What's the value of TREC - is there a gap to jump or a chasm to bridge?[*]

Karen Sparck Jones
Computer Laboratory
University of Cambridge

Abstract

Work within the TREC Programme has concentrated on generalising, not particularising. Now is the time to think about particularising, that is, to address not further generalisation across information-seeking contexts but context-driven particularisation. This note develops this argument from an analysis of TREC work, applying notions taken from discussions of evaluation for language and information processing in general.

1 Introduction

The TREC Programme has been very successful at *generalising.* It has shown that essentially simple methods of retrieving documents (standard statistically-based VSM, BM25, InQuery, etc) can give decent 'basic benchmark' performance across a range of datasets, including quite large ones. But retrieval contexts vary widely, much more widely than the contexts the TREC datasets embody. The TREC Programme has sought to address variation, but it has done this in a largely ad hoc and unsystematic way. Thus even while allowing for the Programme's concentration on *core* retrieval system functionality, namely the ability to retrieve relevant documents, and excluding for now the tasks it has addressed that do not fall under the *document* retrieval heading, the generalisation the Programme seeks, or is assumed to be seeking, is incomplete. However, rather than simply continuing with the generalisation mission, intended to gain more support for the current findings, it may be time now to address *particularisation.*

My case for this, which follows, is based on the notion of *micro* variation, and on the distinction between system *environment* and task *context.* I will use the evaluation framework ideas developed in Sparck Jones and Galliers (1996) - hereafter SJG96 - and in Sparck Jones (2001) - hereafter SJ01 - to analyse the TREC experimental programme and to support my argument for a new direction for TREC. This argument is that the Programme needs to go beyong micro, or even *macro*, environment variation in order to move from the current type of Programme outcome to a future, potentially more valuable one, that is, as a way of moving from 'If you do this in whatever your retrieval world is, you'll be alright' to 'If this is what your retrieval world is like, do this.' To make this move the Programme needs to extend the narrow notion of environment that is currently deemed adequate, indeed appropriate, for experimental purposes to draw on the broader context within which retrieval is done.

(I will take Voorhees and Harman 2005 as an umbrella reference for the TREC Programme.)

[*] I am grateful to Ross Wilkinson for stimulating this note. It is written in a personal capacity and not as a member of the TREC Programme Committee.

2 Environments and contexts

A convenient way of summarising the Cranfield evaluation paradigm is in terms of *environment variables* and *system parameters*. In the controlled, laboratory experiment case for core retrieval, the environment consists of documents (D), requests (Q), and relevance assessments (R), where requests and assessments are taken to embody information needs. This embodiment indeed is taken to cover not only information content but other need attributes like literature level and so forth. These attributes may not be very visible: getting at need in all its aspects from just the overt D * Q * R data has been the motivation for machine learning in retrieval research; and though there has been no evidence so far that more recondite need attributes can be so learnt, many believe that they can.

The simple historic model for environment variation within this laboratory paradigm was of *micro variation*, i.e. of change to the request set - say plain or fancy, or to the relevance criteria and hence set - say high only or high and partial, for the same set of documents; less commonly there has been change to the document set while holding request or relevance criteria/practice constant. The richer classical model covered changes to both document set and request+relevance set, usually driven by the fact that radical changes of document set naturally provoked new requests and hence relevance sets. However all of this environment variation has been accompanied by the presumption that D * Q * R encapsulates all that it is necessary to know about environments for retrieval testing.

Specifically, D * Q * R tends to be taken as encapsulating the entire information-seeking *task* (T), rather than just the *experimental* task (X). For example, if the information-seeking task is to gather material on which to base a short report on a specific topic for someone who is not knowledgeable about it, this task context is taken as adequately represented by (D *) Q * R. Even with formal acknowledgement of the fact that the system's role is 'only' core retrieval, there has in practice been a good deal of conflation between the system *function* and the *purpose* that the *setup* - system plus context as a whole - serves.

