Assumptions and issues in text-based retrieval

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK
February 1992

This paper in its final form appeared in *Text-based intelligent systems*, (Ed. P.S. Jacobs), Hillsdale NJ, Lawrence Erlbaum, 1992, 157-177.

1 Introduction

This paper is intended to provide an analytical background for those seeking effective text retrieval systems, and more specifically for those advocating the application of techniques drawn from artificial intelligence (AI) and natural language processing (NLP) for this purpose. A new wind is blowing through the world of information retrieval, and it seems that some of the apparent limitations of existing methods for characterising and retrieving text-based information can be overcome. These existing methods refer primarily to information retrieval in the sense of document retrieval. Much of what is proposed falls, explicitly or implicitly, under this broad heading; and it is also useful to approach other forms of information retrieval from document retrieval, to make their distinctive properties and implications clear. My aim is therefore to lay bare the nature and conditions of document retrieval as these have hitherto appeared, in order to provide the context within which new and hopefully better retrieval strategies can be defined and developed. The experience of the past shows that information, i.e. document, retrieval in general is an intractable task, and thus also an intractable task for automation seeking a high level of performance. This implies that it may be harder than expected, in the general case, to make radical improvements with new techniques. But these techniques should certainly be investigated, and they may, as indicated in my conclusion, provide real payoffs in some types of context or in individual applications.

2 Motivation

My starting point is therefore that developments in computing technology, and in artificial intelligence and natural language processing, have stimulated interest in information retrieval from those outside the established library and information science community, and have led to suggestions that the time is ripe for new approaches to retrieval. These are particularly associated with the use of the full texts of documents, which are typically not available in conventional retrieval services, and with the idea that AI and NLP offer distinctively new approaches to text characterisation and searching not found in conventional systems. There is also an interest in types of material, for instance news stories, not generally covered by conventional bibliographic services, and in direct searching by end users, typically armed with high-class workstation facilities.

It is often assumed that what is done conventionally, or has been done in past information retrieval research, is inadequate or irrelevant in these new contexts. But as this assumption may well be based on lack of knowledge or experience, it is most important that, when

approaches are claimed as new, they should be related to important distinctions and justified accordingly. In this case, these distinctions are between doing something automatically that has hitherto been done manually, producing the same type of output intended for the same type of use; doing something automatically which is quite different from what has hitherto been done manually or automatically, but is still intended for the same sort of use; and doing something novel automatically which is also intended for novel uses. In the present context, these distinctions are crucial for document characterisation, i.e. indexing. In the first case novelty is only in the means, not the end, and can only be justified by better (or cheaper) retrieval performance. In the second case novelty is in the means as well as in the end, but has still to be justified in the same way, by better (or cheaper) performance in the same generic context. In the third case the nature of the new context, and especially the nature of new information uses rather than just of new materials, has to be understood. It is further necessary, in this case, to establish appropriate methods of performance evaluation, and also to check the performance of new indexing and searching resources designed for the new contexts against older approaches rejigged for the new contexts.

This paper spells out properties of, issues for, and experience with, document retrieval, to provide a background for developing and evaluating new approaches to information retrieval, and specifically approaches which stem from the application of NLP and the use of full text. It therefore considers the findings of past retrieval research and the potential role for NLP in document and text indexing; the implications of past retrieval experience and of retrieval constraints for NLP-based indexing; the consequences of alternative applications of NLP to create autonomous information bases; and the requirements to develop the necessary evaluation techniques for retrieval performance in novel contexts, and especially those involving highly interactive searching and mixes of different information-seeking activity.

3 Automatic indexing research

Information retrieval (IR) has conventionally referred to document retrieval, and specifically to automatic document retrieval. It has normally excluded searching for known items, like finding the storage location of a known book using an author or title catalogue, and has thus focused on finding documents relevant to information needs as expressed by subject or topic requests.

In the initial development of automatic retrieval systems, the basic assumption was that documents would be *indexed*, i.e. would be represented by brief subject or topic characterisations on which searching is actually carried out. Intellectually, the essentials of an automatic system were the same as those of manual ones, focusing on indexing as the summary indication of key document content, and hence on strategies for providing good descriptions and for finding document descriptions appropriately matching request ones. Automation nevertheless allowed two practical novelties, with long-term intellectual consequences. One was the ability to permute and select, so descriptions could be decomposed and reconstructed to allow multiple views of topics. The other was the ability was to search text directly, for instance abstract texts, so document descriptions could be formed, through matching, at search time.

