Further reflections on TREC

Karen Sparck Jones
Computer Laboratory, University of Cambridge
Cambridge, England

December 1998

Abstract

The paper reviews the TREC Programme up to TREC-6 (1997), considering the test results, the substantive findings for IR that follow, and the lessons TREC offers for IR evaluation. The paper focusses on the adhoc retrieval task, with discussion of other test tracks as appropriate. The paper summarises the structure of the TREC work and analyses the experimental data in some detail. The analysis of the tests is presented through a series of key questions about indexing models, document and query descriptions, search strategies, etc. The assessment confirms that statistically-based methods perform as well as any, and that the nature and treatment of the user's request is by far the dominant factor in performance. One implication is that TREC should move into a new phase targeted on key comparisons and task specifications designed to deliver substantive new information, in particular shifting towards situated IR that addresses the user's context and contribution to searching.

1   Introduction

The Text REtrieval Conference (TREC) Programme, now in its seventh year, has established itself as the IR community's major evaluation exercise, involving many teams in many countries in a series of related tasks and tests. In "Reflections on TREC", written after TREC-2 (Sparck Jones, 1995), I tried to determine what TREC had shown when, after the 'debugging' effort of TREC-1, the first solid tests had been done.

Since then, TREC has developed in many ways, all significant for IR research. First, the number of participants, and hence strategies investigated, has increased. Second, the main task, *Adhoc* (one-off) retrieval, has been subjected to variations designed to make it more realistic and to check whether results can be consolidated. Third, many subsidiary *tracks* have branched off from the main line into other areas, with different data, tasks, or methodologies. Collectively, the tracks now represent a major part of the TREC enterprise, since they serve on the one hand to reinforce the main stream findings, and on the other to extend the range of TREC findings.

One problem, indeed, is the sheer volume of publication and mass of local detail, reporting on a vast amount of work but making it difficult to see the wood for the trees. This fine grain is nevertheless a useful reminder of the many *environment variables* and *system parameters* that affect IR, and of the fact that conclusions about the merits of retrieval strategies have to be thoroughly grounded. This paper, like its predecessor, is thus an attempt to find the main trees in the wood after TREC-6, or at least some of them, and also to see what sort of retrieval house one can build from them. Though TREC has done a great deal, there are large and important matters that TREC has not yet addressed. TREC has confirmed general conclusions to be drawn from many earlier tests, for instance that statistically-based indexing and searching techniques are cheap and competitive, and in particular has done this on a large collection scale. But the TREC work has been carried out within the traditional 'abstract' laboratory paradigm that is hard to

relate, as much for modern Web browsing as for old-fashioned libraries, to users trying to find their way on the ground. While TREC requests and relevance assessments are (fairly) real, and some investigations within TREC have been on live human searching, nearly all of TREC has been cut off, albeit in the very valid cause of controlled evaluation, from genuine, hands-on users. Even where manual query formation is involved, this has often been motivated more by an interest in establishing best performance levels than in capturing normal user or intermediary effort.

1.1 Paper coverage

The TREC tracks are so important in themselves, and cover so many distinct problem areas, that they require full treatment in their own right, and it is hoped they will get this in a separate issue of this journal. I shall therefore focus in this paper on the main line of TREC, Adhoc retrieval, and will consider the tracks only in relation to this and as buttresses for the TREC evaluation effort as a whole, rather than as independent areas of investigation in their own right. I shall in particular not consider track detail except where it is especially appropriate. (I use "Adhoc" rather than "adhoc" to refer to this task as specified within TREC; similarly for "Routing".)

I take as background the detailed characterisations of the data, task specifications, evaluation paradigm, and team participation, in the successive TREC Proceedings (TREC-1 - TREC-6, 1992-in press) as well as, of course, the actual results obtained and descriptions of the strategies that delivered them to be found there. I shall therefore only consider, in the next section, the main facts about the data etc, that are relevant to my analysis of TREC here. I shall also take advantage of the Overview papers by Donna Harman and Ellen Voorhees in the Proceedings volumes, which make significant comparative and general points about the approaches studied and results obtained in the individual conferences. Note that information about TREC-6 is taken from the Conference Working Notes, before the participants' full papers were available, and may be subject to correction.

My analysis of TREC-1 and TREC-2 in "Reflections on TREC" - hereinafter "Reflections", and in particular of the test results, was based on a series of questions: about indexing and retrieval models, indexing vocabulary, document descriptions and indexing sources, queries and query sources, search strategies, match scoring criteria, output forms, and the role of learning. I shall use these questions here, to throw light on what the TRECs after TREC-2 have shown, but also add some new questions as justified by the longer time span and increased range of TREC work. Maintaining this continuity is not just presentational convenience: the questions are about important issues, and one of the major points about TREC is whether the vast test effort is effectively consolidating older results as well as usefully advancing on new problems,

At the same time it is necessary to ask, after six TREC cycles, whether TREC is (or should be) addressing IR issues that modern developments, notably those associated with the Web, are raising. After considering the detailed results in the previous paper I offered an assessment of TREC as a whole, and I shall do the same in this paper covering, as previously, both the substance of what has been learnt about retrieval and matters of evaluation methodology. This paper is therefore organised as follows. In Section 2 I give facts about TREC, concentrating on the tracks and the data as major features of the whole enterprise; in Section 3 I present the analysis of results, from different points of view; and in Section 4 I summarise and assess the performance findings and methodology the lessons of the TREC Programme so far, leading to my conclusions on where TREC should go next.

2 TREC facts

Note: Counts may vary slightly according to the precise source of the information used (e.g. Proceedings overview papers or results table sets) and date (e.g. before or after some files were cleaned up and purged of duplicate documents, queries without relevant documents, etc): most of my figures are taken from the official results tables.

2.1 Organisation

TREC is jointly sponsored by DARPA and NIST, and is managed by NIST under Donna Harman and Ellen Voorhees' leadership, supported by NIST staff and a standing Programme Committee drawn from the participants. With only a few, contingent exceptions, teams are not funded to participate by these agencies. The important consequences of this organisational structure for TREC as an evaluation enterprise are as follows:

1. The TREC philosophy, which combines a search for viable technology with a desire for analytic understanding and a commitment to 'kiss and tell' on what was done.

2. The TREC strategy, which combines a get-up-and-go approach with experimental controls in task specifications, and offers both user-friendly and formally-grounded evaluation measures.

3. The TREC materials, which combine extensive data sets with intensive detail.

4. The TREC activities, which combine extensive attacks on novel problems with intensive study of familiar ones.

TREC's sponsored data provision, which allows a scale in document and request sets and, most critically, in relevance assessments, is especially significant for its present and long-term future value to the IR community. Virtually all of the data is now publicly available (see http://trec.nist.gov/data.html). This is moreover an accumulating resource since data sets used in successive TREC cycles can, in some cases, be legitimately combined. More importantly, TREC's funded commitment to publishing its voluminous detailed results is also building a major resource for future reference, analysis, and comparative performance test, with the added convenience of public electronic access.

2.2 Participants

Participation in TREC is open, to anyone satisfying modest formal constraints and able to muster enough effort primarily to take part in the Adhoc retrieval test. Though not all who enter actually succeed in submitting runs, the number of participating teams, defined here as those which complete, from TREC-2 onwards has been high, as shown in Table 1. The main Adhoc task is not obligatory, but it is strongly encouraged and indeed is likely, through its manifest centrality, to attract participants. Further, while new teams were permitted before TREC-6 to test with a reduced data set (Category B), or for reasons of commercial confidentiality have a somewhat shadowy existence in Category C, most teams have fallen into Category A, using the full Adhoc data. This is important because the generic results and findings, to which reviews such as this necessarily refer, are based on a very large number of individual runs, representing many different approaches to indexing and searching and individual system specifications, that are open to detailed inspection. The fact that participation is international, and by both academic and commercial organisations, shows that the TREC work is seen as solid, and pertinent to operational retrieval. Further, many teams have participated more or less continuously throughout TREC, which has served both to validate their own particular ideas and to reinforce the overall trend in choice of methods and pattern of results.

Table 1 shows the number of participating, i.e. completing, teams from TREC-1 to TREC-6, for full Category A Adhoc, all Adhoc, all Adhoc plus teams participating in the Routing task only, and all of these plus any participating only in one or more tracks. Overall, the numbers have been rising. These figures alone do not altogether reflect the level of interest in TREC, since many sign up who are in the event unable to complete the taxing minimum amount of work required to participate at all. More importantly, they do not reflect the research contribution made, since with the development of the tracks since TREC-4, many teams undertake several (in a few cases all) the track tasks; and while for some tracks this is not significant extra work (e.g. the High Precision track), in other cases it is very substantial (e.g. the Multilingual tracks, or Interactive). The numbers of teams taking part in the Routing task, and in the other tracks since they were

introduced in TREC-4, are given in Table 2. This shows both that in many cases the range of comparisons per track task is fairly large, and also that the track results taken together do much to extend the results coverage of TREC as a whole. (I am grateful to Dawn Tice of NIST for these figures.)

It should be noted that over TREC-1 - 6 there have been about 25 non-North American participants, with the number rising in later TRECs, while there have been 70-80 participants overall.

Thus it must be emphasised (as earlier with "Reflections") that the number of participants that have been involved in TREC is so large that, except where specifically stated, when individual teams or their results are mentioned in this paper, they should be taken as representative of groups rather than unique.

2.3   Design

TREC began with two tasks, Adhoc retrieval and Routing, and with the aim of identifying the best way of doing these through comparative experiments across different strategies, administered in a laboratory environment and with performance characterised using conventional measures based on Recall and Precision.

As noted in "Reflections", automation dominates in the sense that file inspection is automated and document indexing, where this is prior to and hence independent of request indexing, is also automated. Manual processing can figure in two ways: less importantly in the development of such search aids like thesauri, and much more importantly in the development of the actual search *query* from the user's initial *request*, whether before or during searching. For convenience I shall use the term *system* to refer to any combination of indexing or searching *strategies*, which in turn exploit a number of *devices*, so systems will differ at least at the device level. For instance, if we respectively (a) auto index requests on non-stop words, weight, search, apply Y/N relevance judgements to top output, reweight, and search again; and (b) manually select content words, weight ... as in (a); we define two systems, since the difference in initial query term choice is a device-level one. Devices may be regarded as system parameters, but in discussing systems I shall stay above fine parameter definition. As the examples indicate, I shall use the term "system" even where there is non-trivial and possibly quite intensive manual processing, for example in building a search query using a range of aids, i.e. in an extended sense. I shall use *strategy* rather broadly, to refer to what is strictly a type of strategy, e.g. statistical weighting, or the use of compound index terms versus simple ones, and devices similarly, for instance to refer to the *tf*idf* type of weight or to compounds determined by syntactic rather than purely locational methods. Parameter definition and tuning then deals with e.g. the precise form of a generic weighting function and choice of setting for a component constant. One important goal for TREC participants has been to establish reliable formulations for generic devices, rather than detailed parameter definitions. Finally, I shall use *approach* as an overall characterisation, as in "the statistical approach".

TREC began with the aim of determining, by rigorous comparative testing, what the best general system specifications were, as defined by strategy and device choices, and with an interest in as full automation as possible. TREC was primarily a *black box* evaluation, with systems compared as wholes for performance through their team run results, though of course individual teams could engage in glass box comparisons at the particular strategy, device or even parameter level.

Each TREC cycle would however, limit the system comparisons for each task to a single fixed environment, as defined by its sets of documents, requests and needs, and would impose consistency on the system performance evaluation by applying particular measures that were chosen to characterise core facts relative to Recall and Precision in a set of complementary ways.

Further concentration was achieved by limiting the number of permitted *official runs* per team submission to two and also by the mechanics of each evaluation cycle and the tight timetable. Thus following a general DARPA paradigm applied in other evaluation programmes, participants were able to work with *training data*, i.e. full test collections similar to those to be used for test, and were then required to apply their strategies to *test data* consisting of documents and requests

without relevance assessments, within a short time period, and to deliver output up to a size limit for assessment at NIST.

Overall, therefore, if evaluation is seen as populating a *test grid* with environment and system axes, TREC supported comparisons on the system axis from the beginning, with more gross ones typically between teams and more refined ones within a team's run set. TREC-1 also added a major change on the environment axis, when set against previous IR evaluations, through using much larger document files and also full text ones: thus TREC became a new 'meta'-environment for testing all the generic approaches and specific system instantiations hitherto tried out only with much smaller data. Then from TREC-2 onwards, there were changes to the environment in each cycle, giving different values to the major variables, most notably through new request sets but also with new documents of different form or subject types.

2.4   Tasks

As noted, TREC began with two tasks, Adhoc retrieval and Routing. The former represents the most common literature retrieval task, searching for documents on some topic on some particular occasion. *Routing* notionally represented the familiar selective dissemination of information (SDI), or filtering, task, though it was rather awkwardly instantiated in TREC as a two-stage adhoc operation, with query learning on accumulated training data followed by query application to a new data set.

