# Reflections on TREC

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street
Cambridge CB2 3QG, England
sparckjones@cl.cam.ak.uk

July 1994

**Abstract**

This paper discusses the Text REtrieval Conferences (TREC) programme as a major enterprise in information retrieval research. It reviews its structure as an evaluation exercise, characterises the methods of indexing and retrieval being tested within it in terms of the approaches to system performance factors these represent; analyses the test results for solid, overall conclusions that can be drawn from them; and, in the light of the particular features of the test data, assesses TREC both for generally-applicable findings that emerge from it and for directions it offers for future research.

This paper will appear in *Information Processing and Management*.

# 1 Introduction

TREC is a new ballgame for IR research and development. It is richly interesting in itself, can be expected to have many consequences for IR study and practice, and should be noticed and built on by everyone interested in automatic indexing and retrieval. We can learn both from what it has done and what it has not done, and thus focus future research.

My purpose here is to review TREC as a whole, considering it both in its own right and in the context of IR R&D over time. I shall, further, examine TREC both at the object level of its results for indexing and retrieval, and at the meta level of test methodology in IR. This paper, as a review, is intended to complement Harman's factual accounts in Harman(1994a, 1994b) and has to be supplemented by the individual papers, the questionnaire responses and, most obviously, by the tabulated test results given in the TREC Proceedings (Harman 1993, Harman 1994a). Thus my aim is to draw some general conclusions about the significant findings in TREC, and about the directions to pursue in the future: What have we found out? Where should we go next?

*But I must emphasise that this review is a personal one constituting a first attempt at an assessment of a large and continuing enterprise. In particular, in trying to draw conclusions about what the work to date has shown I have been forced to simplify, so my judgements may be crude and controversial. My presentation of results and findings should thus be seen as a snapshot, successful if leading not only to further, closer analysis of the existing materials, but also to further, more discriminating tests.*

## 1.1 Paper perspectives and organisation

It is possible to assess TREC from several different points of view. I shall start by noting my frame of reference for the TREC *programme* as an officially-organised activity, and indicate what I see as the major relevant features of the TREC *enterprise* as whole. The TREC programme has specific characteristics from the IR point of view, constituting what I call the TREC *case*, both providing opportunities for and imposing constraints on retrieval itself. Thus TREC has its particular motivation, leading to choices of test data and evaluation strategy with material effects on the inferences to be drawn from the TREC findings. While some of the object-level TREC data characteristics are broad ones that would also hold in other applications, some are more particular so their external implications may be less clear. The actual retrieval results obtained in the TREC tests have thus to be correspondingly carefully assessed for their wider application. Again, the TREC programme involves, at the meta level, a strongly-shaped evaluation methodology relying on an evaluation paradigm that has to be judged both for its own soundness and for its pertinence to any wider application of the TREC findings. At the same time TREC has a generic context, in the current state of operational IR practice and IR research, that has to be borne in mind in reviewing TREC. Finally, the TREC *effort* itself represented by the work and tests done by the participants, involving many different teams doing very different things, has not only more obvious implications for the range or solidity of the TREC findings but also other interesting features. Thus it is possible not only to ask what's old about TREC, in the sense of what past findings it endorses and helps sigificantly to establish, but also what's new about TREC, in its concerns as well as its findings.

In my analysis of the TREC effort and discussion of findings I shall make use both of the well-established notion of *performance factors* affecting IR, and of the *environment / system* contrast and the test methodology to which it leads, involving a decomposition generating a *grid* design with cells representing *runs* each for some request set and document set (cf Galliers and Sparck Jones (1993)). The environment/system contrast distinguishes factors according to whether they are environment *variables* with *values* that are fixed for any given set of runs for which system *parameter settings* are altered or, alternatively, for which system parameter settings are held constant through environment changes. As this suggests, the implications of this methodology are that reliable and informative tests require modifications to factors in both grid dimensions that are also subject, in a third dimension reflecting a kind of evaluation meta environment, to the use of different performance *gauges*.

My review is organised as follows. The next section, 2, describes the TREC case, supplying the background for the subsequent description and assessment of the research done. My descriptive analysis of the TREC work first lays out the performance factors that have been investigated, in Section 3, and then, in Section 4, summarises the actual test *results* from this point of view, using a series of factor questions to structure the analysis. In Section 5 I present what I see as the main *findings* that emerge from the work when the experimental results are assessed in the context of the TREC data and evaluation framework, and in Section 6 draw my conclusions about the future IR work that needs to be done.

## 1.2 Terms of reference

I shall use the term 'TREC' as a global label for the succession of Text REtrieval Conferences, each covering a preparation and test *phase* terminating in a reporting workshop. TREC-1 was in 1992, TREC-2 in 1993, TREC-3 is underway with its meeting in November 1994, and TREC 4 is already being planned. TREC meetings are essentially restricted to participants, but there are published Proceedings (Harman 1993, Harman 1994a). As already indicated, it is convenient under the TREC umbrella term to distinguish the official organisation of the TREC programme from the actual work effort put in by participants, and also to recognise that something that can be called the TREC enterprise is developing that subsumes not only programme and work but a range of supporting or derivative activities. SMART at Cornell and elsewhere is an earlier analogue, but as will be considered more fully later, the scale of the TREC tests, size and heterogeneity of the TREC community, and the fact that the teams are working concurrently in a short time scale, mean TREC has special properties and hence effects on the field.

As TREC-1 was so novel, while the time available for TREC-2 was too short for extended experiment, it is sensible to treat TREC-1 as a debugging phase with consolidation in TREC-2. I shall therefore take TREC-2 as representative of the work done and progress made so far and thus, unless explicitly indicating otherwise, be referring in using the term "TREC" to the tests done in TREC-2 and reported in the TREC-2 Proceedings.

# 2 The TREC case

While TREC falls within a well-established experimental IR tradition, there is novelty in the stringency with which these experimental conventions are being applied, in the nature and scale of the test material, and as just mentioned, in the size of the participating community. So though the TREC operational details are fully described in Harman (1994b), I shall briefly summarise the main features of the enterprise for convenient reference here, and comment on points that are particularly important for my attempt to assess the TREC contribution to IR. This is also desirable because, while the TREC meetings are restricted, all the materials, both data and results, will be made publicly available and may - indeed ought to - be exploited in further research building on the TREC work, just as test collections have been in the past. These new materials are, however, much richer than the old ones, and also have the great advantage first, of being in full machine-readable form and second, of including output search results as well as the input test data.

I shall characterise the TREC case with respect to its *motivation*, *test data*, *evaluation paradigm* and *generic context*.

## 2.1 TREC motivation

It is important in assessing TREC, to take account of its sponsoring motivation and situation. The stimulus for TREC has come from US government organisations concerned with intelligence operations and R&D sponsorship, notably (D)ARPA, responding on the one hand to the vast growth in communications (parties, events and matter) and on the other to the development of computing, language and information processing technologies that make it possible to attack real world rather than laboratory problems. ARPA has been promoting competitive task programmes since the mid eighties, first on speech recognition (SR), subsequently in spoken language systems (SLS) and in message understanding (MUC), and more recently in machine translation (MT) and combined message retrieval and information extraction (TIPSTER - subsuming TREC and MUC type tasks). (For a convenient overview see the materials in Bates (1993).)

This background is important because it has established a style of operation coloured by the properties of the initial SR case and affecting the implicit goals and actual data for the TREC case. Thus while the IR community has its convenional evaluation methodologies, unlike the NLP community which has had to make a major effort to develop evaluation strategies, the background to TREC leads to a pressure to perform, and specifically to win, in a fairly crude way that is not necessarily beneficial either to the TREC sponsors' own long-term goals or to the field as a whole. The SR case is especially dangerous because performance there can be be simply but uncontroversially defined using single number measures. The established IR mainstream evaluation style, in fact illustrated by TREC, naturally fits the TREC programme *zeitgeist*, but nevertheless has its limitations, as considered later.

## 2.2 TREC data

The salient, relevant points about this are as follows (for more detail see Harman (1994b)). The document set is large, over 1 million items, though it is made up of three large blocks which have been used separately and have not been treated as a single whole. The material is from seven sources, individually varied and collectively various in topics and genres, though with much news story material. In general there are documents from each source in each block, though the material from any source is not necessarily a full or correctly ordered set. The documents are predominantly full text, but some sources do not have rich headers of the kind familiar from journal articles, or even any significant headers at all.