I am referring here only to the Cranfield paradigm as adopted and developed for automatic retrieval e.g. by SMART, and as writ large in TREC. I am thus excluding the mass of research reported elsewhere, especially in the LIS literature, that has paid more attention to the larger context within which retrieval is done. But this has generally been investigative rather than systematically experimental, especially on TREC-style scales. I am also excluding TREC's manual searches and interactive studies, since these have in fact been captives within the paradigm rather than representatives of foreign cultures. Within TREC, following a long tradition in retrieval system research, the practice has been to assume that the D * Q * R environment subsumes all that it is really necessary to know about the system's surroundings, even if lip service has been paid to a larger notion of context. Most of the TREC Programme's energy has gone into experimenting with different internal system parameter settings within such narrowly defined environments.

Changing all of D, Q and R might seem to imply more than micro variation, or at least could be deemed to do so if the *type* of request and/or *style* of relevance assessment changed, not merely the actual document sets. Such variation, embodying a new form of need as well as new documents to draw on in trying to meet it, might be deemed to constitute *macro variation* rather than micro variation, and therefore as implicitly enlarging the TREC Programme's reference to contexts. Thus if the Programme has included macro as well as micro environment variation, this would seem to imply that it may have become more valuable in terms of messages for the larger world of practical retrieval applications.

The analysis below shows that TREC has been mainly engaged in micro-evaluation. However as also illustrated below, even within the document retrieval area, and setting aside the QA track which clearly represents a large shift, there has been macro variation within TREC. Filtering, and the Web track and its successor Terabyte, constitute macro variations. Genomics represents a perhaps more conspicuous macro variation, with the new Legal track potentially a further one with, in particular, a twist on relevance criteria.

But while there has been a gently increasing amount of macro variation within the TREC Programme conditions, the interesting questions are

a) how the participants have responded to this; and

b) to what extent this enrichment of the basic environment model can be taken seriously as importing real attention to context into the TREC Programme.

3 TREC strategies

In general, TREC participants have sought to adapt, or extend, their existing system apparatus to the new environment variable values. This is not merely a natural practical reaction, but seems a proper intellectual one: a successful outcome represents further generalisation of their basic method (VSM, BM25, InQuery, etc). It is not, however, naturally or even typically the outcome of reflection on what the new retrieval situation does, or may, require. Of course, in practice things are not so crude. Thus on the one hand, trying out an existing system to see how it behaves is one way of exploring the new situation. Post-retrieval failure analysis is another. Eyeballing the collection data, or exploring its statistical properties, are others.

But quite apart from whether this continued emphasis on generalisation is a principled intellectual response to TREC's growing macro rather than micro variation, my claim is that these modest moves from micro to macro make only very limited reference to context. Thus even if we assume that the context aspects of need can be sufficiently captured in initial requests (as TREC topics have done for just one, content-oriented aspect of need in their Narratives), there is typically nothing in TREC requests that reflects such context factors as those illustrated in the earlier example, namely that the returned documents are to support report writing for a non-expert. Indeed, this context characterisation should properly be enlarged to include other factors like whether the document reader/report writer is themselves fairly well informed.

This TREC failure is hardly surprising. In TREC, as in many other retrieval experiment situations, there is normally no material access to the encompassing wider context and especially to the actual users, whether because such real working contexts are too remote or because, fundamentally, they do not exist as prior, autonomous realities. But while there may be such legitimate reasons for an inability to engage with the pertinent retrieval setup as a whole (along with the core and control assumptions underlying the whole enterprise), the natural tendency for TREC participants has been to treat the key challenge as one of adaptation to new circumstances (whether by full-scale machine learning, or systematic tuning, or plain tweaking.) TREC participants' attitudes have thus reinforced the empirical restriction to environments, narrowly conceived, that the difficulties of getting data, desire for test control and so forth have imposed, with which the TREC Programme has had to live.