Adhoc searching has retained its primary status, both intellectually and as an officially encouraged task, throughout TREC. Its importance as a retrieval task makes it significant for evaluation, whether directed towards immediate system assessment or eventual system implementation. This status is enhanced by its traditional research role as the arena for comparison between very different approaches and strategies, a matter of concern for sponsors and participants alike. Its central significance is further marked by its function as the source of relevance information. TREC necessarily, because of file size, relies on a *pool* basis for Recall calculation, and it is essential for both immediate and future evaluation to ensure the pool is as good as possible. The more, and the more different, systems contribute to the pool the better it will be, in coverage and lack of bias. Both sponsors and participants thus have an interest in maximising the number of contributors to the pool. The pool has been formed from the top 100 ranked documents from each official run. In studying the TREC pools, Harman (1996) and Voorhees and Harman (1997) have found that variety in the strategies contributing to the pool has been very important, and also that pool coverage of the 'true' relevance set is probably good. Voorhees (1998) also shows that systems can be reliably compared using the TREC data, although relevance assessments vary. However Zobel's (1998) study suggests that while a limited pool allows reasonable system comparisons, using a larger measurement than contribution pool (1000 vs 100) presents problems, and that TREC recall coverage may be lower than assumed.

2.5   Tracks

After TREC-2, however, it became apparent there were important issues not addressed with Adhoc evaluation as it had been conducted, and there was increasing dissatisfaction with the way the Routing specification was distorting the filtering task, which should properly be treated as a stream-based document allocation, not a trained-up version of the Adhoc task. So after some modest initial trials in TREC-3, alternative track tasks were instituted in TREC-4. As Table 2 suggests, these have by now become a major source of research initiative and participant effort in TREC. Informally, TREC as a whole can be seen as a wheel, with a central hub (the Adhoc task) and radiating spokes (tracks) linked both to the hub and round the rim to one another. New spokes can be inserted, existing ones strengthened, or worn-out ones removed. We can thus look for a continuously improved TREC IR wheel, and indeed not just with more and better spokes, but with hub and rim redesign and reinforcement as well.

In practice, track growth has been limited by logistic and workload constraints, especially in the key area of relevance assessment. But this had probably had more good effects than bad,

since even though it has sometimes made it impossible to implement optimal track designs (as in providing continuous assessment for the filtering task), it has promoted care and consolidation: care in the specification of the tracks, and consolidation through gradual development from year to year and through the ranges of comparisons that come from good participation rates.

Altogether, as shown in Table 2, TREC-4 - 6 have covered 11 designated tracks, 12 if the tracks are taken to include Routing, which since TREC-4 has been effectively treated as a track and terminated at TREC-6. The tracks have included some extremely taxing ones e.g. Cross Language, and several with many participants. As noted, the tracks are now a major element of the TREC Programme as a whole, and a significant extension in the research area that TREC covers. The misleadingly labelled Multilingual tracks are for experiments with retrieval in languages other than English; the Interactive track addresses aspects of searching not studied within the main Adhoc task; Database Merge is concerned with the treatment of several distinct document files; Confusion is about corrupted data, as with OCR material; Filtering is the 'real' form of the Routing task; Natural Language Processing examines the value of indexing making explicit use of syntactic and semantic interpreters; the Cross Language track is for systems covering several languages at once, allowing requests in one language and document output in others; High Precision is a study of strategies focussed on Precision performance; Spoken Document Retrieval (SDR) addresses adhoc retrieval for speech files; and the Very Large Corpus (VLC) track is tackling document files much larger than those used for the main Adhoc task.

As the Table shows, some tracks have already passed the limit of their useful life: for instance Multilingual Spanish can reasonably be deemed done in demonstrating that techniques proven for English can be carried over in a straightforward way, and quite possibly with less effort than expected, to (at least some) other languages. Some tracks have in the event appeared better suited to other communities, for example the OCR one. Others, notably Database Merge and Natural Language Processing (NLP), have not caught on, whether because suitable data is not available or because the track specification has not proved attractive: thus for NLP, participants have to have uncommon processors, and be convinced that, in spite of considerable evidence to the contrary, NLP really has something to offer IR. Routing has been replaced by Filtering for TREC-7 because, though Routing was important for the early TRECs, especially in illustrating the role and limits of training data, it does not capture the key issues for continuous information dissemination.

Track types

As the foregoing suggests, the tracks can be assigned to classes defined by different purposes, though this is a rather informal characterisation and individual tracks may fall into more than one class.

The first class deals with more *realistic* forms of tasks already under study in TREC. This applies to the Filtering track, the Interactive one to some extent, and the VLC one. In the Filtering track, as just mentioned, continuous document distribution replaces the inappropriate batch search with ranked output of the earlier Routing track. The Interactive track attempts to overcome the limitations of the Adhoc specification, which fails to engage users directly in searching for their own needs, and to address styles of document presentation or forms of interactive display as supports for the user's search. However though the Interactive track has addressed user behaviour in searching, it has done this in specialised ways and with many elements of the user's interactive environment omitted, so the track falls more naturally into a later class. More generally, while the Filtering and Interactive tracks introduce more realism than in their predecessor Routing and Adhoc tasks, these track studies are still limited and artificial laboratory ones. The VLC track is addressing realism along a different dimension, namely scaling up to the ever-larger document set sizes now familiar in operational situations.

The second class covers tracks that can be described as *variants* of basic Adhoc. For instance, the performance measures used for Adhoc are neutral with respect to users' interests in Recall or Precision; they indicate a range of relationships between the two, but do not encourage work on strategies specifically designed to respond to a user's bias, say towards Precision. The family of utility measures adopted for the Filtering task, on the other hand, does allow for such user

preferences. The High Precision track is deliberately designed to test systems' responses to users' preferences for high Precision output, and is thus reflecting the widespread perception that the users of modern systems are far more interested in Precision than Recall, especially with large files where there can be many non-relevant documents that match queries quite well. For practical reasons, to reduce assessment effort, the Confusion and SDR tracks have used known-item searching, so these tracks can also be seen as conducting evaluations of system performance for a legitimate variation on the Adhoc task. They have not, however, really concentrated on this task in its own right, and SDR in TREC-7 has returned to ordinary topic searches.

The third track class is *focussed* on aspects or components of retrieval systems. The Interactive track is best seen in this light, since as the detailed specification makes clear, it involves careful logging and study of what users are doing in developing their queries, for instance adding or deleting terms, inspecting particular documents, etc. The other focussed tracks are Database Merge and NLP: the former was intended to study strategies for situations where there are multiple distinct databases (not just a distributed single database), as is often the case with bibiographic search services, and thus where it may be sensible to give priority to some bases over others for individual queries. The NLP track sought to isolate the effects of indexing that could only be achieved through complex, linguistically-based document and query analysis. It should be noted, however, that though these tracks can be conveniently grouped as focussed, they differ in other ways: thus Database Merge concerns a situation while NLP concerns methods.

The fourth, and largest, track class can be labelled *extensions* of the Adhoc task to other kinds of data than the 'normal' (though far from uniform) TREC document file. One such extension, as in the Confusion track, is that where the file data, whether at the individual document or whole file level, is corrupted, as would be likely with OCR material. The second, somewhat similar, case is SDR, where it cannot be assumed that the speech recogniser will (actually or logically) deliver wholly accurate transcriptions of the original speech. This material also often differs from ordinary TREC data in genre, since it may include dialogues, but this has hitherto been seen as a secondary factor in relation to retrieval performance. This class also includes extensions of the Adhoc task to other languages, rather than to other manifestations of English. The extension to Multilingual Spanish is relatively modest, but to Chinese is more demanding in covering a radically different script system. It therefore clearly raises the question of whether in practice, as opposed to broad principle, indexing and searching methods developed for English, e.g. those using statistical techniques, are readily transferred to and are effective for other, very different languages.

One important concomitant of the tracks has been their collective influence on the range of data used in TREC. Insofar as participants apply the same, or very similar, techniques in different tracks, they are extending the variation in environments under which their systems are tested. This applies when the same system is used for different tasks, as in Adhoc and Filtering, or more broadly when the same general approach, though with some detailed modification, is used for different purposes. Again, when different performance measures are used, these can be seen as an indirect reflection of different system environments, since they in principle reflect distinct user needs and system functions. However these environment variations in TREC have not been applied coherently or rigorously, and are rather the byproducts of separate track specifications. This aspect of TREC is considered further later.

It is further useful, though in TREC parlance (e.g. Voorhees and Harman, 1997) all the tracks have their particular specified "task", to distinguish tracks where the study aim and design necessarily imposes some 'unnatural' experimental constraints on the task as a retrieval task, as in the Interactive track, from tracks where the task is less tightly controlled, as essentially in all the others. This is a distinction different from that between intrinsically natural tasks, like that with which the Interactive track is concerned, and artificial ones, like Routing as defined in TREC. We can also distinguish those tracks where the task necessarily differs from the Adhoc one, as with Filtering, from tracks where a different task has been used for practical reasons, as with the known-item searching adopted for the Confusion and TREC-6 SDR tracks. These distinctions, recondite though they may seem, and the fact that the TREC tracks fall into different classes, are important when assessing TREC results overall.

2.6 Data

Overview

While the data details for TREC are given (though not always very fully or clearly) in the TREC Proceedings, it is helpful to note the salient points about the data sets used.

Since one of the main TREC objectives has been to test retrieval strategies on a large scale, the Adhoc task in individual TREC cycles has used very substantial document files, typically with more than half a million items. The combined files from TREC-1 - TREC-6 amount to some 1.63 M documents. The material is made up of various blocks of documents from different sources with rather different subject, size and genre characteristics. The data sets used in the cycles differ sufficiently from one another to support claims of generality for technologies and performance, though since they also overlap they indirectly allow a kind of retrospective validation for the new queries in each cycle. However while some of the data types vary in subject, notably the news material, there is nothing like the document range to be found in e.g. AltaVista, and some familar types of document, notably scientific articles, are poorly represented, especially in the numbers that are challenges for the major bibliographic search services.

The Adhoc requests, i.e. *topics* in TREC parlance, are normally used in batches of 50. Their characteristics have changed, and as these are important in assessing the TREC search results, they are considered in detail below. They have, however, the important property that they include a specific statement on what makes a document relevant (the 'Narrative' field), which makes it possible to assess documents for relevance over a long period of time, and adds to the general resource value of the data for future research.

The relevance assessment sets are very substantial, typically providing tens or even hundreds of relevant documents per topic, and are a major factor in making the TREC Collection (or rather family of collections) an extremely valuable data resource. As noted earlier, the search output sets used for these assessments have been subjected to study at NIST to establish pool coverage and judgement consistency.

The Routing task has mainly used subsets of the same document and topic files as the Adhoc task. The tracks have in some cases also used subsets, in others completely new material. The Very Large Corpus track in TREC-6 has enlarged the overall TREC document Collection substantially.

Documents

Table 3 and Table 4 give details of the document sources and collections used to make up Adhoc test data, and from which Routing sets have also been drawn. The Adhoc and Routing data have been gathered onto a series of discs, D1 - D5, and in accounts of TREC it is common to refer to the document sets used by Disc names. The first part of Table 4 gives this distribution; the second includes for reference information about the further documents that were used together with D1 - D5 for the TREC-6 VLC track, namely DAT1 - 4. By *source* I mean e.g. AP, Associated Press, or FR, Federal Register. Material from a given source may be divided to form more than one specific *collection*, as is the case with AP material, or with NEWS, which is sometimes separated by Disc, or may be formed as separate files in a single repository, as with NEWS01 - 04. Table 3 also groups the sources by generic type, and gives the numbers of documents of each type. Table 4 shows the size of the individual collections and repositories in both document numbers and gigabytes. Finally, Table 5 shows which of the TREC Discs were used in the individual Adhoc and Routing cycles. (The figures are taken from the VLC reference sources for TREC-6: Hawkins and Thistlethwaite, 1998, and the track WWW page).

Thus as these Tables imply, TREC-1 - 3 used about 750 K documents for Adhoc testing, TREC-4 - 6 about 550 K. TREC-1 - 3 used the same document sets, so it was thought necessary to vary the data in TREC-4 - 6.

Overall, it is evident that the TREC experiments have been carried out on, and across, substantial and varied document sets. However as Table 3 makes clear, news material (from wire service or newspaper sources) makes up more than half the whole, and while this covers an enormous range of topics, it is only one genre of many; in particular, most of the documents are quite

short and 'straightforward', though they are tricky from the proper name and abbreviation point of view. The next largest category, ZIFF, are in a single broad subject area, computing, but varying from regular scientific articles to popular and more miscellaneous though still technical items. DOE, Department of Energy, are predominantly technical material, over a wide subject range, but are only abstracts. FR, Federal Register, are official publications, not necessarily technical. CR, Congressional Record, is parliamentary proceedings, on a variety of topics and again a different genre. PATENTS have their own varied technical vocabulary and style.

The TREC experiments clearly demonstrate that the systems tested can cope with heterogeneous and bulky files, although ones that have been got together as could best be done, without the luxury of planned coverage. At the same time, the difficulties of obtaining data and unavoidable practical constraints on the supply of relevance assessors mean that the TREC files are dominated by news material, and do not contain large amounts of the kind of conventional scientific material covered by operational services on the very large scale (as with medical literature) that is widely regarded as presenting the real challenge for retrieval: this leaves open the question of how well the systems tested would perform if presented with vast files of technical papers. Past experience, e.g. with Medlars, and good performance with the TREC ZIFF and DOE subfiles suggest that systems might perform better with such technical material than with TREC in general.

It is also noteworthy that the VLC track materials include USENET NEWS, which offers the kind of document data characteristic of the Web, and thus can constitute a useful test set for future studies of strategies oriented to Web searching.