Operational automatic retrieval systems have developed in two ways. One has been to retain manual indexing using subject heads or thesaurus descriptors, i.e. controlled language terms, combining this with the search time exploitation of Boolean request structure, and providing support for the selection of indexing and search terms through index language

classification schemes embodying hierarchical and other relational structure. The other development has been in free text searching on keywords, though normally again with requests constructed using Boolean operators. The perceived, and real, problems of both of these have been associated on the one hand with the opacity of controlled index languages, on the other with the weakness of uncontrolled natural language, and on both hands with the rigidity of Boolean requests.

Information retrieval research over the last two decades has suggested, indeed demonstrated to the limit in non-trivial experiments, that controlled and natural language indexing and searching are competitive in fair comparisons, achieving the same middling level of performance (Cleverdon 1967, 1977, Salton 1986, 1991, Salton and McGill 1983, Sparck Jones 1981, Willett 1988). This research has also indicated the value of much more flexible request formats than conventional Boolean formulae, with free term coordination offering ranked output, and has shown that statistically-based keyword weighting is useful. The research has further demonstrated that relevance feedback techniques of an essentially statistical kind can also be very valuable (Salton and Buckley 1990, Sparck Jones 1980, Sparck Jones and Webster 1980). Iterative searching is of course normal in conventional contexts, but this research has shown that it can be effective with little effort on the user's part. This is important because it is hard to provide effective support in search development for the end user.

These superior techniques stemming from research have begun to be implemented, though not widely in conventional system contexts (Debili et al 1989, Doszkocs 1983, Harman and Candela 1991, Porter and Galpin 1988, Sanderson and van Rijsbergen 1991, Stein 1991, Willett 1988). It is essential to recognise that all the experiments done so far have shown that the research methods are superior to those implemented in normal operational Boolean keyword systems, which have given natural language in information retrieval a bad name. It has however to be accepted that these newer natural language techniques have not been rigorously tested on a really large scale. Thus the largest serious experiments have been with data of order 150 requests and 30,000 documents, and most comparative testing has been with much smaller sets.

4 Opportunity and challenge

The main new development of recent years has been the growth of full text sources. This is taken to open up striking new possibilities for improvements in information retrieval. Thus it is widely believed that searching full text directly, without the impediment of index descriptions, will provide both immediate and superior access to the information the text embodies, and is thus naturally to be preferred to working with index descriptions instead.

At the same time developments in both NLP and in AI appear to offer appropriate strategies for capturing this text information and making it accessible to the topic or concept-hunting user. The approaches stemming from NLP and AI can be broadly labelled *meaning-oriented* and *fact-oriented* respectively. This paper is primarily concerned with the first, i.e. with meaning-oriented information management, so it considers fact-oriented information retrieval only, later, where the comparison is important.

The starting assumption is therefore that what is required is to determine and represent the meaning of a text, so retrieval, operating on a similar representation of request meaning, is a matter of establishing sameness or similarity or some other relationship of meaning between document and request representations. Thus to take a not very extreme example,

a document might be represented as a structure of syntactically normalized, semantically resolved propositions, and a request as a similar but much smaller set.

4.1 Indexing

The crucial issue here (assuming that this sort of NLP can be done) is apparent in the question: what does a request-document match imply? That is to say, supposing a request sentence and one of the document sentences convey the same or sufficiently similar propositions, what does this tell us about the relevance of the document to the request? It may seem obvious that the document is relevant, but this is not necessarily so.

The reason why things are not so simple became apparent when full text was offered for keyword searching. Word matching on titles or even abstracts could be as effective as matching previously-constructed index descriptions consisting of lists of manually selected words because, on the whole, words in titles or abstracts reflect the importance of the concepts they refer to in the underlying full document. This is not the case with matches straight against the text. A word can occur in a text but be very unimportant for it. The same holds, though somewhat less disastrously, for a proposition. Thus those engaged in keyword indexing were obliged to invoke statistical selectivity measures designed to distinguish important from unimportant word occurrences with respect to individual texts. For instance a word occuring with medium frequency in a collection of documents as a whole, but with very high frequency for a single text, may be taken as a significant content indicator for that text (van Rijsbergen 1979).