Figure 1 - TREC TRACKS

```
List excludes tracks which ran only twice in the past, includes recent startups.
Tracks grouped by experimental task similarity, within an overall time ordering

Ad Hoc         D mixed collections, as could be got, not from real setups but assumed ‘realistic’
               Q analyst style originally, assumed but never invoked analyst-pertinent setups;
                 later loosely emulating ‘ordinary user’ offerings of short requests
               R dominated by test collection creation interests, loosely assuming some user
                 seriousness (analyst/ordinary)

Routing        All-synthetic variant of ad hoc. Replaced by Filtering

Filtering      D moved towards real material
               Q analyst style assumed realistic
               R per document like Ad Hoc
```

```
MLing         Ad Hoc and like

XLing         Ad Hoc and like, but Q a gesture towards supposed real setups wrt users'
              language competence

Speech        D realistic as news stream
              Q analyst style 'realistic'
              R gesturing towards analyst

[Query        Studying request properties, indirect relation to real]
[Robust       Studying poorly-performing requests, also indirect]

QA retrieval subtask, excluding QA per se
              D realistic as genuine news material mass
              Q assumed analyst style, contributing role to encompassing larger task
              R assumed analyst-style, encompassing task motivated

VLC/Web       D semi-realistic, as genuine pages but selected set
              Q Ad Hoc and Like;
                short requests derived from real;
                realistic service-seeking and home/named-page requests
              R limited set assessment semi-realistic

Terabyte      Like Web

Novelty       D Ad Hoc
              Q Ad hoc
              R realistic judgements for document and sentence novelty in successive documents

Genomics      D real
              Q real requests for topic searches;
                real categories
              R almost-real assessment;
                real categorisation

Enterprise (begun 2005)
              D real
              Q realistic known-item, email discussion, find-an-expert
              R realistic

Spam (begun 2005)
              D real
              Q two real categories
              R real

Interactive   Ad Hoc material, realistic element through human participation in search process

HARD          Similar, though less, human involvement in search
```

Figure 1 gives a more detailed (though still fairly summary) picture of the TREC tracks, including Ad Hoc as a notional track. It shows the track D, Q, and R characterisations and hence test environments, and illustrates how limited these have been both individually and collectively, and how little a larger take on context has figured in the Programme as a whole.

"Realistic" in the figure implies that data and/or experimental task have some fairly direct connection with a real context but are not "real" i.e. simply drawn from such a context; document sets are real only if they are drawn as substantive subsets from a real context; requests are real if they originated with real users; but assessments are not real, as opposed to realistic, unless done by real request originators.

As Figure 1 shows, in general in TREC there has been little rethinking of the experimental task X in a larger sense as a retrieval setup task T, seeking to dig into or extend the environment significantly.

In some cases the experimental task, especially for newer or more lateral tracks, has involved deeper analysis of putative background context. Filtering can be taken as an early semi-example (in material contrast with Routing), since while it might have seemed initially to require only a minor adaptation of existing system behaviour, thinking about shifting from ranked output to yes/no decision making invited a rethink about X as it might be motivated by T. The Web track service-finding experimental task recognised a distinctive type of setup task. Question answering also tacitly invokes its wider contextual motivation even in its retrieval component, as well as in its question-answering component proper, if only a rather modest and limited way compared with analysts' full information-seeking behaviours. Nevertheless, from the beginning for more than two-thirds of the Programme, development was primarily by micro-variation, including reuse of static document or request sets.

More recently, the experimental tasks have moved closer to reality in both data sets and, more importantly, requests and relevance assessments. Early TRECs, and also later ones using early material for study purposes, used document sets that indeed consisted of genuine material but were not natural collections as exploited by some service or engine. The Filtering Reuters data was an exception here, though the repeated use of news material could also be justified by its generic significance for many real users in real contexts (like analysts). More importantly, though the early requests and assessments were specially created, the shift to short requests was welcomed as implying closer engagement with the world that real 'ordinary' users occupy, and the VLC/Web track moved further towards real contexts through the various request types (very short, named-/home-page etc) that it explored, though less with its document files. The Genomics track can, however, be taken as embodying a more marked acknowledgement of real context, both in choosing its document data and in defining its particular test tasks; this is also apparent in the recent Enterprise track, while the arguments that raged about the experimental tasks and performance measures for the Spam track clearly reflected issues and requirements in what setups in spam filtering require. The new Legal track is also being obliged, in order to specify its test tasks, to try to understand something about what real lawyers do and need.