The Routing task has required a rigid separation of training and test data. In TREC-1 - 3, as Table 5 suggests, the document and topic sets were related in a semi-patterned manner, with new topics and old documents for Adhoc and old topics and new documents for Routing, accumulating in a sort of backward merge into a specific training set for Routing as well as general-purpose experimental utility. The more restricted document sets used for Routing in TREC-4 - 6 were primarily due to the difficulties of getting new documents, but were also stimulated by a desire to study Routing in a more discriminating manner, since the earlier TRECs had shown that with very different as well as carefully specified topics, and very large training sets, a high performance level could be achieved.

The tracks, apart from VLC, have in some cases used the ordinary current-TREC Adhoc or Routing file, in some used subsets taken from main data and chosen for their propriety for the track focus, and in some necessarily brought in new document collections. In general, the tracks have used smaller document files than the main Adhoc tests, though these have still been respectable in size. The details are given in Table 6. As the table shows, where tracks have not used Adhoc or Routing data, they have typically had quite small files, and new data files have often been news wire or paper material: VLC is the striking exception, though it has still included documents of this kind.

Overall, while some TREC experiments have used specific document types, there has been little controlled experiment on the effect on performance of systematic variation in the document type environment variable, especially when taken in conjunction with variation in e.g. request length. TREC has tended to focus on system/strategy robustness and consistent relative performance in system/strategy comparisons across broad data changes. How far TREC should adhere to a grid-style experimental paradigm, whether to pursue information about system behaviour or to guide system adaptation to different situations, is an important general topic to which I shall return in my conclusion.

Topics

The TREC topics (i.e. original 'user' requests) are the most important factor in the TREC tests. Their characteristics have changed over the TREC Programme, with significant consequences for the test results and the inferences to be drawn from them.

For the Adhoc experiments, as already noted, they have been used in batches of 50. This is not as large as is really desirable, since it is impossible to select, from any one batch, subsets with different properties for comparative study that are themselves individually large enough for

confidence in test results with them. Larger sets can be constructed by amalgamating comparable batches, but this precludes direct comparisons with published TREC results. Each batch does, however, cover a range of topics and needs, and invokes relevant documents spread round the document file.

The TREC-1 and 2 topics (as illustrated in "Reflections") were carefully formulated and elaborately structured, reflecting their origin in an SDI environment and also the requirement that they incorporate a statement of relevance document criteria (the 'Narrative') to permit relevance assessments at a much later time or by others than the original request writer. Even though TREC-1 was not a fully debugged evaluation, it was already evident by TREC-2 that the quality of the topics was so high that it was possible to reach a very good level of performance, and to do this without unduly demanding indexing and searching strategies. In particular, while it was reasonable to work with such careful profiles for routing, the topics were felt to be unrealistically refined as inputs for adhoc searching, even when viewed as illustrating the old-fashioned model with substantial query development before any searching; and this model was in any case no longer really applicable. From TREC-3 onwards, therefore, the topics have been prepared with an eye to greater realism, while recognising diagnostic interests in laboratory evaluation.

As shown in Table 7, the original topics had 4 fields, Title, Description, Narrative, and Concepts. In TREC-3 the Concepts - key notions, especially ones expressed by phrases - were dropped, as having made life far too easy. However the Narrative was retained not only for assessment purposes but because it was thought to reflect information the user would give an intermediary searching on their behalf or apply in their own query formulation. In TREC-4, with yet more realism as an objective (especially in the light of observation of operational systems, notably Web engines, where opening requests are typically extremely brief), the Narrative was provided only for the assessors, the Descriptions were much shorter, and Titles were dropped. This had a striking impact on performance, but also inhibited study of the effects of varying amounts of request information on performance, and could be criticised as limiting, in an unrealistic way, the amount of user information on which manual query formation could normally rely. Thus in TREC-5 the topic structure was as for TREC-3, but with shorter topics; and TREC-6 was as TREC-5.

But in assessing the impact of the TREC topic properties on retrieval performance, it is also necessary to take into account the official task specifications, governing the conditions for submitted runs, in the various TREC cycles. These were always designed to allow manual as well as automatic searching, both so as to be able to compare these and to encourage commercial systems employing manual searching, and also to allow search system feedback, to study the impact of document assessment (even if this was by members of the participating teams, rather than the official assessors). Thus in TREC-1 - 3, as Table 7 shows, runs were labelled *automatic* or *manual* or either with *feedback*; and both automatic and manual searching could use the whole topic. In TREC-4, in keeping with the general austerity of the Adhoc design, and also because there was a separate Interactive track, the only options were automatic or manual with no document assessment allowed for either (though 'blind' feedback where top ranking output from a first search is assumed relevant is always allowed).

However in TREC-5, when the topics became fuller again, the alternative specifications for query base and formulation became more complex, in an endeavour to separate automatic and manual modalities and to cover the effects of having more or less topic information. The design recognised that allowing manual searching without feedback was somewhat artificial, while the detailed study of interaction had been shifted to the Interactive track. The options allowed were therefore automatic query formation using only the Description field (obigatory), optional automatic on the full topic, and manual on the full topic with feedback if desired.

The TREC-6 design was as TREC-5; however changes to the way the topics were built unintentionally introduced a new variation: the topic builders often used the Titles not just to label the topics but to give topic cores in key terms not necessarily duplicated in the Descriptions. So those experimenting with these as 'Very short' topics analogous to typical initial user requests obtained good performance, while those constrained to the 'Short' topics defined by the Descriptions did no better. A better specification structure would have been Title alone as 'Very short' and Title+Description as 'Short', along with the whole topic including Narrative as 'Long'. TREC-

7 is handling this better, since the Description includes the Title terms and is thus effectively Title+Description.

These points may seem over-refined, and they do not necessarily influence the gross findings as to generally effective strategies. But they make it impossible to base more detailed conclusions as firmly over a coherent *series* of tests as is desirable.

The other important point about the provision of topics over the TREC Programme has been more care in ensuring that assessments are normally made by topic originators, which did not hold initially, that more different people are involved, and that they provide topics in a manner closer to genuine adhoc users, rather than as professionals replicating their own professional work. These points are discussed in detail in the TREC overview papers, e.g. Voorhees and Harman (1997): the issue of how far the TREC 'user requests' are typical of other than professional inputs is a matter of concern in assessing the predictive value of the Programme results.

Many of the track experiments have used ordinary TREC topics, as shown in Table 6, sometimes in sets with consistent properties, sometimes drawn from batches with different properties, but normally (though not always) with the T+D+N structure. Where topics have been specially provided, they have ranged from simple sentences for known-item searching (as in the Confusion and TREC-6 SDR tracks), to straightforward versions of the normal three-part topic, as in TREC-6 Cross Language. However the topic sets for several tracks have been small, e.g. a set of 25, and in some cases, notably Interactive, very small, though it has been argued that the particular test design for this track, with multiple searches on the same topic, ensures the results are valid.

Relevance assessments

For the main files, these have typically been done by the topic providers, but in some cases, e.g. VLC in TREC-6, have been done by others. Voorhees and Harman (1997) report studies of consistency, but questions have been raised about the accuracy of the judgements: it is possible that the very large numbers of documents to be assessed, and the effect of e.g. text highlighting to help the assessors locate topic words in document texts, means that some assessments are unconvincing. However it is quite possible that the overall impact is no greater than that of the 'mistakes' that users, when rapidly interacting during a search, may make, and that while the relevance sets may not be perfect as a base for future experiment, they are good enough for reliable findings about generic retrieval strategies.

Review of the data

In overall conclusion on the TREC data, in relation to their consequences for the test results, the following points have thus to be borne in mind:

1. The document files are posthoc assemblies rather than natural gatherings. Further, while explicit attention has been paid to obtaining subfiles with distinct properties, the range of data types is still somewhat limited. Thus while the document data does support claims that systems can survive heterogeneous, and also full text, files, the finer picture for the impact of document properties on retrieval is not clear.

2. The topics, though not taken from a working retrieval operation, do reflect genuine information-seeking practice. However in general there has been a topic building procedure of at least a relatively self-conscious kind, and some control on the formation of the topic set as a whole. So though, especially with Adhoc, there has been an explicit concern with naturalness and range of topics, the topic sets are not simple samples from a working situation.

3. The changes to topic design for the Adhoc task over the TREC Programme, while individually well motivated, and in the event very informative, make it difficult to establish clear relations between request properties and system performance.

4. The care with which the basis for relevance assessment is provided for each topic, though essential for key evaluation purposes, may be unrealistic in implying tighter specifications than are often found in practical searching, where evolution of need is observed.

Altogether, the TREC data is a mix of the considered and the fortuitous. But though this may preclude some investigations, it is natural and rich enough to support a mass of valuable experiment.

2.7 Performance measures

The measures used for evaluating performance for the main Adhoc and the track tasks in TREC are summarised in Table 8.

TREC has throughout relied primarily on thoroughly conventional performance measures using Precision (P) and Recall (R), in the style originating with Cranfield and made familiar by the Cornell SMART Project. This has been justified by tradition and familiarity. which have made it possible to place the TREC results in their wider context, which is that defined by prior laboratory experiment. TREC is designed to throw light on the system elements of retrieval, abstracted from particular application contexts and direct involvement with users, not on operational systems in their environments. The link between laboratory experiments and practical situations is provided primarily by the use of Precision as a performance measure: this is important for actual users, or at least seems to be more important than Recall, especially with the growth of search files where seeking high Recall implies getting far more documents than users want to inspect. Modern developments, notably with the Web, moreover, make it far easier than formerly to supplement basic searching by following inter-document trails: these offer means of improving Recall without the 'normal' loss of Precision.

The details of the measures used, and their specific implementation using SMART procedures, are given in the Proceedings Appendices (e.g. Appendix A, TREC-5). They are Precision at standard document output ranks, Precision at standard Recall levels, and two single number measures, Average Precision (AveP), and averaged Precision at query output rank corresponding to the query's number of relevant documents to retrieve (R-Precision).

The main specific innovation in the treatment of Precision and Recall with TREC has been the systematic use of Precision at standard document ranks as well as Precision at standard R. As collections become larger the implications of Recall levels for the numbers of documents to be inspected become less apparent, and document cutoff is much more transparent. However taking both together gives an informative (approximate) correlation, though it is necessary to bear in mind the fact that R is relative to the pool, not absolute. The TREC practice of using multiple ways of treating basic P/R information is also valuable in emphasising the important point that the same underlying search facts can be arranged or summarised in different ways; and because they can, that the absolute numerical values associated with any specific measures do not capture the only truths about performance. Thus to draw conclusions about TREC, consistent trends across different views have to be evident. This further implies that, as in "Reflections", overall remarks about results and findings can only be broad, informally-expressed ones; and that statistically significant differences, even if in principle necessary and in practice available, are not always sufficient for big picture conclusions, especially from the point of view of implications for designers or users of operational services. This applies particularly to Average Precision, which is highly reductive and is computed over far more documents than humans are usually willing to consider, even if AveP has research value as a reliable relative performance predictor (C. Buckley, personal communication). Certainly, while showing that differences in AveP are statistically significant is both methodologically important and informative, they may not be reflected in user's perceptions about 'real' performance differences.

In general, performance differences on the scale of interest for this paper hold across the measure set used. *For illustrative purposes*, for the reasons mentioned above, I have used Precision at document cutoff 30 (PDoc30 or, sometimes, Precision at DCut30) as a convenient way of characterising performance, as in the next section.

Significance testing is a problem at several levels for TREC. The first is that the best available tests (t, Wilcoxon and sign) are not wholly suited to the IR case (for TREC-related attempts to do better see Tague-Sutcliffe and Blustein, 1995). The second is that while individual teams may apply them to particular comparisons of interest to them, the official run submission does not directly involve comparison: thus to determine, for example for the present paper, whether two sets of results are significantly different requires a specific new application of tests to the data - a potentially feasible but practically daunting enterprise. The third is that, as already mentioned, significance testing is not directly applicable to the broad cross-TREC comparisons with which this paper is concerned. I shall therefore just assume that differences of note for my present purposes are in fact statistically significant, whether directly or indirectly in that all legitimate instances of the comparison are so.

Track measures

The measures just discussed are those used for Adhoc, and also for the Routing task. The same methods have been used for some of the tracks, for instance Multilingual. For other tracks, as shown in Table 8, different measures have been used, for two reasons. One reason is that the track goal itself requires different measures: this is most clearly illustrated by the Interactive and Filtering tracks. The other is that the practical difficulties of obtaining enough relevance assessments in a short time forced cheaper alternatives, carrying with them in most cases a change of the test task, as in the use of known-item searching: while this is valid in its own right as a task, as mentioned earlier, it has been used as a surrogate for normal Adhoc searching, with a corresponding change to performance measures, as for the Confusion and SDR tracks.

From one point of view, these measure variations add to the difficulty of drawing clear conclusions about relative strategy value. But from another, they can add to the support for one strategy as superior (or not) to others.

In some cases, formal measures have been accompanied by other data gathering, for instance event logs for Interactive, data about timings for VLC.

3 Results

As noted, it is very hard to see the wood in the trees, given the vast mass of TREC results, even when considering only the official submitted runs without reference to other reported experiments. In this section I shall attempt to present and analyse the main results, concentrating on the Adhoc case. I shall first give a performance summary; then review the results in the light of questions that have been asked about factors, specifically system parameters as embodying strategies and devices, affecting retrieval; then examine the critical contrast between automatic and manual query indexing; and finally consider the evidence from the track tests.