The request set totals 150, but has only been used as subsets. The TREC requests, or *topics* in TREC parlance, have very distinctive properties that need to be taken into account in any attempt to draw conclusions for wider application from the TREC tests. From this point of view it is also useful to bear in mind the distinction between a user's information *need*, a literature *request*, and a search *query*. Information needs are often not fully or explicitly expressed, especially in writing; and historically there has been a shift from the old-fashioned case where an initial informal natural language request is replaced by a formal boolean query, perhaps using controlled language terms, to the case where request and query are much closer because the latter is a simple list of terms derived from the former that is applied in searching to deliver a ranked output. In accordance with common IR convention, and also to avoid confusion with other uses of the word "topic", I shall normally refer to the TREC topics as requests. However as the example shown in Figure 1 makes clear, they are long and detailed, and much more elaborate than requests usually are, with elements in them of both need statement and query. The TREC requests have been very carefully formulated, even honed, and have a complex structure with several distinct fields. The Narrative field, in particular, clearly indicates what properties documents must have in order to be deemed relevant, thus reflecting the user's needs; and the Concepts field contains a set of chosen, often compound terms, already constituting a 'proto-query'. At the same time, there are 'implicit' category terms like "US company" in Narratives that are a problem for indexing, and proper names that are a challenge for matching. Thus the TREC requests look like standing interest, or routing, profiles, and were indeed designed with this task in mind: in the TREC tests they have however been used both for routing and for one-off or 'ad hoc' searches. Finally, as the foregoing suggests, the TREC requests were provided by professionals, experienced in request formulation and, further, by users whose primary task is to do information analysis, i.e. are not typical of many user communities.

With respect to relevance needs in TREC, one important point is that documents do not have to be wholly, or even primarily, about a request topic: a document may be relevant given only the facts conveyed in two independent sentences. The implication is also, though this is not stated but follows from the routing background, that high recall is the normal requirement.

The relevance assessments are on the combined search output for all the participants' official submitted runs. The official specification limits the number of runs per participant, and sets a standard limit on output size; at the same time, there are many participants applying very different approaches; the *relevance pool* is therefore likely to be less strongly or arbitrarily biassed than for many collections in the past and, since the various outputs retrieved different documents, is likely to have pulled in many, even most, of the relevant documents. The number of documents assessed for relevance per request was large; the assessment was not done by the requester, but by similar professionals. (In the rest of this paper *relevant document* always refers to those so known from assessing the pool.)

Reviewing the TREC data from the point of view of *test collections*, the major points are as follows.

First, it is actually a *family* of related collections: the evaluation design described below has meant that tests have been done on various combinations of document and request subsets, with a document subset defined as one of the three blocks D1, D2 or D3, and with a request subset one of the three subsets numbered 1-50, 51-100, or 101-150. The subsets have been set up over time, but though some sources are not represented in each document subset and neither document nor request subsets have been drawn as systematic samples from prior wholes, they are similar enough or at any rate sufficiently typical of collection growth, to be taken together. This is important because while the document sets are large, the request subsets are small and only taken together form a fair set, so caution may needed in drawing inferences from subset performance. For convenient reference in this paper I shall refer to whole body of material as the *TREC Collection*, and use *TREC collections* when the division into subsets is relevant.

Second, as a test collection, the TREC Collection is strikingly large, both absolutely and by comparison with those previously used for IR research. This refers not only to the number of documents, in particular, but also to the fact that they are largely full text ones. A test collection of this size constitutes a completely new test environment for the field.

Third, the Collection is a constructed rather than natural (or proper sample) one. This description applies in various ways. It applies to the aggregation of material from very different sources, not all of which might be searched directly with requests of the type used but rather excluded via checks on database descriptions. It also applies to the requests. These were specially developed for the TREC programme and were not drawn from ordinary retrieval operations, though they were intended to be like those used in practice. They were also made by only a small number of experts, so the number of requests does not represent the same, or even a large number, of users. Finally, the relevance assessments were not made by the request providers, and the data gathering design for these also meant that there is no reflection in the assessments of the number or choice of documents that actual users would have identified in practice. Of course the important other requirements for relevance data for test collections that lead to pooling and obtaining as many assessed documents as possible are themselves an unavoidable source of artificiality; but the lack of reference to original requester's judgements makes the TREC Collection like former collections where proxies were used, even if the requests themselves provided unusually detailed guidance for the assessment.

The difficulties of forming truly 'real' test collections are well known, and relevance data in particular have to satisfy fundamentally conflicting demands. The points just made are not intended as criticisms: the TREC data provision has been a heroic effort, thoughtfully grounded and carefully executed, and drawing on past IR test experience. My point here has only been to note salient features of the Collection possibly bearing on future conclusions from the tests. Moreover, while the TREC Collection is perhaps most accurately defined as semi-real rather than real, some of its properties, for instance the document heterogeneity, may be indicative of future search conditions and also constitute so taxing a test environment that good results obtained in it will carry over elsewhere. Overall, the TREC requests are the most critical element in the test data: the fact that they are rich in content, carefully designed, and elaborately structured; are intended for document routing; and come from information professionals, clearly has implications for inferences from TREC. Thus it is important to consider the extent to which performance results in the tests reflect generalised *tailoring* to the type of data found in TREC, notably the full-text documents, or more specific *tuning* based on the particular properties of the test material, notably the requests.

Equally, while it is certainly helpful to have large relevance sets based on exntensive pooling, it would be desirable to have some study of the nature and degree of bias there may be in the relevance sets.

## 2.3 Evaluation paradigm

The evaluation paradigm being used subsumes both what the IR task is taken to be, and hence what *assessment gauges* are used, and the test organisation. In terms of the analysis of system evaluation given in (Galliers and Sparck Jones 1993), the TREC tests have a *remit* determining the test goal, and a *design* specifying gauges, data, and procedures, with remit and design together contituting the evaluation *scenario*. The goal stems partly from the TREC motivation, but meshes with conventional approaches to IR testing done outside the real working environment of users and operational system. Thus the goal is taken, strictly, just to be capturing relevant documents, which leads in a very natural way to the use of *recall (R)* and *precision (P)* as evaluation *criteria* with specific performance *measures* using output cutoffs and application *methods* for these measures averaging across values for individual requests. The measures give $P$ at standard recall levels, and at standard document output size (rank) levels, from which single number average $P$ and single number $P$ for total relevant can be respectively derived. (Values for *fallout* are also computed, though these have not been accorded much attention.) The general approach is that common in the field, and popularised by the SMART Project, the different measures reflecting slightly different views of $R$ and $P$; but document cutoff is better suited to intuitive interpretation and more realistic: $P$ performance for very high $R$ is not normally of practical interest. There are also well-known problems in assessing iterative search performance: in TREC, where comparisons across participating teams cannot be tightly controlled, results have been computed using the frozen ranks method.

The test design has two distinctive and especially important features. The first is the way requests have been designated as *routing* or as *adhoc*, and the closely related way in which request and document subsets have been combined for tests. The requests were, as mentioned earlier, designed for routing. Routing normally implies allocating documents from an incoming stream to members of a user community. But searches for these requests were in fact done on static document sets as wholes, thus changing the IR situation, with implications for performance conditions and assessment: the sense in which the requests were properly treated as routing ones was in the way relevance information for searches on earlier document sets provided *training sets* that could be used to develop and modify queries for application to new *test sets*. At the same time, the requests were put to double use since they were also treated as adhoc, for one-off searches of an existing document collection, though this was not the purpose for which they were designed or would normally be put. This is not to suggest that a single request may not be legitimately put to very different search uses, but in assessing TREC it has to be borne in mind that the adhoc requests might not be typical, because of being developed with the care and attention that routing requests can justify.

Thus in the TREC evaluation, the same requests are used either in adhoc or in routing *mode*, according to whether relevance information from *previous* document sets is available for building queries from them. This dual-purpose modus operandi is associated with the elaborate succession structure for tests and their definition in terms of individual test collections. In general, for some specific query set, this can be applied to some new document set in *adhoc mode*, for adhoc performance evaluation. This set, together with the relevance assessments made for the adhoc test, can then be taken as training data for the same queries, for future application in *routing mode*. This routing application is therefore of old queries to a new document set, to which a new set of queries is also applied in adhoc mode. The evaluation strategy for the TREC programme as a whole thus involves a kind of rolling procedure over related collections, with the two-mode operation, in a sense, enlarging the test collection.

The other important feature of the evaluation paradigm, at least as administered within the TREC programme and so constraining its participants, was the limit placed in the number of official *submitted results* per phase: no more than two routing runs and two adhoc runs per team. This has obvious and immediate advantages for comparative purposes and generally telling the wood from the trees. It has of course not stopped participants doing and publishing other runs, and the submitted results should emphatically not be taken as all there is to say about the IR approaches they represent. But it is important to recognise that as TREC is a multi-party enterprise, its evaluation paradigm has this wider manifestation that is relevant to assessing its results. (Other aspects of the TREC programme methodology, like the questionnaires, that bear on research practice rather than evaluation per se, are considered later.)