It is nevertheless the case that such references to the encompassing context are made primarily for the purposes of defining a viable experimental task, and specifically one that will be sufficiently controlled and measurable. The outcome defines the system environment in the narrower sense and guides the experimental design. The context as a whole is encapsulated in the environment variable values and participants are either unable, or uninterested, in seeking to go back outside into the whole setup again. (Or at least they are only interested in this when they cannot understand the relevance assessments that are handed down.) Thus we can describe TREC as shifting from micro to macro variation, by addressing a larger range of, or more markedly different, environment values, but still as distancing itself from most of the detail, and especially interconnected detail, that seems to characterise real information-seeking contexts and their tasks. Context is not embraced, but reluctantly and minimally acknowledged, like an awkard and difficult child.

This applies even where explicit attempts have been made to include users (real or surrogate) in the experiments themselves, whether in the earlier Interactive track or the recent HARD one. The users in these cases are little black boxes separated - as far as TREC system builders are concerned - from larger operational contexts. TREC participants have also been able to exploit other context-pertinent data resources, both specialised, like MeSH and non-specialised, like WordNet. But these are extremely weak setup attributes.

4 Facing up to context

As noted, the classical experimental paradigm is systematic variation of either environment variable values or system parameter settings. An analysis of TREC up to 2005 on this basis shows a far from systematic range of different environment variable values, and though the collective work of TREC participants has shown a huge range of system variations for any given environment as well as for the whole set of environments, this does not compensate for the patchy state of environment variation.

One possible line forward would be to increase the range of environments, while maintaining the narrow view of environmental adequacy as requiring only micro variation, or only such macro variation as can be absorbed, by a kind of reduction, within this tight variation process. The justification for this strategy would be primarily the core and control model, and also the argument that continuing the adaptative approach will provide solid automatic retrieval systems that can operate reasonably in all but totally unusual setups, and can do so at low startup and maintenance costs as well as low user effort in request formulation. A further argument for this view would be that far more of the running in the modern Web engine (and large bibliographic database) world has been made by developments in the nature of the documents in the file than in other constituents of the environment. Moreover, in this modern world, the growing volumes of log data of one sort or another can provide increasingly substantial and detailed indicators of users' contexts, even though much of the external setup within which the user works is bound to remain inaccessible.

But it is not clear that the only conclusion to draw from these arguments is that there is no need to investigate motivating contexts more directly. Thus both contexts themselves, even just IT contexts, are becoming more complex and richer, and contexts or context types may represent significant user communities for whom one would wish to offer a better than 'basic benchmark' system, for example whole large biomedical or information analysis communities.

In my view it is therefore proper to suggest that TREC needs to engage, more positively and fully, with context and the nature of the whole setup information-seeking task T rather than just the experimental task X. That is to say, while TREC has been slowly acknowledging the influence of the larger setup, I believe that it is important for TREC's system-building participants to be encouraged to work forward from a fuller knowledge of the context, rather than limiting their attention to the attenuated form of the context that the D * Q * R environment normally embodies, and recovering what may be distinctive about this - and hence somewhat indicative of significant features of the larger context - by working backwards from system results.

This is not to imply that there has been *no* reference to real contexts in the TREC fifteen-year Programme. In recent years, in particular, there has been an increasing pressure to get real. But the modus vivendi (except in QA with its sister AQUAINT effort and, to a non-trivial extent in Genomics) has been to a large extent responsive to random observations, to the communities informal collective perception, even to gut feelin,g as well as to its own members' experiences as actual retrieval system users, about where the real retrieval world is now; or if not thus quite informal, has been driven by the need, having thought of using a new class of document, to have requests and assessments that consort with them. The result, from the point of view of that understanding of retrieval that the TREC Programme seeks, has been essentially arbitrary. Real contexts, or rather real context clases and types, figure only accidentally, and partially, in the TREC tests, insofar as they can be pinned down within the standard D * Q * R model.