3.1 Adhoc results summary

As an anchor for the discussion of TREC results and findings, I shall use Tables 9 and 10. Omitting TREC-1 as a startup effort, they summarise performance up to TREC-6, following conventions I have used for these comparisons as published in Appendices to the TREC Proceedings from TREC-3 onwards. (The tables are those given in the TREC-6 Proceedings, with apologies to Verity for having earlier omitted them for TREC-3.)

The intention is to give a broad picture of levels of and trends in performance. The results are shown as Precision at Document Cutoff 30 (PDoc30), as proposed earlier. However the P values are grouped, to emphasise the point that small differences, even if they may be statistically significant, are not necessarily important in real terms, though the grouping also has the consequence that close original values may be split between adjoining blocks.

The conventions followed are that the best of two official runs is taken, regardless of strategy used, where there are two that can be taken as alternatives in the same category. The figures

are truncated, not rounded. The team names per block are *not* in merit order, but in published run order, and have been assigned simple, hopefully identifiable names. The tables are confined to Category A participants, and cover only the higher levels of performance, not all the runs submitted.

Individual team status in these tables, when shown over time, is clearly of interest, but the main points to be made about the table data apply across teams. It should also be emphasised that as many teams have not participated throughout, or have recently concentrated on tracks, no inferences should be drawn where teams figure only occasionally in the tables. Moreover, while some teams have tended to do consistently well, they have sometimes not profited from experiments, while others who started less well have improved their performance,

Unfortunately, as the earlier discussion of the TREC topics implies, the changes to topic styles and run specifications make it hard if not impossible to make detailed points about strategy/topic correlations that apply across several TREC cycles. These changes are reflected in the division between Table 9 for TREC-2 - 4, where teams could use either automatic or manual query formation, and Table 10 for TREC-5 - 6, and within the latter, where the automatic and manual *modes* of query formation are distinguished (a/m) and where the split between *versions* of the topics (V/S/L) has also to be taken into account. The Tables nevertheless support a number of important general points about the TREC tests and their results.

1. Ad hoc best performance improved from TREC-2 to TREC-3, even though the TREC-3 topics were less rich. The latter may, however, also have been rather undemanding, so the sharp fall in performance for TREC-4 may be attributable to harder as well as minimal topics. 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 'harder': see Voorhees and Harman's Overview, TREC-5 Proceedings. 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, but the S results for TREC-4 and TREC-5 are not 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.

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 for automatic, 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. First analysis of TREC-5 - 6 suggests this remains true, but the picture is complicated by other system differences (see further Section 3.3 below).

Conclusions from the summary

The overall conclusions to be drawn from these Tables 9 and 10 are:

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

2. The general findings about retrieval strategies for the early TRECs reported in "Reflections" 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,dl* [1] paradigm with relevance feedback refinement. But even with good data (as illustrated by the TREC-6 L version topics), PDoc30 is more often than not below 30 %. For the collection data used 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. In particular, 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.

In the next section I shall present and analyse the TREC results in more detail, from different points of view.

3.2 Analysis: questions and answers

In "Reflections" questions about important system parameters were used both to impose a categorisation on the very broad range of individual tests and to isolate key environment variables or variable values that might account for performance differences. As mentioned earlier, these questions remain helpful as a way of throwing light on the outcomes of the whole series of evaluations from TREC-2 to TREC-6. I shall thus use them again here, but concentrate primarily on the answers to them to be found from consideration of Tables 9 and 10. I shall ignore anything that relates to less well performing teams in individual TRECs since their results may be attributed to start-up factors, rash experiments, or simple goofs: most teams that have persisted with TREC have reached a respectable performance level; and after six TRECs it is important to look for trend information as displayed by consistently better results. At the same time, for detail, I shall concentrate on TREC-5 and TREC-6 where we have the combination of most evidence for automated searching, with many teams, and not over-helpful topics.

I shall begin by re-running the questions used in the previous paper, assuming their background and motivation there, and answer them with the focus on automatic searching (and systems as

---

[1] I use this name rather than the more usual *tf*idf* to emphasis the now common incorporation of document length.

wholes), though taking manual search effects into account. I shall also, since the volume of test results is now so great, give general answers, without referring to or elaborating on individual systems and team performance except where this is especially appropriate.

For convenience I shall group and label questions as in "Reflections'; and I will abbreviate them.

There are also now some new questions to ask. These are marked with a *.

Indexing and retrieval models

The first questions in "Reflections" were about indexing and retrieval models, the underpinnings of systems:

M1:   Are linguistic models better than statistical ones?
M2:   Are there performance differences within either class?
M3:   Are more refined or rigorous models in either class superior to crude?
M4   Are linguistically grounded compound terms better than conjoined simple?

The linguistic models referred to here are those explicitly applied in automatic processing: manual query formation may tacitly rely on linguistic processing, but only informally.

By now there are only a few teams engaged in the heavy-duty natural language processing (NLP) that requires document text as well as query processing. The most notable are GE/NYU and Clarit (CLARITECH), with the former illustrating more extensive processing. The general indexing is however in much the same form: along with single terms it uses compound 'syntactic phrases', consisting of stem- and order-normalised term pairs. Since document analysis is expensive, UMass earlier applied a simpler strategy, analysing queries for proper phrases but requiring only document proximity for phrase components; but this has been abandoned. In general, though teams employing NLP have reached good performance levels, these have been no better than those reached with 'statistical phrases' defined by adjacency or proximity, either simply from the given request, or with reference to a vocabulary established by document file processing with frequency filtering. In fact performance with syntactic phrases is not demonstrably superior to that obtained simply with coordinated single terms. With very short topics, the scope for NLP-based query indexing is indeed limited, so the main issue is whether a phrase vocabulary is helpful for query expansion: this question is considered below.

The TREC NLP track in TREC-5 and TREC-6 was intended to encourage analytic experiments designed to identify the role and value of NLP-based indexing independent of, or in conjuncton with, other devices. The track results suggest that though performance gains can be made with NLP-based indexing, they are only small, and it is difficult to make more specific assertions because other factors, e.g. manual query formation, or devices, notably feedback and weighting, are more valuable and dominate performance.

The answer to M1 is thus as before, namely 'No'; and similarly for M4, 'No'. With respect to the class of linguistic models, the answers to questions M2 and M3 is 'No'.

(The place where some attention may have to be paid to the linguistic properties of the material is in dealing with special properties of a language e.g. characters in Chinese, and in cross-language retrieval: these are considered later.)

Turning now to non-linguistic models, there have been teams applying sophisticated statistically-based document processing, e.g. the Bellcore work with Latent Semantic Indexing, and both CUNY and IRIT use connectionist approaches. Again, these have not proven especially superior to simpler approaches using statistical information only for straightforward weighting or term association purposes. Connectionist approaches using feedback make learning explicit, as regression analysis also does within the probabilistic framework. This is considered later under questions about learning. Thus for non-linguistic models, the answer to M2 and M3 is also 'No'.

Some systems, especially commercial ones, may offer a range of search devices and may allow these to be combined in complex ways in elaborately structured queries. But these are typically associated with manual searching and do not clearly reflect a model-based approach to retrieval.

However the use of methods that combine several forms of evidence to reach a document's matching score can be viewed as applying a model of indexing as properly multi-facetted. These methods may involve index terms of different types, e.g. single terms and phrases - cf GE/NYU's streams approach in TREC-6, - perhaps with differentated weight values; or they may involve query formulations of different types, as in UMass's INQUERY system. This type of *fusion*, emerging with a single initial score per document, has sometimes been labelled "query combination", to distinguish it from "data fusion". The latter is defined as obtaining a ranking, with each document's final score, from the separate initial scores delivered by different systems, which themselves need not use fusion (Belkin et al. 1994). But the distinction is not clearcut, so I shall simply refer to *query fusion* as using multiple treatments of the user's request. There are other forms of fusion: see scoring criteria, below.

On the whole, however, though query fusion could in principle cut across even broad distinctions between model types, it is normally applied in TREC within the framework of one generic model, as for instance with the GE/NYU streams approach, which is applied within a linguistic framework. Query fusion has in any case not been demonstrated, though tried by several teams in TREC-5 and 6, to be of special value.

These negative answers, after six TREC cycles, to the questions above naturally suggest a further, new question:

    *M5: Does having a model matter?

It is not at all evident by now that having a specific model in the sense of Vector Space versus Probabilistic, for example, does make that much difference: what seems to be more important is whether whatever model is used covers critical factors about term occurrences in the mass. The problem with NLP-based approaches to classical IR is that, in contrast to the statistical ones, they do not naturally incorporate this type of information, though they may additionally exploit it. Thus a rather weak answer to M5 is that models have a role primarily as clean ways of characterising IR, or as suggesting new, useful things to try. However the fact that strong model-based approaches, like City's Probabilistic one, have consistently been upper-level performers suggests the more positive answer, namely that having a model has real value in in supplying the right system grounding.

Indexing vocabulary

The vocabulary questions asked in "Reflections" were:

    V1:   Does a holistic approach to the indexing vocabulary pay its rent?
    V2:   Is linguistic sophistication important?

Essentially, over successive TRECs these questions, traditionally perceived as important, have faded into the background. While use may be made, especially in manual query development, of any vocabulary aids in the form of thesauri, terminology lists etc, the notion of a free-standing indexing vocabulary where the status of individual terms and their relations to one another are decided, seems not very pertinent (or even if believed pertinent, is unattainable). It is not clear whether the use of vocabularies as an aid in manual query development is an essential contributor to the generally better performance obtained with manual queries. Experiments with manual query formation have generally involved many 'system' (ie. user+automated system) parameters and parameter settings, and have typically not involved careful, controlled studies of the effects of individual parameters. It is therefore impossible to determine whether, for example, a good result is attributable to the use of vocabulary aids or just to spending a lot of time on query formation. The TREC Interactive track is specifically designed to allow studies of this sort, but has not yet accumulated enough data to answer questions like V1.

Manually-formed vocabularies of conventional kinds, e.g. Wordnet, may be exploited for automatic searching, but so far without clear benefits. Certainly it is not evident that particular forms

of vocabulary search aid are especially useful (the subject of much research angst in the past), as opposed simply to having some suggestive thesaurus, terminology list or whatever to prompt the user. There is some use of automatically-constructed phrase vocabularies, whether linguistically or statistically motivated, playing a helpful but not necessary role. The value of a phrase vocabulary, as used e.g. by Cornell or CUNY in statistical form, appears to be in making reliable, i.e. collection-motivated, phrases available through document descriptions for query expansion. Statistical indexing techniques like LSI treat the indexing vocabulary holistically, in a very abstract way but, as noted above, without evident performance benefits. The main form of vocabulary treatment manifest in TREC, also responsive rather than prescriptive, but only very minimally holistic, is the use of *idf* weighting: the evidence by now - though there is no longer much direct test of the proposition - is that it is moderately helpful. Such vocabulary 'relativisation' is also achieved by learning.

Thus the answers to V1 and V2 are both 'No'. The implication of TREC must appear to be that when faced with full text, the benefits of responsiveness and flexibility outweigh those of an engineered descriptor vocabulary. So the new question

*V3: Does having an indexing vocabulary matter?

would seem (at least for the monolingual case or for non-specialist technical vocabulary) to be answered 'No'.

Document descriptions

The questions here relate to the implications of the models for the nature of individual document descriptions, and specifically for the form and selection of terms:

D1: Is linguistic indexing superior to statistical?
D2: Are compound terms superior to single?
D3: For full texts, is document-specific weighting useful?

We have already covered most of what can be said about the two alternative bases for terms. With particular respect to document descriptions, the main point of importance is whether explicit reference is made to discourse-specific or other properties of documents to justify term assignment. But there is little to report here. The answer to question D1 thus follows from the answers to the Model questions, namely 'No'. With respect to compound terms, however defined, the TREC papers show a general commitment to, and some modest evidence for, the utility of compounds but, as noted, these need be defined by no more that constant conjunction in the file or query word proximity. The answer to D2 is thus 'Yes, slightly'.

The TREC work has, on the other hand, shown both the value of document-specific term weighting and the importance of suitable document length normalisation: the latter is well illustrated by the progressive modifications of the Cornell formulae. There has also been a general trend to shift weighting from the matching score function to document and request descriptions, also a trend observed over time with implementations of the Vector Model. This general shift seems to have been practically motivated: in some model-based approaches, e.g. City's, the scoring function has generally been expressed as a sum of (document- and query-specific) term weights; but in others is is not so clear they can be thus decoupled. An example of a more complex function is provided by CUNY: here decoupling requires each term to have four separate weights, two relating to the document and two to the query. A *tf* component figures in most systems, and correspondingly a *dl* (document length) one in some form or other; but equally from year to year there are reported experiments with variations designed to best respond to the range of document sizes typical of the files. The straight answer to D3 is therefore 'Yes', though it has to be accepted that with some systems where elaborate, manual query development is undertaken, as with Waterloo, there may be no term weighting or it may be incompatible with the query format; but performance as just as good if not better.

This discussion also leads to a new general question:

*D4: Do documents need descriptions?

And in concordance with the general shift to request-based indexing, the answer appears to be that with full text, there is no requirement for autonomous document descriptions, in the classical sense exemplified by the provision of a subject characterisation at file time.

It should be noted that it is desirable to pay some attention to the treatment of names, abbreviations, acronyms etc: real carelessness on this appears to depress performance, but fully automatic procedures are not completely effective. (It follows there may also be value in specialised name vocabularies giving variant equivalences.)