## 2.4 Generic context

The TREC programme and enterprise has naturally, like all the (D)ARPA evaluations, evolved a house style. This in part follows from its motivation and evaluation paradigm, but there is also what may be called its generic retrieval system context to take into account, as there are constraints or presuppositions associated with this that may influence the approaches adopted by the participating teams. This *generic context* is not only, and obviously, that indexing and retrieval are essentially, even if not wholly, automatic, but also that users when receiving matches for routing requests or engaged with adhoc ones will be interacting with a computational system. Though the nature of the interface itself has not been considered at all, and tests with humans doing interactive searches have so far been very limited, there is an underlying presupposition that computational interaction is the norm. This may indeed cover the entire use of the full text, and not just preliminary or superficial scanning. (Both the interactive tests and also the relevance assessment have required interfaces with some of the properties that would be needed for end users in normal circumstances, but this has not been the prime design consideration in either case.)

The computational context in itself implies both opportunities and restrictions for indexing and retrieval. But the size of the document set, the fact that in the real case it would grow and change, perhaps quite drastically, and the provision of full text, jointly define a generic retrieval context that has not been extensively analysed in evaluations to date, but has important consequences. While these do not always apply in a strictly routing situation, they can do, and as TREC itself, and others wishing to exploit its findings are not so confined, the computational features of the TREC case deserve notice. They imply

1. It is foolish to rely primarily on any kind of relatively fixed, controlled, indexing language.

2. But it is also dangerous to assume that any indexing vocabulary, like a set of term classes, depending on properties of the document set as a whole can be reformulated at intervals.

3. Even if it is feasible, it is not going to be worth applying elaborate (pre)processing to each document designed to provide a full and complete index description, especially as an autonomous surrogate replacing the source text.

4. At any rate, it is not sensible to assume that elaborate initial descriptions can be reformulated, for entire files, to suit changing document or request file conditions.

5. It will be too hard, or too costly, to construct sufficiently powerful explicit Boolean queries for effective retrieval (apart, say, from long-standing routing profiles).

6. But it is essential to supply resources and mechanisms for request formulation and query development.

The consequences of these points taken together are first, for indexing and retrieval philosophy and second, for operational system design. Thus with respect to retrieval philisophy, there is a presumption that natural language is straightforwardly the language of indexing; that indexing is request-based: documents are indexed via their query matches; that output ranking allows discrimination and flexibility; and, therefore, that the focus of attack in retrieval should be on the development or modification of source requests and the actual search queries derived from them. The prime questions are thus what form of natural language indexing, and of query manipulation, is either generally optimal or is best for given conditions.

From the operational point of view, the issues are first, what forms of information have to be gathered at document file time and what implications search uses have for how this information is stored and accessed; and second, what information resources, and display facilities, query development requires.

While this generic context seems to be largely taken for granted by TREC participants, the TREC effort has included some explicit investigations of practical questions relevant to document file handling; and the TREC enterprise as a whole can be seen as implicitly testing all the presumptions just listed by showing that while they may imply a somewhat crude or limited approach to indexing and retrieval, this can still deliver reasonable performance, even if it may not be the best. The retrieval style just described also supplies a natural *baseline* for performance in the form of single terms (word stems), with simple weighting using inverse document frequency and perhaps also within-document frequency, and sum or product scoring for matches. The immediate questions then to be answered for a TREC-specific baseline are whether within-query weights are also to be used, and what parts (fields) in the request.

Given this baseline, the challenge is thus to improve on it by moving in any or all of the following directions: using more complex terms; being more selective on terms; exploiting other or additional weighting; utilising supplementary vocabulary resources or query expansion. These possibilities apply to both routing and adhoc requests, from a static perspective. There is also the dynamic perspective allowing learning, referring on the one hand to system *adaptation* over time and applying to all requests e.g. through changes in term associations, and on the other to *feedback* bearing on individual requests: this may be over time in the routing case or through search iteration in the adhoc case, and normally exploits relevance information in *relevance feedback*. The baseline case here may be deemed to cover only the automatic changes in inverse document frequency weights that naturally follow collection changes, and *minimal feedback* relying just on user document assessments to reweight query terms. The corresponding lines to pursue for performance improvement therefore cover first, elaborations of this basic reweighting and second, combinations of learning with the other possibilities just mentioned referring to complex terms, term selection and weighting criteria, supplementary resources; they thus lead to dynamic term classification, for instance and, most importantly, to query alteration by expansion or deletion: this has a whole range of options with more or less decision by the system or user.

Finally, the generic context for TREC makes technology considerations worthy of study in their own right. Though the test environment imposes its own data organisation requirements, for instance high modularity to allow system alternatives, it is essential to investigate data management issues, like file structures, paging etc, because the data bulk imposes operational constraints and the choices for data organisation interact with indexing and retrieval ones to affect not only efficiency but effectiveness.

### 2.5 TREC effort

The final important feature of TREC, in relation to IR R&D, is the sheer scale of the evaluation effort in terms of the number of teams involved and the consequent range of approaches being tested. Since the programme in itself explicitly involves a common treatment of output for common performance measures, the provision of replies to a detailed procedure questionnaire, and workshops at the end of each phase, there is a very rapid dissemination of results to, and interaction within, the participating community, leading to a quick response to results and also, via the active programme committee, to continual development of the whole assessment strategy. From this point of view TREC is not merely a large effort but a de facto cooperation as much as competition, even if the cooperation is indirect through changes in individual approaches rather than through explicit joint work. This has however also occurred, and there is a further, important direct form of community cooperation in the work of the programme committee. All of this has meant that a very large amount of ground has been covered, and much more quickly than in previous IR work, with some consequences for assessing the TREC findings.

In the next section I shall consider the actual work done by the TREC participants, and in the following one summarise their test results; then in assessing the results, I shall try to establish what influence the TREC case had on these and hence on the conclusions to be drawn about needs for further research and for the wider exploitation of the results in operational systems.

## 3 TREC work

The participants in any TREC phase are divided into two categories with Category B participants in 'startup' mode doing smaller scale tests. As in reporting the TREC results in the next section I shall focus on the full official tests, I shall not consider Category B work here other than where it is especially appropriate. I shall also, in this overall review, not single out individual teams for concentrated attention; however as references to teams are needed, I shall cite these by simple institution names, for which a key is provided in Appendix 1. Also, unless specific citations are made to other publications, the references for teams' work are to their contributions to the TREC-2 Proceedings (Harman 1994a).

## 3.1 Performance factors

As noted, the TREC(-2) participants have pursued a whole range of approaches to indexing and retrieval. It is helpful, in reviewing the TREC tests and the lessons to be learnt from them about the different approaches, to characterise these in relation to the evaluation framework elaborated in Galliers and Sparck Jones (1993) and mentioned earlier, where *performance factors* are distinguished either as *environment variables* with *values*, or as *system parameters* with *settings* [1]

*Environment variables*

In TREC the environment for all the participants in any one phase (and category) is the same, being the documents, requests, request mode (routing or adhoc), relevance needs, and relevance data for previous document sets. Thus for each phase there are two environments defined respectively by mode, each with their test collection defined by request set (with relevance needs), document set and (subsequent) assessment set, and also prior test collection now available as training collection. However while the phase environments for a mode are technically distinct they do not, for the reasons mentioned earlier, cover any serious variation in the values of the environment variables: the successive test collections are designed to be like one another. The TREC tests do not therefore involve the significant changes of environment that are properly required to check on the behaviour of system parameters and fill out the test grid cells. Moreover while the mode environments are distinguished, in particular by the supply of relevance information, the fact that the same requests are used, albeit in different set combinations, makes the environments less different than appears. Thus while the main thrust of the TREC tests over time may be taken as intended to establish that systems perform consistently across test collections (setting aside changes to the systems themselves), the collections have not been varied enough to demonstrate this or, as research understanding requires, explain why they perform as they do. It is thus legitimate to characterise TREC, at a coarser level, as having a single, constant environment, and the TREC work so far as not involving any serious attempts to explore the effects of environment changes with e.g. very different types of request.

*System parameters*

The real variation in the TREC tests has come from the differences in system parameter *settings*, both at major and minor levels across and within the different teams' tests. Here TREC has filled in many test grid slots, though these are

[1] "Variable" and "parameter" were used the other way round in my earlier papers on IR tests, such as Sparck Jones and Webster (1979): the present use fits better with other conventions.

somewhat unevenly distributed and are largely arbitrarily related to one another, especially at any one level of detail, with many gaps. The TREC evaluation thus addresses the question: what differences do these variations make, in fact at several levels of granularity, both coarse- and fine-grained, where coarse-grained may be illustrated by the contrast between linguistically- and statistically-defined compound terms and fine-grained by choices of constants in weighting formulae.

Further, as noted earlier, the TREC research enterprise also applies a fixed evaluation context for all participants as defined by the standard evaluation scenario used; so while $R$ and $P$ are natural first performance criteria there has not been any system comparison on other effectiveness criteria and only very limited studies of efficiency.