The important point about index descriptions, in other words, is that their function is not simply negative. They are not a regrettable substitute for full text, which can be jettisoned with more and cheaper machine storage. They have a vital positive function which is to indicate the important, main concepts or message of a text. This still leaves open what exactly is meant by this, how much is selection, how much generalization, and so forth, questions which can similarly be asked about abstracts. The major difficulty about indexing, illustrated by comparing indexing done by different human indexers, is that what is important is not unequivocal, or permanent.

Index descriptions are thus reductive, simply because not everything in a text is important. But index descriptions were formerly, and still are, also reductive for the simple good reason that human beings cannot read every text to find out what it is about, i.e. index descriptions have the same vital filtering function as titles. They will still have this function in any system involving significant user interaction and non- trivial amounts of material. Moreover even where users are happy themselves to work directly on full texts without prior filtering, there may be a subsequent or supportive role for abbreviated descriptions in internal file structuring linking one document with another. Thus even if users always want in the end to access a full document text not just to read it for its information content but (possibly at the same time) to assess it, in fine detail, for relevance to their need, index descriptions have an essential role as prior filters embodying a condensed characterisation of a document. At the same time, index descriptions may need different forms for human and machine consumption. This may be a matter simply of presentation, for instance offering keywords in phrasal rather than alphabetical order; but real differences may be justified by the intrinsic differences between the ways humans and machines manipulate information.

There is, however, a further constraint on index description, which the earlier work on the full-text keyword indexing served to establish, though it was also recognised in keyword

operations with e.g. title terms. This is that it is not enough for a document description to be a good description of the document itself. It also has to be discriminating. Thus given that descriptions are reductions, they naturally reduce the difference between documents, in just the way that the same two or three words may be used as the title for very different books. However as the need the user has will often, though not necessarily, be for relevant information at the more detailed level of the full text, descriptions should as far as possible balance accuracy of description with distinctions between descriptions.

What all this implies, for those who believe that what is needed in information retrieval is a "one-for-one" representation of a text established by using NLP, is as follows. Full representations are either required in their own right, or as a means to the end of reductive indexing. In the first case, as defined by data properties or search purposes, retrieval means retrieval on full representations. This may be done directly or in two stages via reductive descriptions derived from the full ones; but either way, it is necessary to show that the file data or the search purposes force the use of full representations in order to serve retrieval needs adequately (utilising them because they are cheap and good enough, and selectivity for extra performance is too costly, is a separate matter that has to be justified in its own terms). In the second case, full representations are the necessary precursors to reductive index descriptions. However it is then essential, if the full representation is only a means to the end of reductive indexing and is not preserved, to demonstrate that the desired nature and/or quality of the indexing cannot be obtained without going through the full representation.

In this context it must also be emphasised that if the user is directly involved in searching, he must either be able to understand the form of a representation or have it translated for him, and that this is especially critical with full representations.

5 Potential roles for NLP

Now consider the case where indexing is explicitly accepted as the goal of the full-text NLP, so intermediate representations of whatever sort, and not just full ones, are jettisoned when they have been exploited to provide index descriptions. The presumption is that NLP will give better indexing than the current keyword standard. But it is essential here to be clear about the exact nature of the claim that is being made.

One form of the claim is that NLP analysis (and perhaps generation) will give better indexing (and associated searching) than, for example, conventional Boolean keyword systems. But this is misconceived goal, since while these systems for a variety of reasons do not perform well, they can be improved on by the superior word-based strategies of information retrieval research. Thus the correct comparison is with these research-based techniques for term selection and weighting (just as new cars should be designed to work better than this year's cars, not last year's.) It is also necessary in these comparisons to make proper checks on the starting points. Thus sensible request formulation is vital for reasonable performance: this is part of the controlled language operation with a skilled intermediary, and needs to be provided for in other ways (even with relevance feedback as a boot strap) with text-based approaches - as it was not obviously provided with Blair and Maron's STAIRS investigation (Blair and Maron 1985). These points are general ones: there may be circumstances where, given an institutional Boolean system, performance might be improved by using NLP to give better keywords for Boolean searching which explicitly combines different search fields (Rau and Jacobs 1991).