Indexing sources

The "Reflections" question here is:

S1: Are subdocuments useful sources?

I noted in "Reflections" that in conventional systems it may be possible to restrict the search (i.e. index) fields, for instance to abstracts as providing helpfully concentrated content material; and modern Web systems may offer the possibility of using format restrictions e.g. to titles or subheadings. Such system parameters have not been open to systematic investigation in TREC, largely because much of the file material does not allow it. However there has also been a general presumption that whenever full text is available, it is always preferable to use it, while the file properties have precluded studies of relative subdocument weighting.

The main focus of study on subdocuments has therefore been on *passages*, whether natural paragraphs or arbitrary-length (and overlapping) segments. True passage retrieval, as a substitute for document retrieval, cannot be studied in TREC since there are no passage relevance judgements. Work with passage matching has thus taken two different forms: using them as a surrogate for documents in scoring, and scoring them in conjunction with full documents in so-called global/local scoring; the motive in both cases is to prefer documents where query terms are locally concentrated, implying genuine treatment of the topic in question. The former would also naturally fit with a selective display of passages to the user. For the reason mentioned, however, the global/local case has been the main one studied. It has been a staple of successive TRECs, and has included dynamic subdocument indexing where documents are simultaneously indexed by passages of varying length, with a final score determined by their best alternative. The evidence, e.g. from MDS/RMIT's work, is that passage-level indexing, whether as the sole or one of the scoring bases for documents, may provide some modest performance gains. Thus the answer to S1 is 'Yes, slightly'.

Passages have also come to have somewhat different roles. One of these, which has become increasingly important - see further below, - is as a bounded field round matching query terms from which query expansion terms are drawn. Though passages are not a consistent feature of the better-performing systems, evidence from e.g. Cornell and UMass suggests that, when interpreting S1 from this point of view, the answer is also 'Yes, slightly'. It should however be noted that both in this case and in the previous one, working with passages presents problems of weight formulation that are not yet fully understood.

Passages also have a obvious display role in interactive searching. But the Adhoc task specification does not address the user interface and output presentation, and the Interactive and High Precision tasks have not involved systematic study of system functions like display.

Queries and query sources

With the increasing emphasis, for both theoretical and practical reasons, on query as opposed to document-based indexing, the questions put and answers given under this heading are the most

important ones. In addition, as the earlier discussion of the TREC data implies, the nature of the TREC test requests (topics) and their changing character over the sequence of TRECs on the one hand, and the changes to the specification of the official Adhoc runs on the other, make queries the crucial area of concern in considering the detailed TREC results and wider TREC findings.

The "Reflections" questions about queries (slightly modified) were:

Q1a: What form of index description is best?
Q1b: Is complex indexing superior to simple?
Q2:   Is manual query formation better than automatic?
Q3:   Is query expansion valuable?
Q4:   Is feedback in general useful?
Q5:   Are statements of relevance criteria a good source of terms?

However taking Tables 9 and 10 together with Table 7 shows a complex situation with a lack of clear comparability over TREC-2 - 6. But it is also most useful, in assessing the lessons of TREC for adhoc retrieval, to concentrate on TREC-4 - 6, and to consider elaborate, honed topics like those used in TREC-2 and 3 only in the context of the Routing task. Several teams, for instance Apple, Lexis and Waterloo, have concentrated in TREC-5 and 6 on the Short versions of the topics, and more specifically in TREC-6 on the Very short versions, precisely because these resemble those typically submitted by real users, e.g. in Web searching.

Though, as noted earlier, manual indexing may also be associated with complex query structures, when manual and automatic query processing are taken together the test results as a whole do not show that any form of query description is best, or that complex indexing is unequivocally superior to simple indexing. This applies whether e.g. Boolean formulations or query fusion, combining different types of query, are concerned. Thus the answers to Q1a and Q1b are 'No'. The performance trends over successive TRECs do, however, show that manual query formation is advantageous even when fairly minimal, as long as other good devices are also applied e.g. CMU/Clarit. That is, while in TREC-4 a few automatic results were comparable with manual, for short queries, the trend in TREC-5 and 6 is clearly for manual query construction, normally done with Long topics, to outperform automatic query formation from the same Long sources. The answer to Q2 is thus evidently 'Yes'. The table also, not surprisingly, shows manual query formation outperforming automatic when the latter is restricted to Short or Very short topics.

Query expansion has been a widely exploited device in TREC, whether done in the conventional way before searching, via a thesaurus, or after searching, using retrieved documents. One important development has been in the use of *blind*, or *pseudo-relevance*, feedback, where top-ranking documents in an initial 'pre-search' are assumed relevant and are exploited to develop the full search query. If feedback by definition involves users, blind feedback is not in fact feedback, or alternatively any use of collection data as in *idf* weighting is feedback. However as in practice blind feedback is seen as an approximation to user-driven pre- or mid-search feedback, we will retain the term. In early TRECs, the rich topic information made so-called massive query expansion viable, but expansion has been more carefully limited in later TRECs in order to maintain focus on the essential request topic. The overall evidence is clearly in favour of expansion, with gains even from fully automatic searching using blind feedback. That is, the answer to Q3 is 'Yes'. The answer to the closely related Q4 is also 'Yes'.

The investigations done under the Interactive track should provide further evidence on these query management issues, but it is still too soon to draw solid inferences from these track studies.

Overall, the experience of TREC reinforces the trend towards query- rather than document-based indexing, reflected in the new question:

*Q6:   Is query development the most critical factor in retrieval?

to which the answer is unequivocally 'Yes'.

The final original question (Q5) was natural given the character of the early TREC topics. It is difficult to be sure that the particular quality of the Narrative component of the Long topics

is significant, as opposed simply to the supply of more information (words) about the topic. The answer to this final query question must therefore be 'Don't know'.

Search strategy

In general, any form of query development, for example expansion, is a search strategy. But in "Reflections" it was treated as a separate factor, motivated by the classical view that refers to broad or narrow search strategies etc, that are deliberately chosen, especially in interactive searching. However "Reflections" noted that in the early TRECs there was little systematic study of alternative search strategies; searching has been merged into indexing, and in general search devices have been used in a standardised manner: thus expansion covers both precision and recall promotion. Again, as mentioned earlier, some teams have focussed on strategies designed to to help with very short queries, on the grounds that this is the usual case. Strategy application is a natural area for the Interactive track, but the conditions of the Adhoc tests have meant that strategy choice and development, a dynamic process in human searching, have been covered in a different way. In automatic searching different strategies have been so to speak conflated into the single query description (or set of alternatives as in fusion), or have been pursued in a no-holds barred manner in order to develop the best possible manual query.

TREC participants have indeed recognised the need to develop methods for automatically choosing indexing and searching strategies that are suited to individual query properties, i.e. do more than adapt a given strategy to an individual submitted query, as in preferring some terms for expansion. But little material progress has been made on this difficult problem. Given the large files, there is a natural bias in TREC towards Precision strategies; however the main area where there has been a more concentrated focus on Precision strategies has been in connection with getting a good set of top-ranking documents for feedback, especially blind feedback: see e.g. Cornell in TREC-6. However these have not been systematically evaluated in their own right, trading off with Recall. The High Precision track, and also known-item searching as in the Confusion and TREC-5 Spoken Document Retrieval tracks, would be expected to stimulate appropriate strategies, but there are no solid conclusions to be drawn about these from the limited tests so far.

Thus the answer to the strategy question:

> Y1: Are uniform (and automatic) strategies good enough?

so far has to be 'It seems so'.

Scoring criteria

The "Reflections" question here was:

> C1: Are complex scoring functions better than simple sum/product?

The TREC design requires ranked run output. Some teams, especially those engaged with manual searching and with commercial systems, e.g. Excalibur/Conquest in TREC-4, make use of complex scoring functions. Automated systems are generally simpler but may treat single terms and phrases differently, or apply global/local matching. The latter is document dependent, and can be viewed as a rather different form of fusion, which I shall call *document fusion*. Different ranking formulae suited to documents with different lengths may lead to a need for fusion (Lexis, TREC-5), as may the use of other document features like publication type or collection source. But there is no strong evidence that elaborate scoring criteria treating different elements of index descriptions, or generic document attributes, in different ways, deliver better quality (i.e. ordered) system output. The answer to C1 is thus 'No'.

It may, however, be more appropriate now to ask a slightly different question, namely:

*C2: Can the best scoring functions be reduced to simple inner product?

This question is related to the earlier discussion under document description, with the implication that factor information pertinent to retrieval is better located in descriptions than scoring functions. As noted there, some scoring functions are already in this form. The evidence to date suggests that even if there are scoring functions that are not so reducible, they are no better than ones that are.

Output form

As noted in "Reflections", this is an important system parameter in real life, but it has no place in the TREC evaluation format.

Learning

An IR system may have the ability to learn under many headings, but broadly speaking these can be assigned to the file level or to the query level. Though learning has been implemented in systems in the past, for instance for automatic assignment of indexing categories, the TREC community appears to have been the first to engage in wholehearted learning, or *training*, at various points and for various purposes within a whole system framework. This has been a natural consequence of having very large document files, and substantial past query and relevance sets, to play with. The various forms that training could take are discussed in "Reflections", ranging from whole system *adaptation*, which was there seen as *tuning* to a collection, to individual adhoc query *modification*, for instance by relevance feedback. At the file level, participants focussing on statistical methods, for instance Berkeley, have made heavy use of training through regression analysis, and training is also built in to connectionist approaches as with CUNY and IRIT. The many teams applying relevance feedback are training at the query level, and the Routing task and Filtering track have essentially been studies of the value of training when more assessment data is available then in the usual adhoc case.
    "Reflections" asked four learning questions:

    L1:   Is adaptive tuning to a collection valuable?
    L2:   Is refined vocabulary revision helpful?
    L3:   Does relevance feedback help?
    L4:   Is regression analysis useful?

Though participants using adaptive tuning and specifically regression analysis have been among those doing well, they have not done better than teams not using any training at all, or only minimal training in such forms as trials to determine constant values in weighting formulae. Thus while Berkeley shows that regression analysis alone can compensate for a lack of other devices, the overall straight answers to L1 and L4 are 'No'. There is also no evidence that refined vocabulary revision, e.g. omitting terms with certain properties, is helpful, so L2 is also 'No'. (This excludes the weak sense of revision represented by *idf*-type weighting, which is of some value.) The TREC results do, on the other hand, support the answer 'Yes' to L3, as being valuable if not mandatory.
    As a whole, for the new general question:

    *L5:   Is systematic training of value?

the TREC answer is 'Yes, when focussed on queries'.

3.3   Analysis: automatic versus manual

In the previous section, I considered TREC-2 - 6 from the point of view of questions, primarily about indexing and searching *devices* but also about broader *strategies* and, ultimately *models*

underlying these. In this section I shall consider the TREC Adhoc results as given in Tables 9 and 10 from the complementary point of view, namely what they show about actual performance: specifically, what they show about attainable levels of performance, and about the conditions in which and means by which these levels are achieved. As the points made in the previous section suggest, the contrast between automatic and manual means is of special importance.

Thus, summarising, we see that the more successful participants achieved good performance, even with fully automatic systems, in TREC-2, and further improved on this, though the valuable topic Concepts field was abolished, in TREC-3, presumably through experience. However the short topics in TREC-4 stimulated a bias towards manual query, though two teams, Cornell and City, did well enough by purely automatic means. The best level of Precision performance for automatic systems at DCut30, $> 50\%$, in TREC-2 and even higher in TREC-3, fell off to 40% in TREC-4. TREC-5 shows the impact not only of brief but of difficult topics, with a further large fall in performance to only 25% P for fully automatic Short topics, though a little better with Long. The benefit of manual query construction in this case is marked, with many teams exceeding 35% P. The picture is much the same for TREC-6, subject to the 'confusion' about the relationship between Title and Description field described earlier in Section 2.6.

Further, while Average Precision (AveP) may be problematic as a method of describing, as opposed to predicting, comparative performance, it should be noted that an analysis of TREC-5 and TREC-6 AveP performance, using blocks in the same manner as for PDoc30, shows a similar relative distribution of the teams, with only modest block-boundary crossing. Conclusions can therefore fairly safely drawn from Tables 9 and 10. For instance, the direct comparison for the best three blocks between automatic and manual on Long in TREC-6 shows a performance difference between the range 20-34 for the former and 35-49 for the latter (excluding Waterloo - see later) that should be visible to the user, not just statistically significant, i.e. a difference at middle value in each range between 8 relevant and 12. Getting 8 relevant documents is nevertheless respectable.

Thus we should now consider first the particular methods used in the automatic case, and second the effort and resources used in manual searching. The former has to cover both similarities or differences within the upper block and between the upper and lower ones. It is also necessary to consider interactions between methods and the Very short, Short and Long query lengths.

Automatic methods

In general, as TREC has progressed, everyone in the automatic camp (and also some in the manual, if it is compatible with their general strategy) has come to adopt *tf,idf,dl*-type weighting. This can be seen as a TREC confirmation of earlier research findings, but modified (with respect to document length and, often, specific matching coefficient) to suit the full-text document condition. The forms used by leading teams are local variants that appear to be similarly effective. There is little doubt that this type of weighting is really valuable.