While it is not appropriate to describe individual TREC approaches and results in detail, a broad categorisation of approaches is needed both to assess the scope of the TREC studies and draw conclusions from the results. The categorisation refers primarily to matters like the type of treatment given to documents, that are clearly performance factors. But it is also useful, for a comprehensive view, to extend the categorisation to include the general form of model underpinning an approach, for example whether an overtly linguistic or statistical view of information is being taken. Thus the choice of *indexing and retrieval model* can be treated as defined by a highest level parameter under which lower level parameters fall into groups concerned with the nature of the *indexing vocabulary*; of the indexing strategy for documents, i.e. the nature of *document descriptions*; of the document *indexing sources* for these descriptions; and, since documents and requests may be treated differently, the nature of request descriptions, i.e. of *queries* and of *query sources* including not only requests but also e.g. relevant documents. There are, further, the *search strategy* and *scoring criteria*, clearly separate because the search procedure may affect what documents are inspected for matching; and *output form*.

In some cases there is a strict dependence between settings for these different parameter groups, in other cases natural or habitual associations, in yet others complete independence. These heterogeneous relations between performance factors as represented across TREC approaches makes it very difficult to assess the general implications of the results for particular approaches or even types of approach, because it may be hard to attribute responsibility for these results to any particular features, i.e. choices of setting for system parameters, of the approaches involved. This is well illustrated, for instance, by Buckley's discussion of weighting (Buckley 1993).

Moreover the classes of performance factor listed earlier are themselves informal, though familiar and to some extent hallowed by tradition. Thus in practice, whether one is dealing with indexing vocabulary or document description may be hard to determine, and it is in any case necessary to distinguish logical characteristics from practical implementations. However the categories allow a first cut across the range of approaches with a view to attributing performance findings to system features.

It is also necessary to recognise *learning* as a high-level system feature. This may be applied in relation to various system elements, but is conveniently treated under its own heading.

In discussing the different approaches adopted in TREC for each parameter, in the material which follows, I shall put parameter questions: these are intended to focus the comparisons between results in the next section, where I shall try to answer them.

## 3.2 Indexing and retrieval models

Characterising the models underlying the TREC approaches is not easy. But it may be done in terms of what criteria are used to choose index terms, i.e. how they are grounded as units, irrespective of whether these are in fact single or multi-word units. From this point of view units consisting either of single words or of multiple words held together only by proximity-based cooccurrence can be labelled statistically motivated, while term units defined using explicit linguistic (syntactic or semantic) constraints, e.g. selecting head nouns, are linguistic. These alternatives may be instantiated either through prior vocabulary definition or at file or search time. By extension, word classes may be (explicitly) statistically or linguistically motivated. This statistical/linguistic distinction is one informing an entire indexing and searching process; however some TREC teams apply their model in a much more aggressive and all-encompassing way than others. In many cases, too, underlying models are diluted, overlaid, or mixed through complex combinations of resources and procedures. The categorisation made here is thus only indicative.

The statistical models represented in TREC include the vector model associated with the SMART system (e.g. Cornell, TMC) and the 'straight' probabilistic one (e.g. City, Dortmund); latent semantic indexing (Bellcore) is a type of vector model, probabilistic inference nets (e.g. Amherst, CUNY) are another form of probabilistic model, which also extends towards learning (see below) via connectionist implementations.

These models ramify into all aspects of indexing and retrieval. The TREC linguistically-motivated models are less aggressive and are primarily focused on the use of syntactic or semantic criteria for identifying index terms, especially compound ones, and variant representations of the same concept (e.g. CMU, NYU). Selection criteria based on large-scale discourse structure have also been examined, by Syracuse. In general in TREC, however, there has been little use of linguistically-complex complete index descriptions of the traditional subject-heading type, presumably both because the TREC requests are too extensive to encapsulate in one or two descriptive units, and because this is not appropriate for the documents. Some projects (e.g. Siemens, Conquest) have also used manual thesauri or other classifications implicitly based on linguistic considerations. In general, linguistic models for index terms have been combined with subsequent statistically-motivated processing, e.g. for weighting or scoring. Equally, primarily statistical indexing may be softened by encompassing compound or phrasal terms based on adjacency or proximity (e.g. Dortmund, TRW).

*Model questions*

The first questions to be asked about the TREC tests are thus what differences alternative models make, as follows:

M1: Are linguistically-motivated models superior to statistically-motivated ones?

M2: Are there performance differences between the models in either class, e.g. between vector and probabilistic in the statistical group, between sentence syntax (NYU) and discourse syntax (CMU, weakly, Syracuse) in the linguistic?

M3: From a different perspective, do more sophisticated or refined models contribute anything, where linguistically-motivated term selection criteria may be taken as more refined than statistical ones; or in the linguistic case, do richer linguistic models (NYU, Syracuse) outperform plain ones (CMU); or in the statistical case, are rigorous probabilistic models (City) superior to vector ones (Cornell), or fancy vectors (Bellcore, HNC) better than simple ones (Cornell)?

M4: As the main specific issue here (given that everyone does stemming, and starting from the baseline): are linguistically-grounded compound terms, as opposed to term conjunction in matching, valuable?

## 3.3 Indexing vocabulary

The model distinctions just considered clearly apply to the indexing vocabulary used. The issue for vocabulary is whether a vocabulary is established as a (more or less) fixed entity for explicitly characterising documents, rather than being simply taken as the universe of potential terms embodied in the documents and realised at search time. As mentioned already, operating with an explicit vocabulary is so exigent that it has not been common in the TREC work, though it is exemplified in the statistical case by latent semantic indexing (Bellcore), and in the use of collection-motivated phrase vocabularies by Cornell and Dortmund. (Manual thesauri have been primarily used to supplement, not to form, descriptions e.g. by Verity, Siemens, Conquest.) However vocabulary design is also approximated by term weighting schemes using collection rather than individual document information, since these can be taken as relativising term value for any use. In this form vocabulary manipulation is common to both linguistic and statistical approaches and for the latter to e.g. both vector and probabilistic cases, as well as being a motivation for the learning strategies considered later.

*Vocabulary questions*

V1: Does a holistic approach to the indexing vocabulary pay its rent?

V2: Is linguistic sophistication important?

However it should be noted that the TREC requests (and also documents) include proper name elements with many variations, abbreviations etc, and several teams have made use of special purpose lexicons, grammars or other processing strategies to handle these terms. Whether these are mandatory because matches on these items are demanded, or are just useful as a means of generally improving performance, depends on careful analysis of the individual application. The claim is that they are beneficial, though not hugely so, in TREC.

## 3.4 Document descriptions

The model distinctions clearly apply to the way individual documents are treated, i.e. what the basis is for choosing a term to index a given document. Thus setting aside the virtually universal use of stemming and a stoplist, the main distinction between the TREC teams has been whether document terms are linguistically or statistically motivated, the latter including proximity which is also de facto document indexing even if implemented through matching conditions at search time. The context vectors used by HNC are a complex way of exploiting proximity information in statistical document indexing. Linguistic processing is mostly local, sentence-based (e.g. Amherst, NYU) but, as mentioned earlier, may take account of large-scale document structure, though the Syracuse approach here has not been actually tested in retrieval yet. Using compound terms (whether linguistically or statistically defined) or single terms is itself a major distinction, and wherever compound terms are used, one of the issues is whether their component elements are also assigned as terms or allowed in matching. At the same time, many approaches exploit statistical weighting for whatever document terms are used, with the practical problem for derived compounds defined at search time of having the incidence data to do this.

*Description questions*

D1: Is linguistically-motivated indexing superior to statistically-motivated indexing, especially where this is derivative for individual documents?

D2: Are compound terms (whether linguistic or statistical) superior to single terms?

D3: For the full texts used in TREC, what is the value of document-specific weighting schemes?

## 3.5 Indexing sources

While indexing strategies may be indifferently applied to documents or requests, there may be document-specific constraints depending either on views about document form or on the practicalities of large file searching. As mentioned earlier, a choice that might be thought relevant to the full text case, namely whether full text or only title+abstract is actually used for searching is not often available in the TREC case; and for the same reason, differential weighting taking advantage of the concentration of key information is abstracts is not possible.

Some strategies involving *subdocuments*, e.g. paragraphs (Cornell) or pages (CITRI), have however been investigated within TREC which can be looked at from this point of view. Thus preferring (Cornell, CMU), or requiring (CITRI, CUNY), query matches within page units can be viewed as an opportunistic use of subdocuments as indexing sources for the whole. However retrieving i.e. offering the user, subdocuments only has hardly been investigated in TREC, since though the nature of the TREC requests might suggest it could be appropriate for long file documents, even if not worth it or suited to the short news stories, the TREC relevance data makes proper performance evaluation impossible (CITRI).

*Source questions*

S1: Is using subdocuments as indexing sources, as opposed to the whole text, useful?