Another possible, though less frequently encountered, claim is that NLP on full text will provide better index descriptions than the conventional ones using thesaurus descriptors or subject headings, given the underlying presumption that this sort of indexing is better than raw or even improved keyword indexing. This claim may be associated either with the same degree of reduction as in conventional indexing, when this is in fact done from full-text rather than abstracts, or with less reduction, yielding fuller or more complex descriptions. In the second case the value of more extensive or exhaustive descriptions would have to be demonstrated, taking into account the various factors like increased matching potential which have already been investigated for manual indexing. But these descriptions would still be reductions on their sources, and in general, the advocates of NLP for indexing have not considered how reduction is to be achieved.

However the main thrust of the argument for NLP is either that applying NLP, whether more shallowly or deeply, would deliver the same sort of result as conventional manual indexing, but a better quality one, or alternatively that it would deliver a different and better kind of index description. (Of course these claims could also be made for abstract or even title processing.) These claims are normally based on the view that NLP, perhaps supplemented by AI-style inference, can provide a better concept *identification* and better concept *representation* than has so far been achieved.

The identification claim is typically associated with the view that the component terms of a description can only be properly recognised by using information about syntactic/semantic structure in the text, i.e. about constituent relationships and/or functional roles. The representation claim is typically associated with the view that the representation itself has to have a syntactic or semantic structure indicating the constituent relationships and/or functional roles of its terms. The representation claim may also be associated with the view that description involves normalization, not just of structure but of vocabulary, for the same reason that in conventional thesaurus indexing ordinary language words are replaced by controlled language terms.

These two aspects of description are quite independent, and conventional indexing can vary along both structure and vocabulary dimensions, covering both more or less syntactic structure, more or less regularised syntax, and more or less explicit syntax, with varying degrees of vocabulary control (Chan et al 1985, Lancaster 1972, Lancaster et al 1989). Thus when proposals for more sophisticated indexing based on NLP techniques refer to representation, i.e. the nature of the index descriptions for documents, they can refer to complex natural language descriptions of the same kind as e.g. titles, or to descriptions combining natural language words with constrained or artificial syntax, as in PRECIS (Austin and Digger 1985), or to descriptions with both vocabulary and syntax in a specialised artificial indexing language. Clearly there are quite different implications for the user in these different types of description, and particularly in the use of indexing languages imposing artificial constraints on the form and content of descriptions. However what follows to a large extent applies whichever of these styles of index language is adopted.

The crucial point now is that the view that NLP (without or with AI) is needed to deliver sophisticated descriptions, for the uses that ordinary indexing descriptions are put to, cannot properly be based on hoary examples of the kind that syntax is needed to distinguish blind venetians from venetian blinds. Nor should it be based on the assertion that keyword searching of the kind often implemented in legal services delivers poor results and that more sophisticated indexing would obviously deliver better results.

6 Past retrieval experience

Assertions like these may based on an inadequate grasp of the facts, on the one hand about the realities of retrieval and on the other about the history of retrieval testing. Thus for example, and just to begin with, a collection may not have documents both about physical disabilities and interior decoration; or search descriptions with the necessary discrimination can be very readily achieved simply by adding further terms to the request, like "sight" or "curtains", which is useful anyway since increasing the number of term matches increases the chance of relevant retrieval. Equally, quite apart from the fact that simple natural language indexing can be used more effectively than in conventional keyword services using Boolean queries, information retrieval research since the late fifties has been largely concerned with index language design and performance, and specifically with the design and performance of manual indexing languages and descriptions. The range of languages and methods developed and investigated has been very large, subsuming both approaches applied in serious or large-scale operational services and in more experimental ones. These performance evaluations have covered not only the nature of the indexing resources themselves, but also relevant matters like the effects of care in indexing, and a host of other issues like indexing exhaustivity. One of the major features in particular of the research has been comparisons between different indexing languages and forms of description (Sparck Jones 1981, especially Chapter 12, Salton 1986).

This work is relevant to current proposals for automated indexing and retrieval using NLP and AI techniques for two sorts of good reason.