It is therefore more useful to consider the other devices used, to see which do, or appear to, make helpful contributions to performance. This has regrettably sometimes required inference, since it is far too often the case that TREC participants' papers detail the various alternative strategies studied and compared, but without identifying the particular ones defining the officially submitted runs. [2]

The strategy elaborations, beyond simple single term weighting, that recur and are thus worthy of comment here are:

1. the use of 'phrases', i.e. compound terms (whether 'syntactic', i.e. linguistic, or 'statistical' i.e. associative or proximity-defined);

2. the use of passages (whether to rank output, or to limit the base for expansion terms in feedback);

---

[2]At the fine level of detail it is more than likely that future researchers will not be able to determine precisely what has been done, and hence to replicate it, though perhaps generalisation, even if involuntary, is desirable in the interests of robustness.

3. the use of blind, or pseudo-relevance, feedback, (whether just to reweight terms, or to support query expansion);

4. the use of query expansion (whether via vocabulary invocation, i.e. manual or associative thesaurus, etc, or via fedback documents).

Breaking Tables 9 and 10 down by the detail of participants' systems on the four *devices* Phrase, Passage, Blindfeed, Expansion, we find that over TREC-2 - 6 there has been:

1) a clear increase in the use of phrases, which are typically statistical rather than syntactic. Phrases are normally handled additively, i.e. their component terms are treated as separate single terms in their own right as well.

2) a persistent, though not dominant, use of passages, with increasing interest in their role as context windows for query terms in pulling in expansion terms, rather than as matching arenas for document scoring, though attempts continue to be made to get benefit from the latter.

3) and 4) a very marked growth in the use of blind feedback and of query expansion, and typically of the latter via the former. This is manifestly a response to the need to amplify short queries: the main research topics have been the appropriate degree of expansion and focussing of expansion sources in the documents.

Table 11 gives a snapshot for the use of these devices in TREC-6, indicating that they have become common. Specifically, it shows the devices applied by the teams in the top PDoc30 block for each of the Long, Short and Very short query versions.

It is by now the case, as is clearly shown by the TREC detail, that the four devices are combined, on top of the basic tf, idf, dl weighting, as the default automatic strategy. (The most common additional refinement is then probably query fusion.) TREC can thus be seen as essentially endorsing earlier research findings about useful devices. This applies particularly to the core strategy represented by weighting. But, setting aside passages which could not be fully explored without full text, both with weighting and the other devices the relation between TREC and earlier research is sometimes a subtle one. For instance, with weighting by requiring development to deal with full text, or with blind feedback by emphasising that this is only likely to work when there are enough relevant documents to ensure that they in fact populate the upper ranks on which the feedback step draws.

Manual methods

As noted earlier in Section 2.6, the specification for manual runs has changed over TRECs; however setting aside TREC-4, it has normally allowed feedback as without this manual searching hardly (in modern situations) deserves the name. With manual queries the strategy *elaborations* it is useful to consider are:

1. the amount of human effort (specifically: is this large?);

2. the role of expansion (especially by reference to terminology resources);

3. the use of feedback, i.e. explicit human relevance judgements (normally as a base for suggesting new query terms).

From TREC-2 onwards, good manual performance has been associated with non-trivial, and in some cases very large, human effort. Query expansion is also usual, with by TREC-6 considerable reliance on feedback. It is noticeable that query preparation times can rise to as much as two hours (Waterloo, in TREC-6), and may average 30 minutes. In general, good manual query performance is associated with considerable human effort, whether in constructing complex expressions (as

required by some established commercial systems) or in developing a query's term composition. However two of the better-performing TREC-6 teams, GE/CMU and Clarit have relied on only modest human effort, largely applied to assess search output for feedback. Interestingly these are teams also using NLP; but other teams have done as well with somewhat more (though not necessarily huge) direct human query effort and less heavy system document processing, so the payoff from the NLP is not clear.

Table 11 gives the strategy elaborations used by all the TREC-6 teams doing manual searching shown in Table 10, emphasising the importance of expansion and feedback.

It should be noted that in TREC-6, where Waterloo had very high performance, manual search output was also manually ranked for submission - as in fact permitted by the guidelines - as in the conventional search-intermediary situation. This result can therefore be taken (for this document data at least), as setting a realistic upper bound to performance relying on human effort and judgement and given good starting requests, i.e. Precision exceeding 50% at PDoc30.

Method comparison

Beyond this broad picture, it is not clear that specific refinements make any marked difference. However, both to flesh the picture out, since individual teams do vary in detail, and at the same time to see what value there really is in the automatic devices and manual elaborations mentioned, we should look more carefully at the way these figure in characterising the top block systems for the various alternatives under TREC-6. (It must again be emphasised that these are only *illustrative examples*: mentioning the teams does not imply that there are no other comparable ones). In fact we find that for automatic devices, the characterisation for ANU, Cornell, IRIT, CUNY and Berkeley, with automatic searching and Long topics, as given in Table 11, shows considerable system variation along with similar performance. (Note that pretty well everybody does stopping and stemming and everybody does *tf,idf,dl*-type weighting.) With Short topics and Very short the general pattern is the same, even if the absolute level of performance varies.

Thus the overall conclusion to be drawn from Table 11 is that, as has been observed before, devices like those listed are individually not very powerful, and that even in combination they may not add very much to performance, beyond what a well-founded, basic, single term system with weighting can offer. It has to be borne in mind, as Lexis (in TREC-6) note in commenting on a systematic comparative study of a set of devices, that the TREC evaluation format can mean that the official submitted runs are not those with the best option performance. It is also the case that different teams have not all run the same set of device comparisons, so it is not clear precisely how much even the most consistently-favoured device, namely blindfed expansion, is actually contributing, or can be the only means of contributing. Nevertheless, the devices listed are all pretty robust, and do not need much collection tailoring or formal training.

Finally we may make an analogous comparison for the TREC-6 manual searching, as shown in Table 11. This too shows considerable variation but, not surprisingly, relevance feedback playing an important role, whether to support manual query reformulation or, as is sometimes the case, automatic query modification.

3.4   The tracks and routing

It is now time to consider the various track evaluations as they bear on the view just taken of the main Adhoc task experiments and results. As mentioned earlier, I shall not attempt a detailed discussion of the individual track detail in its own right, especially since in many cases the tests have not been on a large or continuing enough scale to support solid conclusions. I shall consider the Routing task and Filtering track, as dealing with a significantly different task, in the next section.

The track evidence

I shall summarise the track material in the groups introduced earlier. It should be emphasised that it is impossible, because of data differences, to make any meaningful direct comparisons between performance levels e.g. at PDoc30.

Data extensions: The most important extensions to the Adhoc task have been made in the multilingual Spanish and Chinese tracks. Participants in these have in general continued to use techniques applied to English, with similar effects: e.g. UMass used exactly the same methods for English and Spanish in TREC-5.

In Spanish in TREC-4, where there were more participants than in TREC-5, teams could reach a reasonable, comon level of performance with automatic searching, though UMass did somewhat better, presumably from applying NLP to query processing to identify phrases. University of Central Florida did noticably better with manual searching, but by applying very large effort to query formation. In TREC-5 the same applied for automatic searching, with little gain from manual query, even using relevance feedback; however as with the main Adhoc task, better results were obtained with long than short topics.

With Chinese, for the most recent fully-published results of TREC-5, manual processing brought some gain especially with feedback, but very respectable performance could be obtained, as Cornell demonstrate, by simply applying the methods used for English, regardless of whether Chinese characters have the same linguistic properties as English words.

It is however premature, especially taking the wide variations in language pairs tested, to attempt to draw any conclusions from the first Cross Language experiments of TREC-6. The main point for future interest is the extent to which single-language methods can be crudely extended so, for example, there is no attempt at word-sense selection.

The other direction for extension has been with 'noisy' data, in the Confusion and SDR tracks. The former has not attracted many participants, though it may be noted that when using OCR documents in TREC-5, ETH were able to obtain good performance with a probabilistic technique employed for other retrieval purposes too; the SDR track is too new for assessment in the present context. Both of these tracks have used known-item searching, however, which is not directly comparable with the standard Adhoc case.

Focussed studies: These are the Database Merge, NLP, and Interactive tracks. The first two were not large-scale efforts. The last included many teams in TREC-6, but the experiments still need extensive analysis and cannot be simply related to the Adhoc tests.

Variant tasks: This group covers the High Precision track and, to some extent incidentally, the use of known-item searching with Confusion and SDR. The High-Precision track in TREC-6 was designed as a manual search test; the results are quite comparable with the manual Adhoc searching (and the value of the manual search is confirmed by CUNY, which submitted an automatic search through lack of time). The main question is whether the track demonstrates system, as opposed to user, facility in adaptation to the task requirement.

More realism: This group covers the Filtering track, considered in the next section, and also the VLC track of TREC-6 as well as, to some extent, the Interactive track. Informally, since the TREC-6 VLC detail is not included in the Workbook, it is only possible to say that the track reports indicate that it appears possible to scale up the strategies used for the Adhoc tests, but the details need further analysis.

Routing and Filtering

Table 12 shows the upper block performance for the Routing task from TREC-2 to TREC-6, in the same style as the earlier Adhoc tables. As noted in discussing the TREC data sets, after TREC-4 the routing data has diverged from the Adhoc, and has also become somewhat smaller and more

tightly focussed. The Routing task work as a whole has demonstrated that when a large amount of training data is available, and also fairly informative topics (including Narrative sections), it is possible to achieve a very good performance level with the 'mainstream' approaches to automatic retrieval, namely of more than 50% at PDoc30.

However, ranking performance thus measured does not reflect the true binary nature of the task, and the Filtering track since TREC-4 has attempted to evaluate the task in a more appropriate way. But as Lewis (1997) makes clear, when the number of documents retrieved varies, comparability between teams is hard to determine. The measures used for the Filtering task have sought to capture an appropriate notion of *utility*, balancing relevant retrieved against non-relevant retrieved. Lewis' analysis for TREC-5 shows best overall performance for those teams, typically those applying probabilistic models in automatic searching, which is fairly well in line with relative Routing performance. A broadly similar outcome appears to apply in TREC-6; however the need to tackle the normal temporal sequencing of material, and dynamically changing relevance need, that normally charactererises filtering is only being addressed in TREC-7. It is therefore too soon to say whether the approaches and methods associated with Adhoc retrieval really carry over to the filtering task.

Track lessons

The foregoing may suggest that in spite of the very large amount of work that has been put into the tracks, and the enthusiasm that is associated with them, firm conclusions about the issues they address are somewhat lacking. From one point of view this may not matter: it can be argued that the tracks, even more than the Adhoc task, have other functions than rigorous comparative testing: for instance to make initial passes at new problems, e.g. cross-language retrieval, to develop evaluation methodologies, e.g. for filtering; or have broader functions like bringing different communities together, e.g. by linking speech processing and retrieval. However even on the narrow view that tracks have a central testing function, while it is indeed the case that a larger series of tests and/or more participants are required for many of the tracks, it is now possible to make an important general statement referring to all of them. This is that many teams have moved from track to track, tackling the various tests with the same apparatus as with the Adhoc task, especially that used for automatic searching, and have not fallen flat on their faces. The automatic methods are *viable*. This is significant because they are typically not expensive. The same familiar participants recur throughout TREC although, it is also important to note, they are continually kept on their toes by new players with variant or new methods.

At the same time, for those tracks not using data drawn from the Adhoc resources, it is hard to pinpoint the impact of each such environment variation, since these have been little controlled. It may also be difficult to create future test collections based on systematic variation. Thus the VLC data, in principle a potentially rich resource, has limited relevance assessment (pooling was on 20 documents per run).

4 Assessment

My assessment of TREC in "Reflections" was from two points of view: what general findings about IR could be drawn from the detailed results, taking the data features and evaluation formats into account; and what lessons could be learnt about IR test methodology.

4.1 Retrieval findings

The general findings on retrieval strategy in the previous paper, put briefly, were:

1. statistical techniques work alright;

2. they can do much better than the simple term baseline;

3. they can be effective, though individually weak, through combination;

4. they need to subsume some collection training;

5. they are computationally viable in initialialisation and application;

6. relevance feedback is valuable.

It will be evident from Section 3 that these continue to apply. The further, data-oriented finding, that request quality is very important, clearly also applies. However, after the further TREC experience, the implication of the strategy findings just listed, namely that fully *automatic* processing is not merely OK but as good as manual, may not be quite so convincing for poor queries.

Given the scale of the TREC effort indicated in Section 2 the results summarised in Section 3, and the findings just listed, appear modest and dull. Is there really nothing more exciting to be said about the largest programme of tests ever carried out in IR?

But it is important not to be misled by the combination of the apparently bland, negative character of results when presented in the form "On the whole, X doesn't work". Thus seemingly negative statements, like "Elaborate indexing devices don't work" are in fact justifying thoroughly *positive* statements, namely "Simple indexing devices can, and do, work pretty well". Further, these TREC results are *confirming*, with the advantage of much larger scale, the findings of many decades of work by such teams as the Cornell one.

It is moreover equally important to recognise that the broad-brush general conclusions that are appropriate in such an overview as this are based on a very large range of careful, specific tests that have to be studied in their own right to establish the precise base and scope of the generalisations. The participants' reports lay out many particular comparative experiments, of great importance as investigations of performance factors, that require detailed analysis. However it is already evident that this mass of tests has done somewhat more than simply confirm earlier work: the TREC work has been *consolidated*, by giving generalisations more substance, for instance in the appropriate form of query term weighting, or in the mode of employing phrases together with single terms. This applies both to individual devices and to what may be described as a whole style of indexing and searching.