## 3.6 Queries and query sources

While IR of course requires compatible document and request indexing, and some TREC participants adopt similar strategies for both, most of the TREC effort has gone into request rather than document indexing. The complex and non-standard character of the TREC requests means that they offer varied possibilities for query formation, and work has been done both on different ways of treating the various fields (e.g. Amherst, Syracuse) and on which fields are useful. The difficulties of processing the large document files has also encouraged concentration on the requests, where manual query formation may be viable or manual modification perfectly practicable (e.g. Rutgers, TRW, Verity, Conquest, PRC). Of course queries as search specifications may be very complex (e.g. Amherst, CMU, TRW, Verity, Conquest), or have a (semi) boolean form (e.g. GE. TRW, CUNY, VT), though actual document matches are simple. Compound index terms defined by proximity may be specified via queries (e.g. GE, TRW, VT), and may be treated slightly differently in queries and documents, for instance where compound terms are only accepted as such on tight criteria for requests, even if they are more loosely matched in searching: though formally the indexing is the same, the grounding is biased to the request. The same weighting schemes may be applied to documents and queries, but though some types may be less informative for the latter the long TREC requests offer scope for query term weighting, and this has been extensively used (e.g. Cornell, Bellcore, ETH, Verity, Conquest). Apart from the manual queries (which may be linguistically complex), query indexing has been primarily statistical, though some teams have applied linguistic analysis to request texts (Amherst, Syracuse, NYU, CMU).

Further possibilities specifically for queries are associated with mechanisms for query *expansion*, and especially query expansion in relevance feedback. Many teams used relevance feedback to apply training data about documents including relevance information to alter individual queries (as learning this is considered later). This is natural for routing requests, but there have also been a few tests with adhoc requests (City, HNC). Several teams used thesaurus expansion in conventional style (e.g. GE, TRW, Verity, Siemens, Conquest), without relevance information: this might or might not be automatic, and use a statistical or linguistic thesaurus. However TREC offered the opportunity for a deliberate and systematic use of relevance information for query expansion, not just reweighting, and as a device for improving either routing or adhoc queries. The expansion process may be fully automatic (given relevance assessments), semi-automatic (the user may be offered vetted terms to endorse) or manual. Many TREC teams used expansion for the routing mode (e.g. Cornell, Dortmund, Amherst, HNC, Bellcore, CMU), but experimental difficulties limited work on the adhoc case (City).

With respect to query sources in requests, the nature of the TREC requests made this quite a complex matter. The request fields differ in type of information, degree of formality, and linguistic properties (isolated words or phrases versus running text), though they overlap lexically. The range of fields available in TREC stimulated many tests trying various selections from and combinations of fields, with a specific point of interest in the utility of the Narrative field (Amherst, VT). Further, as both requests as wholes and some individual fields are quite long, this has implications for query length and query term weighting as system parameters.

*Query questions*

Thus with respect to queries, there are both questions that are relatively independent of the particular TREC application, and ones that are more narrowly tied to it.

Q1: What form of index description is effective, especially is complex superior to simple?

Q2: Is manual query formation better than automatic?

Q3: Is query expansion valuable?

Q4: Is feedback in general useful?

Q5: For TREC specifically, are explicit statements of document relevance criteria helpful as a direct source of search terms?

## 3.7 Search strategy

The search strategy in its fullest sense has not been a parameter of major interest in TREC, other than via indexing or scoring procedures applied in a routine and uniform manner across all queries at search time. More generally, the TREC evaluation design working with ranked output implies a 'normal' strategy of ordering by decreasing matching score. But there has been almost no study of classic areas of concern like e.g. starting narrow and broadening out, or vice versa, except as an automatic filtering device, or (except for Cornell) of flexible and varied procedures suited to the individual request: everything is standardised and averaged. While some queries were manually constructed, this was not by end users, and apparently with little study of or control over the factors influencing the construction process. Search strategies are also applied in the reformulation of queries in iterative searching (as in the City case), but tests in this area have been too limited for explicit strategy studies, and indeed the motivation for these tests was such that those developing searches were deliberately encouraged to do whatever they thought was sensible. Moreover much of the motivation in feedback tests has been to develop automated methods that reduce effort for the end user. Using subdocuments as a normal target in matching (CITRI, CUNY) can be seen as a type of search strategy; but it has been uniformly applied without variation for individual requests.

The very detailed and specific nature of the TREC requests has, in fact, reduced the pressure to devise productive ways of developing useful queries from tentative and minimal initial requests, or to tailor searches according to types of request, with one strategy for a broad request and another for a narrow one. The TREC enterprise has thus not so far addressed a major challenge for operational systems faced with very large volumes of material and, also, not very experienced users.

*Search strategy questions*

There is a generally relevant question about searching that can be asked of the TREC tests:

Y1: Do uniform and primarily automatic strategies give reasonable results for one-off requests, and also reasonably uniform results across different queries?

Though the TREC requests do not vary greatly in specificity, they still differ enough for it to be legitimate to ask these questions.

## 3.8 Scoring criteria

Scoring criteria, on the other hand, have been investigated in various ways. While matching scores may be strongly influenced by the form of indexing, e.g. by document-derived term weights, there are nevertheless different possibilities for characterising the relation between query and document. Thus the SMART model, used by several TREC teams, embodies a specific definition of what a document-query match means.

However it is also possible to refine on query-document scoring by such strategies as promoting documents that match locally rather than just globally (Cornell, CMU). Some of the TREC participants have, moreover, experimented with *data fusion* ideas, or ways of combining different versions or subcomponents of queries (e.g. Rutgers, Amherst, HNC, IDS, VT, PRC). These naturally follow from e.g. inference net models which can subsume different types of indexing information, but it is additionally possible to assign different scoring values to different information types as a function of query document matches, as investigated by e.g. VT.

*Scoring questions*

C1: Are complex, differential scoring functions superior to more straightforward ones of a plain sum or product kind?

## 3.9 Output form

Output form is an important parameter for operational systems, covering both the way the search output for a query is presented, e.g. in rank order, and the way the user is given information about individual documents. Some of the issues arising in older systems where there was no output ranking, for instance, or only title information in the file, do not apply in TREC. However, though full text systems present their own challenges for output presentation, these have not been investigated in TREC (other than for page output by CITRI), since the provision of output just as ranked document identifiers is geared wholly to the formal evaluation protocol and is divorced from users. The interface provided for gathering relevance assessments, though it might have properties bearing on output forms for users, was designed specifically for the evaluation task and so should not be considered under this parameter heading.

## 3.10 Learning

I have explicitly or implicitly referred to learning under several of the previous headings. Whether or not a system is designed to learn at all is in itself a very generic feature, rather than a system parameter: the different points to which learning is applied, whether of data or process, are those that define parameters. However learning may be more widespread, in system *adaptation* affecting all index term weights and relationships; or targetted on a single component, as in *revision* of the indexing vocabulary; or particularised by events as in individual query *alteration*, subsuming query *feedback* either as routing query *evolution* or as adhoc query *modification*, using *relevance feedback* for reweighting or expansion. These possibilities have all been studied before, but both the TREC data and the structure of the TREC programme offer a valuable context for research on system learning in response to changes in documents, requests, or relevance needs. First, TREC offers very large volumes of text data, providing a respectable corpus for e.g. learning word collocation patterns (e.g. GE); second, there are large relevance sets per request; and third, there is the routing task, for which query evolution is naturally appropriate. In addition, TREC has not depended on the use of a conventional controlled vocabulary or classification of the sort which is manually revised either via adhoc gap plugging or as an occasional, global exercise that relies on informed experience but is not rigorously tied to collection data. With TREC it is in principle possible to investigate continuous and systematic learning, and also to take advantage of the modern technology and learning models (e.g. connectionist ones) that are available. Thus the TREC effort has included work on automatic learning with a fairly serious intention, including overall system index adaptation (CUNY), adaptation centred on vocabulary revision (Bellcore), and extensive work with relevance feedback by many teams. But the adaptation work has not been pushed far as a continuous rather than spasmodic process, and the relevance feedback studies, even though they have been notionally for routing, have also been non-continuous.

Adaptation is of course collection *tuning*, whether within the context of a standard IR strategy or one already *tailored* to a collection type. One important form of tuning for which TREC, through the volume of data it provides, has offered a serious study opportunity is via the application of *regression analysis* to determine the relative weighting to be given to the various elements of indexing (Berkeley, Dortmund, ADS).

*Learning questions*

L1: Is overall adaptive tuning to a collection valuable?

L2: Is refined vocabulary revision helpful?

L3: With respect to requests, does relevance feedback significantly improve performance?

L4: Does regression analysis help tuning?