The first reason is that these earlier proposals and tests referred to indexing notions of the same general kind as nowadays proposed, i.e. with relationally motivated and structured compound terms or complete descriptions, and also studied them in many individual particular forms covering a very wide range of possibilities. Some were indeed implemented automatically, see e.g. Bely et al 1970, but this is not the important point. The important point is that the end indexing styles were the same as those now proposed, so whether they were effective in use is what really matters, not how they were achieved. Thus those advocating modern versions of these methods have the obligation to look at what was advocated in this work, as a necessary preliminary to claims for superiority or difference. Moreover even if implementation quality and consistency has also to be taken into account, the quality achieved in the past has to be established as inferior to that likely to be achieved now, meaning, notably, that past human indexing has to be shown to be less effectively executed than the proposed automatic indexing.

The many studies done in the past in particular showed both that performance for quite different techniques, when seriously applied, was much the same, and thus that simple techniques were very competitive with more sophisticated ones, and that absolute performance is not high (Sparck Jones 1981). Those who want to make legitimate claims about the superiority and novelty of their approaches to indexing need to look much more carefully at conventional indexing in all its variety and in all its aspects - philosophy, implementation, index language design, indexing description principles and so forth (Chan et al 1985).

This is particularly important because the current focus of attack is on indexing and, more particularly, on the way documents are described. The evaluation tests done in the past showed how important other factors are, and in particular how important requests are. It is more helpful to devote attention to determining the user's *need* and to expressing this as a request than to fiddling with individual documents, particularly when searching can be

iterative so that if relevant documents are not found first off in one way they may be found later in another. The details of indexing languages may thus not be particularly important. For example, how much does recall (getting all the available relevant documents) matter to the average user? Languages and descriptions may or may not be designed to promote recall. Thus the real challenge of information retrieval is the indeterminacy, complexity, and variety of users' needs, and the correct approach to developing indexing and searching techniques is to relate these firmly to the properties of users (Belkin and Vickery 1982, Hewins 1990, Saracevic et al 1988).

The second good reason for taking past research on board in the context of current interests in NLP-driven indexing is that this research has served to establish investigative methods and evaluation techniques. Performance testing in information retrieval is far in advance of that in NLP, so those moving from NLP to information retrieval need to know what is involved, for example, in choosing measures or gathering data samples (Sparck Jones 1981). It is true that even the largest tests have been limited, given the size of major operational services, so the results obtained may not scale up. But this is a problem for new NLP-based techniques as much as for older ones, whether conventional or the products of earlier research with e.g. simple natural language term approaches. Even so, major research projects have conducted many hundreds of runs just to establish quite basic propositions (Salton and Buckley 1988, Sparck Jones and Webster 1980, Willett 1988).

It may, however, be that NLP is advocated not as a means of generating sophisticated descriptions as wholes, but as a means of making more sophisticated choices of simple NL terms than the research-based statistical ones. For example it may be thought necessary, given a simple coordinated-term style of indexing, still to allow for terms which are multi-word units, although with implicit rather than explicit relationships. Here again, past research investigating the relative merits of syntactically motivated units, statistical phrases, and simple de facto coordination at search time is relevant (Fagan 1987, Keen 1991, Lewis et al 1989, Salton and McGill 1983). The same applies to the most limiting case of NLP, where analysis is used to identify individual words satisfying conditions like e.g. being nominal heads. This again has to be compared with cruder approaches (e.g. all content words), and like all the other techniques, has to be related to the statistical properties of terms which are relevant in indexing, whether for the whole text, or for the collection. One of the important challenges for any NLP-based indexing is to combine it effectively with statistical information. This may seem simple if collection-based information is used for selection or weighting of individual terms, but is more complex in phrase identification where the components of a phrase have different statistical properties.

7 Retrieval constraints

All of the foregoing has been concerned with indexing aimed at meeting retrieval needs of the usual sort i.e. for documents relevant to some topic, and has been aimed at reducing ignorance about this. Indexing here has to be based on an understanding of the intrinsic problem character of this situation and so, whether applied to documents or requests, has to address the problems of the choice of descriptive items, the internal structure these descriptors have and the structural relations between them, and the lexical normalisation that is required. In general, the closer to the actual text the indexing is, the more matching requirements have to be met by the orthogonal provision of a vocabulary normalisation apparatus in the form

of a thesaurus or whatever. This again has to be grounded either in the view of vocabulary organisation characteristic of conventional thesauri or in more recent approaches based on statistical or relevance associations. Without this apparatus to support matching, the user has to contribute more, by explicitly indicating alternative expressions for the same content.