Thus we can add some further general findings, as follows:

7. moderate query expansion is helpful;

8. simple phrases along with single terms contribute something;

9. some user assistance is needed to upgrade minimal requests.

Finally, and perhaps most importantly, we can draw on the fact that by now in TREC its tests have ranged over a far wider range of requests and request qualities than in the earlier paper. This means that we can, if only tentatively, offer findings about *absolute* rather than purely comparative performance levels. Thus (unfortunately) we have to recognise that the best Precision levels attained in TREC-2, namely round 50% at DCut30, are exceptionally high, implying that even with quite good requests (of the usual text sort, not full topics) one cannot envisage more than 30% Precision with 30% Recall, and that a conservative view of likely performance in 'ordinary' circumstances would suggest 20% with 20%. As a generalisation we can say:

30% Precision with 30% Recall at rank 30 cutoff is doing alright.

Again, though this might not seem very impressive, as Buckley, Singhal and Mitra (1997) observe, over the TREC Programme as a whole the systems have got much better, even if the data has made the Adhoc task so much harder the systems cannot fully keep up.

Some of the results obtained naturally require followup to establish the explanation for observed performance. Though we know a good deal about the form and extent of query expansion under various conditions, it is still necessary to establish the precise conditions for different degrees of expansion. Moreover as mentioned earlier, there has not been as much progress in TREC towards

the goal of designing effective *individual query*, as opposed to *query class*, strategies for the adhoc case. That is, while with routing/filtering using feedback can implicitly develop a query to better fit the individual case, we are not in a position where a system can automatically analyse a request for the implications of its term makeup and *choose* one particular query formation strategy as likely to give best results. Indeed more work is also needed on the best ways to handle different query classes.

The major issue is the added performance value that manual searching in general appears to provide, and specifically whether there is enough added value with minimal human input. There is not quite enough evidence, from the tests done to date, to factor out the different environment variable and system parameter effects to determine this. It is arguable that this is where the main thrust of the next TREC or two should be: addressing the relative performance of automatic and manual searching when requests are not merely (short) or very short, but where the latter in particular are rather less well-formulated than the TREC-6 Very short topics were; and seeking in particular to determine whether there is enough performance gain when the manual contribution is confined to simple relevance assessment, especially if this only covers a small number of documents and so identifies only a few relevant documents. (As a laboratory experiment the study would imply some modification of the treatment of search output, since it would appear to imply a requirement for frozen ranking in the feedback study.) It would also be useful to try to establish the specific value of low-effort initial manual modification of a candidate query.

The implication is that it would be more valuable for the TREC community to engage in a more carefully controlled 'cooperative' evaluation of adhoc searching, exploring the pertinent area of the test grid, than to seek to improve their own best performance under yet another task variation, in the hitherto established style. The aim would be more of an informative, glass box evaluation than the usual black box style for separate team effort in previous TREC cycles. There is a model for this in the way the TREC-5 and TREC-6 SDR tracks specified its official runs.

Data effects

In "Reflections" I concluded that the data used was not very 'natural' and that some of its properties, e.g. topic quality, promoted good performance. The TREC cycles since have introduced more realism, so even if the document files are still artificial aggregations of such distinct types they might never be searched as one in an operational system, there is less reason to treat the performance levels obtained as non-typical.

However the impact of specific document types on strategies, and specifically on performance levels, needs further investigation. Work has concentrated on showing that strategy A does better than strategy B regardless of document type. But we also need to know what the levels of performance across document types for strategy A are. This would again require the type of controlled evaluation envisaged in the previous section.

TREC is also still not testing with real users, though it has been moving towards data which simulates theirs. The investigation of manual searching just considered bears on moving further towards realism in this respect.

Methodology effects

The earlier criticism, of the untoward consequences of the 'routing bias' in the first TRECs no longer applies. But the adhoc methodology is still essentially (for all that some teams engage in heavy manual searching) an anachronistic batch one. The question of how TREC can engage with more natural interactive searching in its mainstream testing, without all the baggage that the specialised Interactive track is lumbered with, needs to be tackled.

Technology issues

It has proved difficult, within TREC, to address the computational efficiency issues and tradeoffs that are important for real systems, for several reasons. The TREC test situation is intrinsically

unrealistic and promotes e.g. the use of data structures tailored to repeated testing. There are also many aspects of real systems, that would enter into efficiency considerations, that are not covered by TREC at all, e.g. index recompilation. There is also the important point that TREC participants' systems are not required, for TREC purposes, to support real-time multi-user operation. Operational and commercial systems have been used by TREC participants e.g. Clarit, InText, but this does not imply that good techniques developed by other participants would simply carry over to viable operational use.

It is nevertheless the case that technological progress is working in TREC's favour, in that where computationally costly procedures are justified by research results it is less and less likely that these will be operationally unviable. From this point of view the VLC track in TREC-6 was reassuring in that its participants found scaling up to be less of a big deal than might have been feared.

Tracks

The foregoing applies to the Adhoc task. As will be clear from the earlier comments on the tracks, though tests in some tracks, e.g. the Spanish and Chinese ones, have been carried far enough to support reasonably solid conclusions, we still await substantive, well-grounded results for important ones. This applies to Cross Language and SDR in the Data group, and to Filtering in the Realism class. In general, it is a measure of TREC's success in raising evaluation quality that the results data from one particular track cycle are nowadays not acceptable as even remotely definitive, even if the cycle covers the same number of tests as would have been regarded as quite sufficient even ten years ago.

4.2 Methodology lessons

The lack of clear operational system guidance from the Interactive track studies so far, instructive though they are in showing how hard such studies are, leads naturally to the review of TREC methodology.

As noted, TREC has stayed firmly within the well-established laboratory experiment paradigm, for good test control reasons, even if these are difficult to satisfy, as is well-illustrated by the problems of designing a respectable (i.e. reasonable) filtering evaluation. The Interactive track is indeed still within the controlled test paradigm, and also illustrates the challenges of proper test design, especially with respect to sampling. The suggestions about future directions for adhoc experiment made in the previous section are also embedded within this controlled test paradigm, even though they are intended to address matters of importance for real systems.

In "Reflections" I noted that the computational technology used in the Programme in general, and more specifically in the processing of official runs at NIST, meant that it was possible to 'view' the search results in different ways, applying some different measures, possibilities that Donna Harman and her colleagues have pushed further in their own overview analyses (e.g. Harman, 1996; Voorhees and Harman, 1997). Even though the individual search outputs are heavily aggregated and averaged, and also represent a black box level of system comparison, the fact that this data can be seen from different angles makes the analysis more informative and conclusions drawn from it more substantial. The main methodological weakness in TREC Programme as such has been the lack of cross-team significance testing, though individual teams may carry out their own tests.

But though TREC gains much of its strength from its methodology, it is arguable that it is now time to place the the application of IR test findings up front, and hence to move TREC more firmly towards studies of retrieval in operational contexts, *even if* this implies, as a necessary consequence of practical and funding constraints, a weakening of the heavy laboratory control. Thus as it seems appropriate to engage in less 'abstracted' forms of retrieval task, say involving real user, but treating a range of user types in a properly controlled way would be too big a deal, it may be necessary to accept some relaxation of controls, and a more observational style of investigation. This need not, however, be too damaging if such studies were firmly built on, or complemented by, a solid foundation of conventional laboratory tests. However one point deserving more attention

is some creation (though it may involve adding further relevance assessments to existing sets), of more systematically formed or varied test collections drawn from existing resources.

5 Conclusion

The first point, in conclusion, is whether it is now possible to be *prescriptive* rather than *descriptive*, i.e. to be able to proceed from the essentially descriptive findings to prescriptions for generally sound and effective system design. In my view it is now possible to say that:

> *The 'core' statistical approach, with tf,idf,dl weighting, statistical phrases, blind feedback, moderate expansion, and application of passage constraints, will deliver useful goods.*

This prescription is for the adhoc task and for track ones that are similar to it. The extent to which prescriptions are possible for the tracks is limited because they are less well developed.

It is essential to emphasise how important, in justifying this prescription, the TREC data scale has been. Though some retrieval strategies found valuable in TREC were first suggested forty years ago, one of the major TREC contributions has been to establish them not only through many individual tests but with very large files and with full text data. However TREC has done far more than this: it has stimulated much new work both in varying old ideas and trying new ones. The possibilities are, moreover, far from exhausted. For instance, the fact that large text files, and accumulating past queries and assessments, constitute a rich data resource that can be mined to improve and tailor indexing and searching offers significant new avenues to explore.

The second concluding point is whether, following the startup-stage of TREC-1 and -2, and the expansion from TREC-3 - 6, it would be rational to modify the generic character of TREC. In the immediate context of the view taken of TREC in this paper, there are three fairly obvious possibilities to explore. First, to concentrate effort on fewer lines (including adhoc as a line) and to pursue these in more depth, perhaps indeed allowing for evaluation tests less frequently than in every year. Second, to address 'realism' by moving at least some of the lines nearer to users and taking more explicit account of context factors, i.e. environment variables. Third, in doing this, to consider other performance criteria and measures, e.g. time in manual searching. In the automated systems world, the TREC model of hands-off IR is very attractive, but it is necessary to recognise that users have necessary roles, and useful contributions to make, in any retrieval situation, whether adhoc or, for instance, filtering. In addition, taking a wider view, it is perhaps time to exploit the experience with basic retrieval gained in TREC to embark on multiple-task studies, where retrieval is just one component of s system that also offers, for example, automatically-produced summaries of retrieved texts.

However there is a larger and more interesting question [3] about the generic character of TREC to consider: this is what the global form of TREC as a (set of) evaluations is.

The 'grid' issue

The dominant paradigm before TREC, or at least the evaluation style to which many tests aspired, was that of *grid experimentation*: i.e. engaging in a series of runs against changes in either environment variable values or system parameter settings (Sparck Jones and Galliers, 1996) This paper has assumed that doing things in grid style was and has remained a TREC desideratum, and has presented its analysis and critique (e.g in 4.2 above) from this point of view.

A great deal of TREC work has indeed been of this form, even if the grid is somewhat loosely drawn; and some particular track and individual team studies have been tight grid investigations. As against this there are powerful influences that appear to have led to a slow movement in TREC away from the grid paradigm as its operational base. These include participants' natural instincts to pursue what works rather than why; the implications of the very free-ranging manual query

---

[3]and I am grateful to Donna Harman for pushing the point

construction allowed for Adhoc; the sheer effort of systematic comparative testing; and (perhaps) the feeling that we know enough not to have to bother with fine-grained validating comparisons.

It is certainly the case that enthusiasm for participation in TREC remains high (as shown by TREC-7), and that while the 'can-do' technological spirit is strong, the level of methodological sophistication and commitment to the 'rational science' grid approach in individual team work is also high. In addition, we can expect the use of TREC materials as test collections in future research outside TREC to help to fill grid cells. It may thus not be necessary to direct each TREC cycle through constraining specifications aimed at promoting grid-style comparisons: such direction could indeed make TREC less attractive to the entrepreneurial spirit among its researchers that has kept TREC renewing itself. The key point, however, is that TREC is so large as a research activity that the way it assesses itself, and evolves, in the future are of great importance to the field.

Acknowledgement

References

[1] Belkin, N.J., Kantor, P., Cool, C. and Quatrain, R. (1994). 'Combining evidence for information retrieval'. *The Second ... (TREC-2)*, 35-43.

[2] Buckley, C., Singhal, A. and Mitra, M. (1997). 'Using query zoning and correlation within SMART: TREC-5'. *The Fifth ... (TREC-5)*, 105-118.

[3] Harman, D.K. Overview chapters in TREC-1 - TREC-4, *The First/ ... /TREC-4)*.

[4] Harman, D.K. (1996). 'Overview of the Fourth Text REtrieval Conference (TREC-4). *The Fourth ... (TREC-4)*, 1-24.

[5] Hawking, D. and Thistlethwaite, P. (1998). 'Overview of TREC-6 very large collection track'. *The Sixth ... (TREC-6)*, 93-105.

[6] Lewis, D. (1997). 'The TREC-5 filtering track'. *The Fifth ... (TREC-5)*,75-96.

[7] Sparck Jones, K. (1995). 'Reflections on TREC'. *Information Processing and Management*, 32 (3), 291-314.

[8] Sparck Jones, K. and Galliers, J.R. (1996). *Evaluating natural language processing systems.* Lecture Notes in Artificial Intelligence 1083, Berlin: Springer-Verlag.

[9] Tague-Sutcliffe, J. and Blustein, J. (1995). 'A statistical analysis of the TREC-3 data'. *The Third ... (TREC-3)*, 385-398.

[10] TREC-1 - TREC-6, Proceedings (1993-1998). *The First/ ... /Sixth Text REtrieval Conference (TREC-1/ ... /TREC-6)*. Ed. D.K. Harman (TREC-1 - TREC-4), E.M. Voorhees and D.K. Harman (TREC-5, TREC-6), Special Publications SP 500-207, 500-215, 500-225, 500-235, 500-238, 500-240 National Institute of Standards and Technology, Gaithersburg MD.

[11] Voorhees, E. (1998). 'Variations in relevance judgements and the measurement of retrieval effectiveness'. *SIGIR-98, Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 315-323.

[12] Voorhees, E. and Harman, D.K. (1997). 'Overview of the Fifth Text REtrieval Conference (TREC-5)'. *The Fifth ...(TREC-5)*, 1-28.