Controlled experiment requires some abstraction. But the abstraction has been from miscellaneous starting points, so the TREC experiments as a series have not been consistent and coherent. Covering the problem of how to design retrieval systems to support information-seeking needs has been more by energetic walking in random directions than by organised quartering of the terrain. But as TREC has already done a good deal of exploration, my argument is that it now can, and should examine real information-seeking contexts more thoroughly, *as wholes*, and work from this to the viable and useful experiments that, even when pared down by the intellectual and practical requirements of a TREC-style programme, will still invoke context more than TREC has hitherto, and hence allow the particularising predictions that we should be looking for.

```
Figure 2 - SUMMARISING FACTORS   (from SJ01)

INPUT FACTORS

a) the FORM of the source
   e.g. academic article; news message; progress report
   this subsumes
     structure e.g. header, amplification ... ; introduction, background, objective ...
     scale (or length) e.g. book; article; 10K words; 1K words
     medium e.g. English; Shipspeak
     genre e.g. description (as in an encyclopedia); narrative

b) the SUBJECT TYPE of the source
     ordinary e.g. daily news
     specialised, alias technical e.g. drug chemistry
     restricted, alias local e.g. organisation plans

c) the UNITS taken as source
     single e.g. scientific paper
     multiple e.g. news stories
   (a single source may have several subunits e.g. book chapters; the distinction is
   whether the set of units to be summarised was intended to go together or consists of
   independent members)


PURPOSE FACTORS

a) the SITUATION, i.e. context, within which the summary is to be used
     tied e.g. to company X's marketing drive for product P on date D
     floating e.g. Computing Reviews - for anyone interested for any reason

b) the AUDIENCE for a summary  can be characterised as
     untargetted e.g. a mass market womens' magazine's readers
     targetted e.g. UK family court lawyers
   (the former covers a wide variety of skills, experiences and interests,
   the latter is much narrower)

b) the USE, or function, for which the summary is intended
     retrieval aid e.g. Web engine results page snippet(s)
     preview device e.g. report executive summary; submitted paper abstract
     refresher e.g. old report resume
     alert e.g. advertising blurb for novel
```

```
OUTPUT FACTORS

a) the MATERIAL of the summary, i.e. the information it gives,
   in relation to that in the source
     covering i.e. all main concepts of source in summary
      e.g. aims, background, data, tests, results, conclusions of biology paper
     partial i.e. some concepts (types) only
      e.g. experimental methods in biology paper

b) the FORMAT of the summary i.e. the way the summary information is expressed
     explicitly structured presentational layout
      e.g. headers and slots in course synopsis; boxed examples in textbook
     running text (primarily)
      e.g. summary review of novel
     non-text
      e.g. graph, cartoon

c) the STYLE of the summary i.e. the relationship to the content of the source
     informative i.e. explicitly giving source content
      e.g. restatement of facts given in an accident report
     indicative i.e. indicating area of source content
      e.g. declaring a biography is about some person
     critical i.e. evaluating the content (perhaps other features) of the source
      e.g. claiming a novel is well written trash
     aggregative i.e. setting parts of the source material against one another
      e.g. laying out arguments for and against a policy in a meeting record

d) the EXPRESSION of the summary i.e. all the linguistic features of the summary
   this subsumes
     language e.g. English; French
     register e.g. technical jargon in legal abstract; popular writing for newspaper
     modality e.g. narrative for newscast; description for weekly magazine

e) the BREVITY of the summary i.e. relative or absolute scale (length) of the summary
     e.g. 5 percent of source; half-page; 100 words
```

To illustrate something of what a fuller view of context might involve I will exploit the factors framework presented in SJ01. This was developed for thinking about automatic summarising, and is shown with examples in Figure 2. It does not, of course, carry over in detail to the document retrieval case. But its main elements apply. There are also problems in seeking to characterise large heterogeneous retrieval worlds with even the only moderate definiteness that Figure 2 shows for particular summarising contexts. Nevertheless, as with the kind of evaluation decomposition illustrated in SJG96, going through the business of trying to specify an entire setup, even if with only limited success and incomplete detail, and examining the result, can be very useful.