Finally, it is increasingly important to address the user interface, and specifically the end user as opposed to professional search intermediary. Modern technology offers great opportunities here, but those engaged with online public access catalogues (OPACs) have already learnt how hard it is to make sure that the non-professional and particularly occasional user is able to search effectively (Borgman 1986, Mischa and Lee 1987). This is an active area of research, but it is as necessary for those offering supposedly superior types of indexing as for those offering traditional forms (whether automatically obtained or not), to show how end users can deploy the indexing information that is supplied effectively. Thus the more complex indexing is, the more difficult it may be to understand and use: this is true even though there are also issues about helping the end user enough, for example to find alternative words, when simple natural language techniques are used, whether these are of a conventional or a research-based kind.

It is therefore necessary to demonstrate that end users are able to manage more sophisticated forms of indexing and their associated retrieval operations, which has not proved easy with conventional subject headings, classification schemes or thesauri, whether of an older fashioned or newer associative kind (Keen 1977). This is an area where expert systems methods have been applied, since these may be used (as in Pollitt 1986 and Vickery et al 1987) to hide the technical complexities of the actual indexing required for the search specifications from the user, while helping him to formulate his need. At the same time, modern interactive technology, with windows and so forth, can make displays more effective and housekeeping during searching more efficient. But though it may, for example, be easier to display classifications with modern technology, they may still not be easy to understand and use. Thus one important area of information retrieval research has been in extracting search information painlessly from users by exploiting relevance feedback, simple judgements of whether documents are acceptable or not without any indication of why, since the system infers this.

So far, I have been concerned not only with retrieval of the 'usual' sort as far as topic specification and matching are concerned, but also with what may be described as typical retrieval contexts, for instance involving retrieval from masses of journal articles. Indexing and retrieval schemes have of course in the past been designed for more specialized situations, whether these refer to the type of material, or to the form of usage (i.e. properties of the user community and its 'requests'). One example is the use of facetted classifications for company libraries. Thus while it may be argued that the need for sophisticated and deep indexing in general contexts has not been demonstrated, this may be required in special contexts. This may follow from the nature of the material or the nature of the needs, but the case has to be carried through, not just taken for granted. Moreover the point just mentioned about whether end users can manage sophisticated, and especially constrained and artificial, indexing language and descriptions still applies.

The essential issues with full text retrieval are therefore as follows. Direct searching on full text, when there is a great deal of text, is either not practical for the human user because he will be swamped, or not sensible because he will fail to reach items that matter. There has to be a means of access, i.e. indexing (and whether this is best, or has to be, done at file or search time is irrelevant here). If there has to be indexing, does better retrieval performance require sophisticated indexing going beyond simple NLP strategies, and especially essentially

statistical ones? If it does, how easy is it for the end user to work with descriptive terms and structures which are not ordinary natural language ones, but are only more or less arbitrarily related to natural language? If the user cannot work with descriptions of this sort, how well can he operate with plain natural language terms, given the mass of data available for them, and the size of the files he is searching? All the evidence is that complex natural language expressions are of no material use as units for searching, however important whole phrases or sentences may be, as they are in the case of titles, as supports for search output assessment. But if the user is left starting from words, how can he manage e.g. extensive collocational or associative information about words, so as to be able to improve his search specification? The real challenge with full text is how to benefit from the opportunity offered by direct, text-based searching without being overwhelmed by masses of easily retrieved material, which is precisely what relevance feedback techniques are designed to do.

Whether sophisticated indexing, to be applied in a way which is entirely hidden from the user, is required and can be supplied in a superior form through novel NLP techniques is a separate matter. It of course has to meet all the criteria already mentioned for overt rather than covert indexing, and has to be justified, as overt indexing does, by rigorous comparative evaluation. However there is also the additional requirement that all of the system's description and search operations exploiting the indexing have to be driven by automatic transformations of the natural language and text data the user sees, and formulating effective searches under these constraints is not obviously easy. This transformation job is what the professional librarian and intermediary does in ordinary information-seeking environments.

But though I have so far been concerned with indexing, i.e. with meaning-oriented information description, it is, however, also possible to see information retrieval in a quite different light, as not concerned with indexing for its conventional access purpose at all. Thus the suggestion that complex indexing descriptions are required may stem from the belief that many information management activities are carried out solely with the document descriptions. This belief takes the traditional use of descriptions as scanning aids to identify source documents to the point where the descriptions can be seen directly as primary sources of information in their own right, just as abstracts may be.