[13] Zobel, J. (1998). 'How reliable are the results of lange-scale information retrieval experiments?' *SIGIR-98, Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 307-314.

```
                            T-1   T-2   T-3   T-4   T-5   T-6

Full (cat A) adhoc task      12    22    22    19    21    29

All ad hoc                   18    24    26    23    29    31

All adhoc + routing          23    30    32    29    34    39

All adhoc + routing + tracks                   33    36    51

    6 teams have participated in every TREC
```

Table 1: Numbers of teams participating in TREC

	T-1	T-2	T-3	T-4	T-5	T-6
Routing	17	25	24	15	16	21
Multilingual Spanish				10	7	–
Multilingual Chinese					9	12
Interactive				5	2	9
Database Merge				3	3	–
Confusion				4	5	–
Filtering				4	7	10
Natural Language Processing					4	2
Cross Language						13
High Precision						5
Spoken Document Retrieval						13
Very Large Corpus						7

Table 2: Numbers of teams participating in TREC Routing task and tracks

	TREC adhoc	TREC tot		VLC	VLC tot	Total
News						
AP (wires)	242,918					
WSJ (papers)	173,252					
SJM	90,257					
FT	210,158			202,433		
LAT	131,896					
		848,481				
			GH	135,477		
					337,910	
						1,186,391
Gov Publs						
FR	101,450	101,450				
			AAG	561,566		
			ADIR	42,481		
					604,047	
						705,497
Plty Procs						
CR	27,922	27,922				
			APLT	421,681	421,681	
						449,603
Tech Arts						
ZIFF	293,121	293,121				293,121
Tech Abs						
DOE	226,087	226,087				226,087
Patents						
PATENTS	6,711	6,711				6,711
Books						
			PGUT	3, 303	3,303	3,303
Net News						
			NEWS	4,400,657	4,400,657	4,400,657
Web Pages						
			AUNI	81,334		
			WEB01	8,531		
					89,865	89,865

AP	Associated Press		CR	Congressional Record
WSJ	Wall Street Journal		APLT	Australian Parliament
SJM	San Jose Mercury		ZIFF	Computer Select (Ziff-Davis)
FT	Financial Times		DOE	Department of Energy
LAT	Los Angeles Times		PGUT	Project Gutenberg
GH	Glasgow Herald		NEWS	USENET News
FR	Federal Register		AUNI	Australian University Web sites
AAG	Australian Attorney-General's Department		WEB01	Miscellaneous Australian Web sites
ADIR	Australian Department of Industrial Relations			

Table 3: Document types used for TREC-1 - 6 Adhoc and TREC-6 VLC track

```
loc     colln          numdocs     size        totdocs     totsize

D1      AP 89           84,678      254.2 MB
        DOE            226,087      183.8
        FR 89           25,960      259.7
        WSJ 87-89       98,732      266.6
        ZIFF (1)        75,180      242.3
                                                510,637    1,206.6 GB
D2      AP 88           79,919      237.2
        FR 88           19,860      209.2
        WSJ 90-92       74,520      241.9
        ZIFF (2)        56,920      175.5
                                                231,219      863.8
D3      AP 90           78,321      237.5
        PATENTS 93       6,711      242.6
        SJM 91          90,257      286.8
        ZIFF (3)       161,021      344.8
                                                336.310    1,111.7
D4      CR 93           27,922      235.4
        FR 94           55,630      394.9
        FT 91-94       210,158      564.1
                                                293,710    1,194.4
D5      FBIS Pt1       130,471      470.2
        LATIMES 89-90  131,896      475.3
                                                262,367      945.5
DAT1    NEWS01         446,106      954.5
        NEWS02         450,027      943.1
        NEWS03         482,395      936.6
        NEWS04         483,145      966.0
                                              1,861,673    3,800.2
DAT2    AAG            561,566     1874.5
        ADIR            42,841      775.0
        NEWS05         590,202     1169.7
        PGUT             3,303      430.3
                                              1,197,912    4,249.5
DAT3    APLT           421,681     1539.8
        AUNI            81,334      724.8
        GH 95-97       135,477      393.6
        NEWS06         571,891     1120.6
                                              1,210,383    3,778.8
DAT4    FT 88-90       202,433      526.7
        NEWS07         520,282     1080.1
        NEWS08         856,609     1727.9
        WEB01            8,513      141.9
                                              1,587,837    3,476.6

                                              7,492,048   20,627.1
```

Table 4: Sizes of document sets, TREC-1 - 6 Adhoc, TREC-6 VLC track

```
          Adhoc                                      Routing

          Document set        no docs    Topic set   Documents     Topics

          D1  D2  D3  D4  D5

TREC-1    X   X                742 K      51-100      D2            51-100

TREC-2    X   X                742 K      101-150     D3            51-100

TREC-3    X   X                742 K      151-200     D3            101-150

TREC-4        X   X            568 K      201-250     special 1     subset 1

TREC-5        X       X        525 K      251-300     special 2     subset 2

TREC-6                X   X    557 K      301-350     special 3     subset 2a


  Routing documents special 1 = subset FR
                               subset ZIFF + 'nettrash'
                         2 = set of FBIS
                         3 = set of FBIS
        topics all subsets of standard; subset 2a almost same as subset 2
```

Table 5: Details of Adhoc and Routing test sets, TREC-1 - 6

```
                  TREC-4                  TREC-5               TREC-6
        -----------------------------------------------------------------------
Adhoc      D   568 K                  525 K                557 K
           T   49                     50                   50
        -----------------------------------------------------------------------
Routing    D   FR 94 +                FBIS Pt1             FBIS Pt 2
               ZIFF (3) + nettrash
               330 K                  131 K                121 K
           T   25 for FR +            39 old for FBIS      38 T-5 + 9 new
               25 for ZIFF etc
        -----------------------------------------------------------------------
Mling Span D   El Norte newspaper     Agence Fr Pr wires
               58 K                   173 K
           T   25 new (D only)        25 new
     Chin  D   -                      Peking Daily Newsppr  as T-5
                                      165 K
           T   -                      19 new                26 new

Interact   D   T-4 adhoc              T-5 adhoc            FT 91-94
           T   25 subset adhoc        12 subset adhoc      6 modif adhoc

DB Merge   D   adhoc (divided)        adhoc (divided)      -
           T   adhoc                  adhoc

Confusion  D   adhoc B (1/4 full)     FR 94                -
               ~ 142 K                56 K
           T   adhoc                  50 new k-item

Filtering  D   routing                routing              routing
           T   routing                49 routing           38 T-5 + 9 new

NLP        D   -                      T-4 adhoc B          T-5 adhoc B
                                       = WSJ 90-92          = FT 91-94
                                      75 K                 211 K
           T   -                      45 adhoc             47 adhoc

Cross Lang D   -                      -                    Eng AP wires
                                                           243 K
                                                           G NZurZ paper
                                                             SudDA wires
                                                           252 K
                                                           Fr SudDA
                                                           142 K
           T                                               25 new

High Prec  D   -                      -                    adhoc
           T                                               adhoc

SDR        D   -                      -                    1451 news
           T                                               49 new k-item

VLC        D   -                      -                    7493 K
           T                                               adhoc
```

Notes : adhoc, routing without detail refers to concurrent year's set
        topics are of title+description+narrative form except where shown;
         but they may vary in length, and the status of the title varies

Table 6: Track datasets, TREC-4 - 6

```
              topic                      query                    runs      notes
              title descr narr  conc     auto  man  feedback
                                                    auto/man

TREC-1   X    X     X    X       X    X   X            2

TREC-2   X    X     X    X       X    X   X            2

TREC-3   X    X     X            X    X   X            2

TREC-4        X    (X)           X    X                2
                                         feedback
                                         man only
TREC-5   X    X     X            X        -X-          1a2m
              X                  X                     1a        oblig auto

TREC-6   X    X     X            X        -X-          1a2m
         X                       X                     1a        opt auto
              X                  X                     1a        oblig auto



              title = 'very short'       (V)
              descr = 'short'            (S)
              title+descr+narr = 'long'  (L)



      average topic and field length

              total     title    descr    narr     conc

TREC-1   107.4      3.8      17.9     64.5     21.2
TREC-2   130.8      4.9      18.7     78.8     28.5
TREC-3   103.4      6.5      22.3     74.6
TREC-4    16.3               16.3
TREC-5    82.7      3.8      15.7     63.2
TREC-6    88.4      2.7      20.4     65.3
```

Table 7: Details of Adhoc topics and run specifications, TREC-1 - 6

```
                   TREC-4            TREC-5            TREC-6
---------------------------------------------------------------------
Adhoc              PRec level        PRec level        PRec level
                   PDoc cutoff       PDoc cutoff       PDoc cutoff
                   AveP              AveP              AveP
                   R-Prec            R-Prec            R-Prec
---------------------------------------------------------------------


Routing            adhoc             adhoc             adhoc

Multiling Spanish  adhoc             adhoc             adhoc

         Chinese   -                 adhoc             adhoc

Interact           no rel in time    aspectual R       aspectual R
                    P,R set based    P                 P
                   best Q,fr ranks   search time       search time

DB Merge           adhoc             adhoc             -

Confusion          adhoc             raw rank          -
                                     mean rank
                                     mean reciprocal

Filtering          utility on pool   utility on pool   utility on pool
                   utility on sample utility on sample utility on sample
                   mean P, R         mean P, R         ave set P = P*R

NLP                adhoc             adhoc             adhoc

Cross Lang         adhoc             adhoc             adhoc

SDR                -                 -                 mean rank
                                                       mean reciprocal

VLC                -                 -                 P at Doc 20
                                                       time for user
```

Table 8: Track performance measures, TREC-4 - 6

```
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        Cornell
       -HNC         -Mead
       -VT

 >= 50  Cornell     -Verity
        Berkeley    -VT
        Dortmund    Westlaw
       -CMU/Clarit  ETH
       -Verity      CUNY
       -Siemens
        CUNY

 >= 45 -City        NYU          -Excalibur/     ==> (best TREC-5 level)
        Bellcore    CMU/Clarit     Conquest
        ETH         RMIT         -CUNY
        CITRI/RMIT -RutgersK     -Waterloo
       -Conquest

 >= 40   ...          ...         -Berkeley
                                  -Clarit/CMU
                                   Cornell
                                  -GMU
                                  -UMass
                                  -InText
                                  -ANU

 >= 35   ...          ...          City
                                  -GE/NYU

 >= 30   ...          ...           ...

 >= 25   ...          ...           ...

 >= 20   ...          ...           ...          ==>


 NOTES : - marks teams using manual queries

        TREC-2 HNC used feedback on otherwise automatic queries
               Siemens offically labelled automatic, in fact manual
        TREC-3 CMU/Clarit manual query performance was lower than
                 automatic
               City ditto
```

Table 9: Adhoc retrieval performance, TREC-2 - TREC-4

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			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

Table 10: Adhoc retrieval performance, TREC-5 - TREC-6

```
Devices and  elaborations, TREC-6

Devices :  Automatic searching
  Long topics

                  phrase      passage     passage     blindfeed  expansion
                              scoring     bounding

    ANU           X                       X           X          X
    Cornell       X                       X           X          X
    IRIT                                              X          X
    CUNY          X                                   X          X
    Berkeley      X

 (IRIT, CUNY and Berkeley are systems with heavy learning)


  Short topics

    ATT                                               X          X
    ANU           X                       X           X          X
    City                      X                       X          X
    Cornell       X                       X           X          X
    GMUetc                                            X          X
    IBMTJWs       X                                   X          X
    IRIT                                              X          X
    Lexis         X                                   X          X
    Waterloo      X                       X           X          X


  Very short topics

    Apple                     (no workbook paper)
    ATT                                               X          X
    City                      X                       X          X
    IRIT                                              X          X
    Lexis         X                                   X          X
    CUNY          X                                   X          X
    Waterloo      X                       X           X          X


Elaborations : Manual searching

                  modest      large       expansion   human
                  effort      effort                  assessment

    Waterloo                  X           X           X
    Clarit        X                       X           X
    ANU                       X           X           X
    GE/NYU        X                       X           X
    Lexis                     X           X           X
    ISS                       X           X           X
    Berkeley      X?
    FS                        X           X
    GMUetc                    X           X           X
    Glasgow                   X
```

Table 11: Search devices and elaborations, TREC-6

```
ROUTING  -  DOCUMENT CUTOFF 30

        TREC-2        TREC-3        TREC-4        TREC-5        TREC-6


>= 60  Cornell       City
       Dortmund


>= 55  City          UMass         City                        ATT
       Berkeley      Cornell       UMass
       UMass         Berkeley      Xerox
       Bellcore      Dortmund
       CMU/Clarit    Bellcore
       CUNY


>= 50  Rutgers       CMU/Clarit    Cornell       City          City
       HNC           Westlaw       CUNY          Cornell       Cornell
       GE            Logicon                     UMass         CUNY
       TRW           TRW
       Verity        Florida
       Siemens


>= 45  VT            Xerox         Logicon       CUNY          Clarit
                     NYU           GE/NYU                      IRIT
                     Verity                                    ETH
                     ETH                                       Waterloo
                     NSA
                     NEC


>= 40  ...           ...           ...           GE/NYU        SRI
                                                 ETH           Berkeley
                                                 Berkeley      UCSD
                                                               UMass
                                                               Verity
```

Table 12: Routing retrieval performance, TREC-2 - TREC-6