The factors framework refers to Input Factors (IF), Purpose Factors (PF), and Output Factors (OF). Input Factors and Purpose Factors constrain the set of choices for Output Factors, but for any complex task cannot simply determine them.

Under IF for summarising we have properties of the source texts. This includes their Form, Subject type, and Units, which subsume a series of subfactors, as illustrated in Figure 2. It is not difficult to see that such factors apply, in the retrieval case, to documents. They will also apply, in retrieval, to requests, past and current, and to any available past relevant document sets. The particular characterisations of documents and requests may of course be different, e.g. documents but not requests might have a complex structure.

Language modelling seeks to capture the distinctive characteristics of D, Q, and R and to relate these, but this cannot be deemed to exhaust all the potentially useful ways of describing D, Q and R, including such old-fashioned but serviceable ones as those illustrated. There is, equally, no reason why the IF set should not be expanded if there are available and pertinent features to consider. For example, while in the automated retrieval case such old-fashioned physical document properties as cover binding might not matter, it is easy to see that factors like originator/authority status might.

It is also the case that IF as envisaged here are essentially static (even if summarising over a set of documents may be over a temporal sequence). If we include request properties in IF we may have to consider how to handle dynamic user interventions such as request modifications, which may be messy but does not seem to be a logical problem.

But the more important considerations are those relating to Purpose Factors. PF include Situation, Audience, and Use or function. Context, as I have been using the term, includes all of these. As illustrated for summarising, Situation includes the features of the context that bear on its particular function. For summarising, for instance, they can include temporal constraints, the presence or absence of other complementary or supporting material, e.g. images with which a text summary is to be associated, the larger role of the whole setup e.g. marketing or research departments in a company. It is easy to see similar setup properties being pertinent to retrieval, as exemplified by filtering and spam detection, and the role of retrieval within question answering also illustrates Situation characteristics.

The Audience and its characteristics refer to the users of the system output, in the summarising case those who exploit the summary, in practice for retrieval systems the user community or identifiable subcommunities as well as the specific user who is responsible for a particular request. Audience is a familiar contextual element for retrieval. The fact that Audience reaches beyond the individual is implicit in the assumptions that Web engines may make about preferred generic page properties. Audience is defined here as the direct consumers of the system output. There may be other people in the setup as a whole with properties bearing on the Situation, for example in retrieval or filtering those who are responsible for the organisational group using the system output.

The third PF factor, Use, or function, is the obvious focus of the whole, namely the precise service that the system output has to perform, Figure 2 illustrates some examples for summarising. It is easy to see the retrieval analogues, and TREC has already studied some possible distinct retrieval functions, for example topic search designed to offer a set of documents about something for user scanning or reading, and known item recovery. It is true that in general with TREC test data we have not known what the precise functionality required for the system output is, and so have tended to fall back on 'offer documents about X' for the user to do whatever they want with, buttressed by the tacit assumption also made by Web engines that if you supply ranked output (especially, in their case, supplemented by metadata and snippets), this is hospitable to a range of functions that are implemented simply by the user's response to the ranking, e.g. pick one item, or slog earnestly down 200, etc.

Finally, Output Factors. As the examples in Figure 2 show, IF and PF will set useful limits on the range of specific options to be considered. Thus a combination of IF and PF might imply a requirement for very short, plain English summaries, but there is still a choice about e.g. telegraphese or full sentences, and e.g. bullets or running text. For summarising the set of OF includes Material, Format, Style, Expression and Brevity. There are analogous choices with retrieval. For example, does the system immediately show whole retrieved documents or just extracts, does it offer fancy window displays or a minimum list of document titles, does it assume the function of the output is to supply enough information to be a substitute for the whole document or just a pointer to it. TREC has studiously ignored, as a direct consequence of its way of measuring performance, this whole aspect of what real systems have to do (except, rather minimally, in the Interactive and HARD tracks). But in practice OF cannot ignored, and have to figure in global performance evaluation.