# 4 TREC results

In this section I shall indicate what, in my view, the TREC *results* are, i.e. what the performance data laid out in the evaluation record, and especially (though not exclusively) in the official submitted results, show. This section is thus primarily a descriptive summary that tries to pull together the very large amount of very varied performance data, especially, as mentioned earlier, for the full tests as opposed to Category B ones. I shall then as a separate exercise, in the following section, review these results, taking into account the participants' own accounts of their aims and activities, in an attempt to determine what the results mean and hence to establish what the TREC *findings* about IR systems are. This broader assessment has naturally to refer to the influence of the TREC case.

## 4.1 Review strategy

My summary of the TREC results has necessarily to deal in rather crude generalisations. The complex test context, wide spread of runs, and hence lack of systematic, fine-grained experimental control make this the only practicable review strategy. It is also important, given the many questions still open, not to attempt a vulgar 'pick the winners' approach.

In comparing performance figures, it is impossible to be anything other than informal: though systematic statistical significance tests are needed they have not yet been done, and it is also not clear how the magnitude of performance differences for this data are to be interpreted. For example, to refer to an earlier scheme, is it reasonable to label performance differences (assumed statistically significant) of 5 % and 10 % as 'noticeable' and 'material' respectively? These conventions also refered to $R/P$ graph areas. In this paper I shall, in my attempt to make a first cut across the result, further focus my comparisons on one *fixed point*, namely $P$ at document cutoff level 30, justifying this on the grounds that this is both a transparent and a realistic performance indicator, at least more so than the others used. I shall then, in comparing two performance results at this point, use the notions just mentioned, namely no difference, a noticeable, and a material difference (symbolised as =, >, or >>), with the last two corresponding (roughly) to 10 and 20 percent increases in $P$: this is appropriate for such simple analysis.

It is, further, useful to summarise the submitted full results for the fixed point *Pdoc30* as shown in Table 1: this assigns each team's best result for each of their adhoc and routing pairs to a performance *block* and and also refers only to the upper performance levels, namely for adhoc the blocks >= 55, 50, 45, and for routing >=

60, 55, 50. (For the comparative record, analogous results for recall cutoff at 30 % are also shown.) Results or approaches where there are no or few full runs are considered only where this is especially relevant: it is not really appropriate, in particular, to discuss performance results for TREC starters who are still getting their act together. It should be noted that, quite apart from any strategy variations that naturally follow from the difference between routing and adhoc retrieval, the results for adhoc and routing for a given team are not necessarily obtained in a similar way: while the general style is normally the same, teams may have investigated distinct approaches in the two cases.

The table of course somewhat crudely divorces adjoining figures, e.g. two teams with results of 60 and 59 respectively are separated, so the table blocks should be treated with caution and only as a way of achieving leverage and insight. It should also be noted that the recall cutoff values are somewhat lower than the document cutoff ones, indicating the danger of attributing too much importance to the absolute numbers obtained, though the relative positions of teams according to the two measures are similar. *It also cannot be emphasised too strongly that these figures provide only one cut across the large mass of experiments and that the detailed accounts, especially for the final TREC-2 Proceedings, make it plain both that sheer accidents can occur (e.g. Bellcore, Cornell), probably or actually artificially depressing performance, and also that fuller work can significantly upgrade performance, for instance in choosing suitable values for constants in formulae. Thus work at City following the official runs submission raised the adhoc performance shown in Table 1 by two blocks, and at CUNY brought routing performance up to the Table levels. These two examples show how the continuing search for performance improvement can be one of the major gains from TREC and similar enterprises, while reinforcing my point that the officially tabulated runs are only a starting point for assessing TREC.*

## 4.2 Global comments

Looking at Table 1, the following global comments can be made:

1. Performance is pretty good: 55 % on hundreds of thousands of documents is very creditable.
2. The routing results are generally better, but not enormously so.
3. The spread from top to bottom even of this subset of better results is quite large.
4. Many teams have very similar performance.
5. Teams generally have similar relative positions for routing and for adhoc.

## 4.3 Question answers

Now, considering the test results in relation to the TREC participants' approaches to indexing and retrieval, I shall first seek answers to the system parameter questions asked in the previous section, and then sum up the major conclusions that can be drawn from the results overall. The question answers attempt to isolate performance elements but, as noted earlier, this can only be done in a rough and ready way given the mass of miscellanous combinations represented and the difficulty of assigning performance responsibility; equally, while individual parameter choices may not seem to contribute much, they may be effective in (some) combinations.

*Model answers*

M1: Heavily linguistic approaches are only represented by CMU, since NYU is in category B. CMU's performance shows a linguistic approach performs perfectly well, but no better than statistical ones.

M2: Different linguistic models were not assessed. The main statistical schools, vector and probabilistic, perform similarly (e.g. Cornell vs Amherst).

M3: (Linguistic not assessed.) Elaborate statistical models e.g. Bellcore, do not better than simple ones (disregarding training effects).

M4: Thus, following the previous answers, there is no gain from using linguistically-motivated, as opposed to adjacency-defined, compound terms.

*Vocabulary answers*

V1: Only one of the TREC teams applied any strongly holistic vocabulary design methods, namely Bellcore, but without obvious benefit. However some (e.g. Cornell and Dortmund) employed a pruned and normalised extracted phrase list (along with ordinary single terms), though this too, alone, does not appear to be significantly superior to simple single-term extraction and individual-item weighting.

V2: Manual thesauri were used primarily as adjuncts, but without noticeable effect e.g. Siemens.

*Document description questions*

D1: There is no clear gain from linguistic processing to select document terms, e.g. compounds, as in CMU.

D2: There may nevertheless be a slight advantage from compounds (e.g. Amherst, Dortmund) as opposed to single terms.

D3: There does appear, moreover, to be advantage in document-based weighting, practised by many participants including e.g. Cornell, Dortmund.

*Indexing source questions*

S1: Work using subdocuments to constrain matching as a basis for retrieval, as in Cornell, CITRI, is inconclusive.

*Query and query source questions*

As mentioned, the scale and structure of the TREC requests offered considerable scope for investigations of query formation, and many participants explored the possibilities. Thus with respect to query formation both manual and automatic approaches were tried (e.g. Verity, VT vs Cornell); also complex automatic strategies using statistics (e.g. Dortmund) or rules (e.g. CUNY) or linguistic processing (e.g. Amherst, CMU); query weighting (e.g. Dortmund, Bellcore); and various forms of query development through expansion and relevance feedback for reweighting or expansion: all of these last were particularly popular.

Q1: There does not seem to be any striking advantage in applying sophisticated query construction methods (as such, independent of collection properties), whether relying on human intelligence, or elaborate automatic strategies.

Q2: There is no clear gain from manual query preparation.

Q3: There does not appear to be any advantage in query expansion independent of relevance information (e.g. Siemens, Conquest).

Q4: There does, however, appear to be utility in relevance feedback, especially via query expansion (primarily tested for the routing case).

With respect to query source (specifically for TREC), the Concepts field (cf Figure 1) is normally recognised as crucial, and Title and Description may be thrown in.

Q5: Whether or not the Narrative field, giving relevance constraints, should be exploited, as a source of terms, term conditions, or weighting, is not clear.

The general implications of these answers, given the unusual character of the TREC requests to which they refer, are considered later.

*Search strategy answers*

Both individual teams, and different teams, obtained quite different performance per request (cf Harman in (Harman 1994a)).

Y1: However uniform strategies applied across request sets gave aggregated results that were 'good enough'.

*Scoring criteria answers*

A few teams explored combination scoring or differential criteria applied at search time, under the headings of multiple queries or data fusion (e.g. Rutgers, HNC, VT). Differential scoring of a more pervasive sort may also be established by learning e.g. through regression analysis as investigated by e.g. Dortmund, and by choices of values for constants in formulae (e.g. City).

C1: There is not enough evidence to show that query-based differential scoring is valuable, though collection-based learnt differences may be useful.

*Learning answers*

The TREC experiments addressed learning primarily in two forms: generalised collection adaptation, or tuning, notably by regression, and individual query alteration in relevance feedback: though vocabularies were constructed for TREC data sets e.g. by Bellcore, they were not continuously revised to match collection development, and though connectionist techniques were used by e.g. CUNY, HNC, they were not applied in continuous learning mode.

L1: There is no strong evidence for adaptive tuning beyond the obvious weight changing that accompanies collection growth and decay.

L2: There is no proper evidence for the value of vocabulary revision.

L3: Use of relevance information for changing routing queries appears beneficial, though the TREC design limited the type of testing available. The TREC design restricted feedback testing too much for proper tests with feedback on adhoc queries.

L4: There is no clear evidence for regression analysis as a specific tuning technique, since systems without it worked as well as those with.

## 4.4 Summary of results

Summarising the results overall, there are no parameter choices that hold across all best performance: there are only tendencies for some system features to be associated with better performance, even if not required for it. However combination effects may be equally important, given how much variation there also is between performance for approaches of similar types.

