SUMMARY PERFORMANCE COMPARISONS TREC-2 THROUGH TREC-7

Karen Sparck Jones
Computer Laboratory, University of Cambridge

December 22, 1998

The context

These comparisons continue my attempt to illustrate long-term performance trends in TREC. My last comparisons, for TREC-2 - 6, appeared as the final Appendix in the TREC-6 Proceedings. This year the tables are confined to adhoc performance, as routing has become less conspicuous in TREC. I have taken the opportunity to include a few minor corrections of earlier tables.

Over TREC as a whole there have been some important changes relating to a major variable, namely the topics (requests). First, their composition has changed; and second the conditions on their treatment for officially submitted runs have changed. Table 1 shows (a) the component fields in the topics in successive TRECs and (b) field lengths in successive TRECs. It also defines different topic *versions* as specified for different runs, ranging from Very short, titles only, to Long, covering title, description and narrative fields.

In earlier TRECs, up to TREC-4, automatic and manual *modes* of query formulation were treated simply as low-level optional alternatives, so in my earlier comparative tables I used the best performing run for each team regardless of mode. In TREC-5, and since, the modes have been treated as distinct conditions. At the same time there have been some changes, during TREC, in the definition of what is allowed in manual searching. These variations in test data and condition mean that overall trend comparisons can only be rather general. Because the search mode has become more important, runs for TREC-2 - 4 are now marked according to whether they were automatic or manual. For TREC-5 - 7 the modes are listed separately. Thus the main performance tables for these TRECs illustrate the pairings of topic version with search mode e.g. Very short with auto, Long with manual.

Table entries

The detailed figures for TREC-4 onwards are taken from the Conference Working Notes. They cover only Category A runs, and only higher levels of performance, not all 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 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).

TREC ADHOC SEARCH RESULTS FOR PRECISION AT DOCUMENT CUTOFF 30

KEY TO TABLE NOTATIONS :

a = fully automatic searches
m = manual searches
\- = manual searches in TREC 2 - 4

Topic fields available as base for queries :

	(TREC-1)	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7
T = title	x	x	x		x	x	x
D = description	x	x	x	x	x	x	x
N = narrative	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
T	3.8	4.9	6.5	-	3.8	2.7	2.5
D	17.9	18.7	22.3	16.3	15.7	20.4	14.3
N	64.5	78.8	74.6	-	63.2	65.3	40.8
C	21.2	28.5	-	-	-	-	-

TREC 2 - 4 did not distinguish queries by any specific sets of topic fields
TREC 5 - 7 distinguished runs by different sets of fields

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

	TREC-2 a/m	TREC-3 a/m	TREC-4 a/m	TREC-5 a S	TREC-5 a L	TREC-5 m L
>= 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			ETH
>= 40	...	...	-Berkeley -Clarit/CMU Cornell -GMU -UMass -InText -ANU			Waterloo
>= 35	...	...	City -GE/NYU			ANU Clarit Cornell GE/NYU GMUetc Lexis
>= 30	...	...	...		City CUNY ETH	OpenText CUNY Berkeley
>= 25	...	...	...	Apple City Cornell IBMTJW	Apple GE/NYU RMIT Berkeley	DCU IBM
>= 20	.....					

	TREC-6 a V	TREC-6 a S	TREC-6 a L	TREC-6 m L	TREC-7 a V	TREC-7 a S	TREC-7 a M	TREC-7 a L	TREC-7 m L
>=60	(best TREC 2-5)								
>=55									Clarit
>=50				Waterloo					ManInst Waterloo
>=45				Clarit					GMUetc
>=40				ANU		NEC	ATT Cityetc UMass	BBN Cityetc NEC UMass	ANU Harris Berkeley Toronto
>=35				GEetc Lexis	Cityetc	Cornell CUNY Fujitsu	Lexis RMIT	ANU Cornell CUNY TwentyO Iowa	GEetc Lexis IRIT
>=30	Apple ATT City IRIT Lexis CUNY Waterloo		ANU Cornell IRIT CUNY Berkeley	ISS Berkeley	ATT Cornell CUNY Fujitsu Lexis NEC NTTData RMIT Waterloo	IBMTJWs	IBMTJWg IRIT	GMUetc NTTData Rutgers Berkeley UNC	FS
>=25	DCU ISS	ATT ANU City Cornell GMUetc IBMTJWs IRIT Lexis Waterloo	City IBMTJWg MDS/RMIT UMass GMUetc	FS GMUetc	ANU Avignon GEetc IBMTJWg ETH Berkeley Maryland			FUB ImperC JHopk NSA	
>=20	MDS/RMIT Glasgow	Apple GEetc IBMTJWg MDS/RMIT CUNY Berkeley Maryland UMass Verity	Verity	Glasgow	FS GMUetc JHopk		Avignon ImperC MIT	NTHU	NTHU

Performance summary

Boiling down the larger tables for a summary picture of performance levels, I have taken the highest performance level reached for each version and mode over TREC-2 - 7 in the diagram below: the numbers refer to the corresponding TREC. This clearly shows high best levels of performance for TREC-2 and -3, and growing differences, for these respective top performers, between automatic and manual modes from TREC-4 onwards, generally reflecting less initial topic information along with more manual effort.

```
          V          S          M          L          L
          T          D        T+D      T+D+N      T+D+N
          a          a          a          a          m

>= 65

>= 60                                    3333333    3333333

>= 55                                               222/777 +

>= 50                                    2222222 +  6666666

>= 45                                               444/555 -

>= 40             7777777    444/777    7777777

>= 35  7777777

>= 30  6666666                           555/666

>= 25             555/666

>= 20

Key: 222 = TREC-2 highest performance level, 333 = TREC-3 ditto, etc
     + TREC-2 also had Concept field
     - TREC-4 did not have Narrative field
```

However it is important take the more detailed information of the main tables into consideration, as follows.

Overall comments

1. Many teams obtain similar performance, even at top levels.
2. Upper outliers are especially likely with manual mode, typically reflecting the amount of effort put into query development or user judgements on search output.
3. Though there has been some convergence on 'default' strategies, similar performance is obtained with very different strategies.
4. Performance trends over TREC clearly show the effects of data challenge, i.e. having less topic information or more difficult ('hard') topics: TREC-4 performance reflects the former, TREC-5 and -6 more the latter, since automatic performance is comparatively low regardless of query

version. Since TREC-7 full topics are shorter than TREC-6, but TREC-7 performance levels are better, the TREC-7 topics are presumably less hard.

5. However performance is not as tightly correlated with topic length, and specifically with version, as might be expected (setting aside the known-problematic TREC-6 descriptions). Thus similar good performance is obtained in automatic mode for different versions.

6. TREC-7 shows respectable absolute levels of automatic mode performance for intermediate length topics (Short and Medium), interestingly as good as for (the unrealistic) Long version; they are also comparable with all but the best manual mode. Performance with Very short is less good, but not negligible. Taking all other factors into consideration, these reasonable levels of performance with shorter versions of the topics must be primarily attributed, over TREC as a whole, to improvements in automatic mode methods.