8 Creating information bases

Using descriptions as information sources in their own right leads to the second major current line or work in NLP and AI- based information retrieval. This treats document descriptions not as access aids, but as substitutes for their sources, giving all the essential information of the sources in a more explicit, or regular, or other more convenient form. Modern approaches to message processing for example, where natural language originals are replaced by instantiated frames (Lehnert and Sundheim 1991, Young and Hayes 1985), sometimes illustrate this strategy, though it was followed much earlier in Sager's work (Sager 1978). In some message processing applications there is no or very little reduction, so the representation can for many purposes be taken as a substitute for the original. Effective reduction is more difficult to achieve (see DeJong 1982's summarising) and it also follows that the sources must remain available. (In some message processing cases, as in Sager's work, the frame fillers may be only slightly normalised, and preserve much of their original natural language character.)

These message processing examples illustrate the case where the set of descriptions can be treated as an aggregated knowledge base, in the way many record catalogues constitute

an aggregate base. The base may, however, be integrated not just in the minimal sense represented by having common fillers for slots in different frames, but in the more thorough sense represented by the explicit definition of frame relationships, as in a hierarchy. It is easy to see that a natural progression from here to full integration would occur when all reference to the particular sources of whole frames or of individual fillers was abandoned. At this point the interest of NLP or AI techniques for document processing is just that of knowledge base derivation, on the assumption that the knowledge base is appropriate for information retrieval which is now interpreted in a rather different way and in turn leads to fact retrieval and full-blown AI.

It is important to recognise explicitly that this step is being taken, and that it is assumed that source documents are of no interest in their own right, e.g. for their expressive properties or character as individual wholes (Sparck Jones 1991). It is possible to combine having a knowledge base with access to backup documents, but this is difficult to manage - i.e. what points to what, and like the full abandonment of the sources, has to be justified by particular information needs. Thus when proposals are made to apply NLP or AI methods to produce text representations or replacements, a proper case has to be made that the specific retrieval needs to be met really require this. It has to be shown, that is, that these needs are not of the usual generic topic kind that indexing in the ordinary sense is designed to meet. Indexing of this sort for document retrieval has developed because long experience has been taken to show that, given the many sorts of imprecision involved in retrieval combined with the fundamental lack of information that retrieval presupposes, descriptive refinement is unnecessary and what is rather needed is proper support for the user in searching. This imprecision stems, in document retrieval in the ordinary sense, from the multi-facetted nature of any topic, the analogous property of ordinary language, and the indirection of access; it has to be counterbalanced by redundancy in indexing and searching, not by pared-to-the-bone accuracy, especially as allowance has also to be made for the imprecision of the user's need.

This is not to imply that retrieval from information or knowledge bases does not allow for non-specific or partial queries. It is rather that if the form the base takes is independently justifiable on good grounds, as it is in the similar case of conventional databases, it may imply correspondingly different forms of interrogation. In general, if the assumption behind having a knowledge base is that the base can directly provide answers to questions in the shape of facts then, as with conventional databases, the inquiry situation is functionally different from the document and text retrieval case we are concerned with here, where the user's constructive interpretation of the retrieval materials is essential and central. It is, however, also possible to envisage information and text bases being used directly for searching with imprecise needs provided, as mentioned earlier, the user fully understands the form of knowledge representation used.

9 Evaluation problems

The current opportunity is that there are new contexts for information retrieval in the broad sense; and these are interesting because they may justify new approaches to information extraction and representation. But with new approaches the concomitant challenge is to devise and conduct appropriate system evaluations.

The root problem here is dealing with interaction. As mentioned earlier, those working in document retrieval over the last thirty years have painfully acquired a set of techniques

for evaluating retrieval system performance which are far in advance, methodologically, of anything normally used in NLP apart from machine translation, at least until the recent Message Understanding Conferences (Lehnert and Sundheim 1991) and similar projects (and the same holds for much of AI, cf Cohen 1991). These techniques were, however, originally developed for offline searching, and though they are still used (for example in SMART- related work: cf Salton and McGill 1983, Salton and Buckley 1989) and are useful, evaluation methods and standards need developing for online and interactive searching. Evaluation methods, especially for performance evaluation in operational contexts, 1are also specifically needed for retrieval from non-text information or knowledge bases; but while this is a tough problem in itself, the real challenge, as in the document and text retrieval case, is in evaluating interactive search performance (Robertson and Hancock-Beaulieu, in press).