5 From generalising to particularising

The foregoing is only an informal discussion: more thorough analysis of retrieval contexts is needed for a factor characterisation to be taken seriously as a basis for system development. My argument here is that in its next stage TREC needs to engage with retrieval context more than it has hitherto.

One reason for this is my claim that TREC has to a considerable extent, at least for the traditional system view of the retrieval task, established its credentials. It has done this for both methodology and results, bringing its way of determining performance, and the retrieval system models these have supported, decently up to date. The other reason is that, as Ross Wilkinson pointed out, this progress, however considerable, has not so far enabled the research community it represents to say: 'If your retrieval case is like this, do this', as opposed to 'Well, with tuning, this sort of thing should serve you alright'. Our way of generalising has in fact served to distance us further from the particular starting point that any actual retrieval case - the setup with both system and context - offers.

The non-TREC literature refers to many studies of individual retrieval setups: what they are about, what they are for, how they seem to be working, how they might be specifically improved to serve their purposes better. The generalisation goal that the automated system research community has sought to achieve has worked against getting too involved in the particularities of any individual setups. My argument here is that, in the light of the generalisation we have achieved, we now need to revisit particularity. That is, to try to work with test data that is tied to an accessible and rich setup, that can be analysed for what it suggests to guide system development as well as for what it offers for fuller performance assessment. We need to start from the whole setup, not just from the system along with whatever we happen to be able to pull pretty straightforwardly from the setup into our conventional D * Q * R environment model.

Trying to particularise, as the first rather than last step, will not be easy. Even if we can describe a setup in detail, this will normally still be an informal account and not one an automated system can exploit. For example, given a contextual requirement for introductory literature on topic T, how can this be expressed so as to be a usable guide for an automated system, and also underpin performance assessment? How can a system build effectively on the statement that all the users in this setup are very busy people and need stuff they can absorb in ten minutes?

This sort of thing is extremely hard. But I do not believe that we should therefore not attempt to do it or argue, in a supposedly more principled manner, that setups within which modern retrieval systems have to operate are so diffuse, or so variegated, that it is a fundamental mistake to address anything but the immediate D * Q * R environment from which solid, transportable, general-purpose retrieval system knowhow can be acquired. In fact, indeed, TREC's newer tracks subvert both of these arguments: even if the lawyers' interpretation of "relevant" as referencing might be inferable from assessment data samples, one feels rather less confident about being able to infer, even with the best modern machine learning tools, that the name of the retrieval game is getting information that "appears reasonably calculated to lead to the discovery of admissible evidence". Being told that this is a feature of the legal context does not make it easy to apply automatically, but taken together with other factors in a full initial characterisation of the legal context is a better basis for system development than starting with some bunch of D * Q * R, concentrating on making inferences from this, and then sucking in more miscellaneous context facts only as these seem to be required. Such full context characterisations are, moreover, more likely to lead to explanations for results than system tuning, even of the most ambitious learning kind. The challenge, however, is not so much that of converting rich, informal contextual detail into something that a system can exploit, as of not doing this in a way, in the interests of controlled scientific experiment, that throws the baby out with the bathwater.

What I am advocating may be deemed just reinventing, or reinvoking, the unattainable holy grail that earlier retrieval system research deemed not merely unattainable but ineffable. There is also no doubt that the very real constraints of obtaining test data (and especially reusable test data), and of conducting scientific experiments, have militated against what may seem like hopelessly soggy desiderata. However biologists know that you only get so far with studying creatures divorced from their ecologies. One might say now that we've learnt quite a lot about the retrieval system creature, and that we need some ecology study in order to design bigger and better bumble bees.

References

Sparck Jones, K., and Galliers, J.R. *Evaluating natural language processing systems*, LNAI 1083, Berlin: Springer, 1996.

Sparck Jones, K. 'Factorial summary evaluation', *Workshop on Text Summarisation*, ACM-SIGIR Conference 2001, 2001.
(via http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/cambridge2.pdf)

Voorhees, E.M. and Harman, D.K. *TREC: Experiment and evaluation in information retrieval*, Cambridge MA: MIT Press, 2005.