
>=30	ATT Cornell CUNY Fujitsu Lexis NEC NTTData RMIT Waterlo	IBMTJWs	IBMTJWg IRIT	GMUetc NTTData Rutgers Berkely UNC	FS	ACSys RMIT TwentyO UMass	IBMTJWs Sab/Crn	ACSys CMU IITetc ImperC JHopk RICOH RMIT Marylnd	RMIT	
>=25	ANU Avignon GEetc IBMTJWg ETH Berkely Marylnd			FUB ImperC JHopk NSA		City/M		UNCy	CMU Dartmth	

Performance summary

To give a final overview of performance from TREC-2 - TREC-8, Table 3 gives the highest level of performance reached in each TREC for the various versions and modes.

As this table clearly shows, the early TRECs with 'good' topics reached high levels of performance in both automatic and manual modes; performance in the middle TRECs declined under the much less favourable data conditions (whether of topic information or relevant document accessibility); then in TREC-7 and TREC-8 performance for automatic mode in particular revived. This must be attributed to superior systems, since best manual performance has remained on a plateau. More specifically, amplifying on Tables 2 and 3, it is clear that the better level of performance in the TREC-7 and TREC-8 evaluations was the same.

TABLE 3 : PERFORMANCE SUMMARY

Highest level reached, Precision at Document Cutoff 30, TREC-2 - TREC-8

	V T a	S D a	M T+D a	L T+D+N a	L T+D+N m
>= 65					
>= 60				333	333 888
>= 55					222 777
>= 50				222	666
>= 45					444 555
>= 40	888	777 888	444 777 888	777 888	
>= 35	777				
>= 30	666			555 666	
>= 25		555 666			
>= 20					

Key: 222 = TREC-2 highest performance level, 333 = TREC-3 ditto, etc

(TREC-2 included Concept field
TREC-4 manual did not have Narrative field)

Overall comments

As before, but even more clearly when the evidence of TREC-8 is added in,

1. Many teams obtain similar performance, even at top levels.

2. Manual query formation can give superior performance to automatic, typically reflecting the amount of effort put in and/or user judgements on intermediate outputs.

3. There has been some convergence, especially in automatic searching, on default strategies; but similar performance is also obtained with very different strategies, presumably reflecting the dominating influence of the frequency data that strategies share.

4. Results in TREC generally illustrate the way in which established teams can maintain and enhance their performance; but it also shows that new teams can take advantage of published TREC experience and the rich training data that is available to get up to speed quickly.

5. Performance is broadly correlated with the quality of the topic information available and the difficulty of the topics.

6. However, as the results for TREC-7 and TREC-8 show, it is possible to do almost as well in automatic searching with the minimal (the Very short title) topics as with much longer ones.

7. The best levels of automatic search performance as illustrated by TREC-7 and TREC-8 are quite respectable, and in particular in many cases are achieved with relatively simple, albeit well-motivated, methods. It may be noted that at Cutoff 10, several teams achieved almost 50% Precision in automatic searching even with the Very short titles in TREC-8, and several reached more than 50% with the Medium length titles+descriptions. Manual searching without enormous effort can do better, achieving 70%, but the time and attention required is nevertheless not negligible.