APPENDIX C

COMPARISON BETWEEN TREC2 and TREC3

Karen Sparck Jones, November 15, 1994

The following table presents Precision performance at Document Cutoff 30 for both Trec 2 and Trec 3. The Trec 2 figures are taken from KSJ, ‘Reflections on TREC’ (prepared for the issue of Information Processing and Management on TREC, also published as Computer Laboratory TR 347, 1994). The Trec 3 figures are from the Trec 3 conference working papers. The data are only for full runs, not Category B.

The conventions are as for Trec 2: 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. However I have indicated the Trec 3 interactive figures separately, as a different evaluation methodology was applied.

DOCUMENTS 30

	AD HOC		ROUTING		INTERACTIVE (ROUTING)
	Trec 2	Trec 3	Trec 2	Trec 3	Trec 3
>= 60		Amherst City Berkeley	Cornell Dortmund	City	
>= 55	Amherst HNC VT	Cornell Mead	City Berkeley Amherst Bellcore CMU	Amherst Cornell Berkeley Dortmund Bellcore CUNY	
>= 50	Cornell Berkeley Dortmund CMU Verity Siemens CUNY	VT Westlaw ETH CUNY	Rutgers HNC GE TRW Verity Siemens	CMU Westlaw Logicon TRW Florida	
>= 45	City Bellcore ETH CITRI Conquest	NYU CMU RMIT RutgersK	VT	Xerox NYU Verity ETH NSA NEC	RutgersB Verity
>= 40					City

Notes:

1) Changed players - Trec 2 only: GE, Conquest
Trec 3 only: Westlaw, Logicon, Xerox, NSA, NEC
also RMIT, Trec 3, was part of the CITRI group, Trec 2
Rutgers had two groups, RutgersB(elkin) and RutgersK(antor)

2) P >= 45 was not shown for Trec 2 Routing in ‘R on T’: it is included here both to cover new players and for comparison with interactive

3) I do not include Minnesota as I do not know which of their many runs to use

COMMENTS:

1) Ad hoc performance shows a slight rise of the top level from Trec 2 to Trec 3.

2) Routing top performance is similar for Trec 2 and Trec 3.

3) Both ad hoc and routing show a slight drift up over classes, but rather more for ad hoc than for routing: with ad hoc it applies overall, with routing really only for lower blocks.

4) Ad hoc performance for Trec 3 is comparable with routing for Trec 3.

5) On the whole the better performing teams for Trec 3 are those for Trec 2: the odd downward movement (where not a goof) probably reflects a willingness to experiment.

6) Some teams that did poorly for Trec 2 (below the bottom of the table) got it together enough for Trec 3 to be with the main pack.

7) Some new players performed respectably (though one, NSA, perhaps less well than for comfort).

8) One team, NYU, showed ability to do NLP on the full scale, not just in group B.

9) Interactive performance was absolutely poor.

OVERALL REMARKS:

1) The General Findings for Trec 2 given in 'R on T' continue to apply in Trec 3.

2) Many different approaches give similar performance.

3) Good performance is obtained with variants of the statistical approach souped up with phrases etc.

4) Routing performance seems to have reached a plateau.

5) The improvement in adhoc performance must be attributed to the stability of the queries (or query types) and the large volumes of training data, as well as the honed character (even after concept removal) of the queries.

6) (Setting aside any evaluation methodology issues) the poor interactive performance must be attributed to the fact that the queries were already well worked up, and mainstream ad hoc and routing performance is good because of quality queries and lots of training data.

7) ALL of these points are broad brush ones: there may be significant differences within performance blocks, and also none between members of adjoining blocks. Without accepting the degree of similarity between different runs that Jean tague's Scheffe tests show, and 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.

8) IN PARTICULAR, all of Trec 2 and Trec 3 is within a very specific retrieval context; and as many of the successful systems are complex overlays on simple bases, it is essential to get a better view of how the Trec IR environment is influencing the settings for the many parameters that are often involved in these systems.