Subject to these cautions, we can now pull together the answers to the particular questions to make some broad brush generalisations across the results as a whole. These are

1. That very different approaches perform equally well.
2. That relatively simple statistical approaches along the lines successfully followed in past IR research perform as well as more complex ones, e.g. exploiting linguistic processing, in the 'long-document, large-document-set' case.
3. That insofar as compound terms are useful, which is not wholly certain, it may be sufficient to define these only by adjacency or proximity.
4. That the systematic exploitation of weighting, applied to different sources of information within the overall statistical framework, is valuable.
5. That tailoring to generic collection properties, e.g. long documents, may be useful, but this is not very clear.
6. That tuning to individual collection characteristics, e.g. through regression analysis, is desirable since though weighting in particular is useful, it is subject to considerable variation in value (cf Buckley (1993)). However such specialised aids as e.g. name lists may not helpful enough to demand explicit provision.
7. That concentrating effort on the request is much more effective and efficient than working (a priori) on document descriptions.
8. That query modification through expansion and reweighting is valuable.

As an alternative view, we can characterise the best performance compared with the baseline in terms of our earlier noticeable/material terms for improvement. For the baseline we can make use of TMC. This is not improper as TMC explicitly addressed technology issues within the framework of indexing and retrieval by the established statistical techniques (though using adjacency phrases as well as single terms), and the fact that TMC's performance was not better is *no* criticism: TMC could expect to benefit, within its technological framework, from the ways of improving performance established by TREC in general. TMC's baseline $P$ is in the group >= 35 for adhoc and >= 40 for routing (for document cutoff). Thus we see in Table 1 that the best adhoc group, at >= 55 has achieved a performance improvement of 57 percent, and the best routing group, at >= 60, has increased performance by 50 percent. These >>> (or, perhaps, >>>>) improvements can legitimately be called *striking*. They also represent, from the user's point of view, a real performance difference: getting 16.5 relevant documents as opposed to 10.5 (adhoc) or 18 instead of 12 (routing) would be welcome in many circumstances. On the other hand, it has also to be borne in mind that many teams achieved ad hoc 50 or routing 55 percent $P$, so the difference between their performance and the best, while a noticeable (10 percent) difference by my earlier definition, would only bring an extra 1-2 relevant documents given some 15 already.

The top group teams, and those near them in the group below, are primarily from the statistical camp, and while they apply different specific combinations of IR device, have much in common, tending to share a use of compound terms, of rich weighting schemes, of careful tuning for numerical values, and in Amherst's case in particular, a deliberate attack on the characteristic features of the TREC requests.

# 5 Assessment

From one point of view the fact that, here again, relatively straightforward statistical approaches work perfectly reasonably confirms previous results in the field. This is important because the main criticism of previous research has been its comparatively small scale: here in the 'long & large' case the results still hold. It is however still necessary to consider whether there may have been favourable biasing conditions in the TREC case and also, more generally, how the TREC case affects interpretation and application of the test results. This needs to be done in relation both to the TREC data properties, and to the evaluation methodology.

## 5.1 TREC data effects

The TREC data properties of importance here are the heterogeneous document sets, elaborate queries, and form of relevance need, plus the fact that all the data - document sets, queries and assessments - were 'non-natural'.

The high quality of the requests may have been of considerable significance both in allowing a high level of performance and in permitting it with 'routine' statistically-based techniques. For instance, in earlier IR tests large-scale query expansion was found damaging, so while it is not completely clear whether the gains from it in TREC were due to the queries providing better anchoring or to superior expansion criteria, it is likely that the former was important. The high request quality, and especially the carefully developed Concepts fields, may equally mean that sophisticated query derivation methods were not required. The next phase of the TREC programme, in TREC-3, is intended to test alternative, much shorter and 'more ordinary' requests.

Again, it is possible that the large relevance sets, perhaps associated with the character of the TREC relevance needs, may have made it easier to retrieve relevant documents than is usual, and future tests with other relevance criteria and sets are needed.

Thus for both reasons, while statistically-based approaches may be as good as others, TREC may have reached a much higher performance level than could be expected with poorer starting requests (cf also Blair and Maron (1985)), with knockon effects for relevance supply and query modification.

The non-natural document sets may be unfavourable to TREC, since in ordinary circumstances data sets may be explicitly distinguished and offered for user preview, on global characterisations. The comparative lack of variety in the TREC requests, on the other hand, may have been favourable to TREC in allowing more tuning than would ordinarily be feasible. There is no doubt that the 'ordinary user' has not hitherto been present in TREC, and needs to be in future tests.

## 5.2 TREC methodology effects

The main issues here are the dual use of requests for routing and adhoc retrieval, the form of the routing tests, and the lack of user participation in reacting to routed documents or in conducting ad hoc searches. As mentioned earlier, the lack of 'reality' with respect to users is more serious now than in earlier tests, when batch searching was also normal in operational systems.

The most important consequence of the first point was mentioned in the last section, since it is very unlikely that even if routing requests are honed, adhoc ones will be. Thus TREC adhoc performance, though lower than routing, may still be far too high. At the same time, routing performance may also be too high because of the large amounts of systematically obtained training data. While in principle streamed routing would allow the gathering of many assessments with little effort, it is not clear whether the assessments would be consistent over time, or in fact very voluminous.

The inferences to be drawn about real routing system design from the TREC tests are also not clear, since genuine streaming and 'online' allocation of documents to users may create different retrieval conditions calling for e.g. different search strategies or query modification. For instance, it is not obvious what the effects would be for sought recall or precision. This point refers both to the more formal aspects of test design and to the lack of direct user involvement in the TREC case: future TRECs need to address at least the former, and a modest beginning will indeed be made here in TREC-3, through the application of thresholds in evaluation.

However the most important weakenss of TREC so far has been the lack of end-user involvement in searching, since this means that there has been no investigation of the way users respond to the 'long & large' conditions and in particular, if they do not cope well, what strategies for helping them would be most effective. This applies, for instance, to whether adequate starting requests can be formulated, or whether effective bootstrapping in relevance feedback can be guaranteed. The planned TREC-3 interactive searches are thus extremely important as a check on the results obtained so far, though these tests too will only be a first attack on this issue since they will not involve 'real' users and will be subject to the artificialities and restrictions that follow from the design need for practically viable and reasonably controlled evaluation.

## 5.3 Technology issues

In a general way, TREC put a heavy stress on virtually all the participating teams' technology, and it is a striking comment on computational progress, and its implications for automatic indexing and retrieval, that even quite small teams were able to carry out these demanding tests on such large volumes of material. Computational strategies were generally organised round the need to conduct many experiments, so not very strong conclusions should be drawn about the operational implications of any of the indexing and retrieval strategies studied. However it is worth noticing on the one hand that some approaches required very substantial processing effort, e.g. Bellcore's computation of latent semantic vectors; and on the other that a few teams deliberately studied the computational efficiency implications of their indexing and searching strategies, with operational performance in mind. These included TMC, working with a mainstream statistical approach, TRW and CITRI, while ETH and ERIM explored reductive coding schemes.

Looking at the TREC tests from this efficiency point of view, it is fair to say that they show that some quite costly approaches are not wholly out of court, though significant scaling up, or frequent revision, of an indexing vocabulary does present major problems. However it is also possible to draw the conclusion that, should computational pressures prohibit significant pre-processing of vocabulary or documents, there are viable IR strategies that do not require these.

## 5.4 Overall research findings

If we now stand back and ask what *general findings* for IR as a whole emerge from the TREC results, subject to the matters just discussed, these seem to be as follows. Note that given the varied masses of test detail, these findings are necessarily broad statements, and ones reflecting majority generalisations rather than universal truths. They are also intended to apply to both adhoc and routing modes, though possibly more strongly to one rather than the other either for intrinsic reasons or because there is more evidence in one case than the other.

On this basis my findings are:

1. That a reasonable level of performance for large sets of full text documents can be achieved by statistically-motivated techniques.
2. That within this type of approach great improvements over the simple 'baseline' can be achieved.
3. That these improvements are apparently promoted by combinations of data and process that are not necessarily individually very helpful.
4. That these approaches require tailoring to generic collection properties, and typically also tuning to specific collection characteristics.
5. That computationally, system parametrisation is practicable while actual retrieval can be very efficient.
6. That the quality of requests (and hence queries) appears very important.
7. That systematically-exploited feedback relying on relevance information is valuable.

As an alternative way of obtaining a findings summary we can make use of the idea of *benchmark* rather than baseline performance. This also has the advantage of allowing for the possibility that the TMC results we have forcibly adopted as a baseline are unusually low for some hidden reason. Thus if we take the kind of slightly richer statistical approach, using several types of weighting information and extending to simple phrases, that is exemplified by SMART and has established performance credentials, as a generic benchmark, this allows a sharper formulation of the TREC findings. This benchmark appears in the middle block for adhoc in Table 1, and it is evident that it is hard to do conspicuously better in adhoc searching. With routing on the other hand it is, not surprisingly, possible to make larger gains using relevance information over the notional baseline or benchmark represented by the adhoc performance for similar collection data.

## 5.5 Overall methodology lessons

The TREC tests to date are essentially just a rerun, on a much larger scale, of the standard evaluation procedures used for many years. As noted earlier, these involve a high degree of abstraction over system behaviour. However both the standard *evaluation results* given for the submitted search outputs in the Proceedings, and Harman's own analyses in these, show how easy it is nowadays to manipulate the raw output data from a variety of points of view. Thus although the official evaluation results have deficiencies, for example those associated with recall cutoff, and significance testing is needed, these data are still very informative. Moreover the sheer scale of the tests makes it difficult to carry out the detailed failure analyses that are ideally required, just as it also makes it difficult to conduct meaningful investigations of the effects of end-user needs or behaviour.

However while it may be very hard to conduct adequate diagnostic tests on this scale, in order to achieved properly controlled comparisons, the volume of test and evaluation results does to some extent justify the rather coarse view implicit in the whole, namely that its what you get, not how you get it, that counts.

Thus perhaps the single most important observation about TREC to date, relevant both to research testing and operational design, is that single 'errors' can have quite striking effects (cf the report by Bellcore), and that it is necessary to do exhaustive testing (and learning) to establish reliable and consistent performance.

# 6 The future

Here the focus is (a) on TREC as an *evaluation exemplar* and (b) on TREC as a *prototyping guide*. The first refers both to the future of TREC itself, and to its role in IR experiment in general; the second to how both the TREC findings as such and the lessons to be drawn from the TREC enterprise as a whole may be applied to operational systems.

## 6.1 Prospective and possible TRECs

As already noted, TREC-3 will begin to tackle other request types and interactive searching, as well as more realistic routing. All of these will need further development. It will also be important in the longer run to engage with a wider range of document genres, and with more sensible database partitions, as well as with, on the other hand, more heterogenous aggregative resources like those flooding the Internet. Looking ahead, the related TIPSTER programme is already working with more than one language, and for future TRECs both tests of methods on different languages and genuinely multilingual retrieval are needed. There is also a need to do further scaling up tests.

For IR experiment more generally, TREC has roles both in providing some performance benchmarking and an evaluation reference paradigm and further, amplifying these, can and should play an important part in fostering research when the actual TREC Collection, test results, and evaluation results are made available, in machine-readable form, to the community: here at last are worthy, if not ideal, successors to the exhausted test collections of former times. In this context, the Questionnaire responses provided by the teams gives valuable supporting detail of a kind too often lacking in the past: though it is not always easy to characterise individual approaches using some standard set of questions, the attempt both to gather explicit statements about detailed choices and processes, and to do this in a way allowing comparisons on specific points, should be a stimulus to better reporting practice and test methodology in the field.

## 6.2 Applying the TREC findings to systems

The main outcome should be a more widespread and wholehearted implementation of the mainstream statistical approach to operational systems. This must however be subject to tailoring and tuning, and to proper support, especially for adhoc requests, for the user in initiating and responding to the search process.

## 7 Conclusion

The major claim made for the ARPA programmes in general is that they hustle research communities very effectively, leading to significant and rapid improvement in task performance. This comes from simple time pressure and from the strongly competitive zeitgeist but also, especially for a programme with as many teams as TREC, from the interactive and communcative synergy that working together in a common framework promotes. One of the major issues is generalisation, and here some ARPA tasks have broader implications than others. TREC is perhaps more generalisable than some other tasks, in part because it is continuing a long research tradition, which other tasks have lacked, and so can confirm past results as opposed to just producing initial ones.

The IR community should indeed be enormously grateful to TREC: it has revitalised IR research and also demonstrated the importance of text retrieval to wider communities, e.g. those engaged with natural language processing. The attention that has been paid to the evaluation in all aspects of its design and execution has set high standards for future work. More specifically, the IR community owes Donna Harman a huge debt for driving the whole.

## Acknowledgement

I have benefitted, as a member of the TREC-3 Programme Committee and as an Advisor to the City University team, from my involvement in TREC, and from the many discussions with Donna Harman and others to which this has led. I am also grateful to City University for the travel support that has allowed me to attend TREC meetings. Finally, I should thank my referee for valuable, constructive comments on the draft of this paper.

## References

Bates, M. (Ed.) (1993). *Human language technology* (Proceedings of the ARPA Workshop, 1993). San Mateo CA: Morgan Kaufmann.

Blair, D.C. and Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28, 285-299.

Buckley, C. (1993). The importance of proper weighting methods. In *Human language technology*, Ed. M. Bates. San Mateo CA: Morgan Kaufmann.

Gallant, S.I. (1991). A practical approach for representing context and form performing word sense disambiguation using neural networks. *Neural Computation*, 3, 293-309.

Galliers, J.R. and Sparck Jones, K. (1993). *Evaluating natural language processing systems.* Technical Report 291, Computer Laboratory, University of Cambridge. (Gzipped copy is available from the anonymous FTP server ftp.cl.cam.ac.uk, as the file TR291-ksj-jrg-evaluating-nl-systems.ps.gz in the directory reports; the file is compressed so must be transfered in binary.)

Harman, D.K. (Ed.) (1993). *The First Text REtrieval Conference (TREC-1).* NIST Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg MD.

Harman, D.K. (Ed.) (1994a). *The Second Tect REtrieval Conference (TREC-2).* NIST Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg MD.

Harman, D.K. (1994b). Review of TREC, *Information Processing and Management*, this issue.

Lewis, D.D. and Sparck Jones, K. (1993). *Natural language processing for information retrieval.* Technical Report 307, Computer Laboratory, University of Cambridge. (To appear in *Communications of the ACM.*)

# Appendix

KEY TO ORGANISATIONS

```
ADS       : Advanced Decision Systems
Amherst   : University of Massachusetts at Amherst
Bellcore  : Bellcore
Berkeley  : University of California at Berkeley
CITRI     : Collaborative IT Research Institute, University
  of Melbourne, Royal Melbourne Institute of Technology
City      : City University, London
CMU       : Carnegie Mellon University
Conquest  : ConQuest Inc
Cornell   : Cornell University
CUNY      : Queens College, City University of New York
Dalhousie : Dalhousie University
Dortmund  : University of Dortmund
ERIM      : Environment Research Institute of Michigan
ETH       : Swiss Federal Institute of Technology
Florida   : University of Central Florida
GE        : General Electric R & D Centre
HNC       : HNC Inc
IDS       : Institute for Decision Systems Research
Illinois  : University of Illinois at Chicago
Mead      : Mead Data Central
NYU       : Courant Institute, New York University
PRC       : PRC Inc
Rutgers   : Rutgers University
SEC       : Systems Environment Corporation
Siemens   : Siemens Corporate Research Inc
Syracuse  : University of Syracuse
TMC       : Thinking Machines Corporation
TRW       : TRW Systems Development Division, Paracel Inc
UCLA      : University of California at Los Angeles
Verity    : Verity Inc
VT        : Virginia Tech
```

Number: 051

Domain: International Economics

Title: Airbus Subsidies

Description:
Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.

Narrative:
A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

(Other possible fields include Nationality: e.g. U.S.; Definition: e.g. of Insider Trading, SALT-II; Time: e.g. future. None is commonly used.)

FIGURE 1 :   EXAMPLE TREC REQUEST (TOPIC)

| ADHOC | | ROUTING | |
|---|---|---|---|
| Precision at cutoff | | Precision at cutoff | |
| Recall 30 | DOCUMENTS 30 | Recall 30 | DOCUMENTS 30 |
| >= 45 Dortmund | >= 55 Amherst | >= 55 Cornell | >= 60 Cornell |
| Amherst | HNC | | Dortmund |
| CMU | VT | >= 50 Dortmund | |
| Siemens | | | |
| VT | | >= 45 City | >= 55 City |
| | >= 50 Cornell | Berkeley | Berkeley |
| >= 40 Cornell | Berkeley | Amherst | Amherst |
| Berkeley | Dortmund | Bellcore | Bellcore |
| HNC | CMU | CMU | CMU |
| Bellcore | Verity | GE | |
| CITRI | Siemens | TRW | >= 50 Rutgers |
| CUNY | CUNY | | HNC |
| | | | GE |
| >= 35 City | >= 45 City | | TRW |
| ETH | Bellcore | | Verity |
| Verity | ETH | | Siemens |
| Conquest | CITRI | | |
| | Conquest | | |

Notes: figures not rounded; teams per block in Proceedings page order
best of pairs if two runs submitted, but two often similar

TABLE 1 : SUMMARY OF BETTER OFFICIAL RESULTS FOR FULL RUNS