The essential point here is that the user is not responding passively to system output, but is revising his search specification in response. This may, and usually will, imply a redefinition of his information need, which has two consequences, one for the individual search, the other for testing in general.

With the individual search, the problem is that as the definition of the need may have changed, it is very difficult, at the end of searching, to evaluate performance for what has been retrieved in relation to what ought to have been retrieved. But while precision (the ratio of relevant retrieved to non-relevant retrieved) may be captured only from what has been retrieved (though even in this case this may involve a somewhat misleading aggregation over the whole search), it is also often important to evaluate performance for an indexing or searching method in relation to what was not retrieved.

The other problem is that whenever comparisons between methods are called for, the individual user has been corrupted by his past experience and so cannot be invited to search for the same need using different methods. That is to say, the user has been corrupted by the relevance assessments he has already made. In older-style investigations, searching was separated from assessment. This corruption problem implies much larger samples of searches to establish system performance properly.

Information retrieval systems, however intelligently adaptive to the individual user they are supposed to be, are essentially driven by averages: indexing or searching devices are adopted because they have generally worked satisfactorily, over many searches, in the past, and can therefore be predicted to perform correspondingly in the future. In essence this also applies to systems offering tailoring to the individual. The prime requirement of retrieval system evaluation is thus to obtain reliable average performance data (whether for different users or the same user at different times), using performance criteria and measures appropriate to the essential nature of the retrieval task.

9.1 Evaluation techniques for novel systems

Performance criteria and measures thus need much more investigation in their own right, as a necessary preliminary to assertions of the value of novel approaches to retrieval. It is at the same time necessary to be careful about a particular point in connection with novel systems. With novel systems, the 'feelgood' factor is important: do people like using them? Asking people whether they do is perfectly legitimate, but the question must be clearly recognised for what it is and not misunderstood as an objective measure of success in retrieving relevant material, any more than saying food tastes good means it is nutritionally adequate.

Then with any novel NLP-based scenarios in the document and text retrieval case, it

is necessary to develop monitoring and measurement techniques for interactive information management, perhaps using the experience being gained with OPACs. Though there have been studies of user search behaviour (Keen 1977, Mischa and Lee 1987), and of notions of relevance as well as of e.g. how their readers use scientific papers (Hewins 1990), there has not been enough investigation of how users interact in an online computational context with end documents. This also applies where abstracts are effectively treated as if they were end documents. It is also necessary, where retrieval is from information or knowledge bases rather than text ones, but where the user's needs are imprecise, to establish the appropriate fundamental concepts analogous to relevance for document retrieval, or rather to give relevance an appropriate interpretation. For instance, if the user is interested in browsing through a frame knowledge base, to see what it can tell him, what exactly is his need and how therefore can success in meeting it be established? Finally, it is necessary to develop appropriate evaluation criteria and methodologies for the multi- purpose or 'hybrid' information environments, combining many diferent types of resource, that are now being developed. Where the user switches not only from one resource to another but from one type of task to another, according to current contextual requirements, how are either the global system's performance, or that of its individual components, to be measured? Some first beginning have been made for the elements of such systems (Croft et al 1990), but much more needs to be done.

But if it is essential to develop appropriate detailed evaluation methods to take account of the new working environment which combines modern interactive and display resources with novel, text-motivated techniques for representing and seeking information, it is also necessary to bear in mind what modern technology offers existing modes of indexing and searching. Modern technology is not the working environment just for novel NLP or AI-based approaches to information retrieval. It is also the context in which the strategies developed in earlier retrieval research are being applied (Harman and Candela 1991, Sanderson and van Rijsbergen 1991, Stein 1991). This may make these comparatively established technologies more effective from the point of view both of formal performance measures and of informal user satisfaction. Thus the advantages that modern technology, say for screen displays, could give to these to these older approaches could lead to higher performance levels for them which would raise the competitive stake for the newer alternative, and putatively superior, approaches.

10 Conclusion

My first conclusion is thus that it is not clear that modern analytic, rather than statistical, NLP techniques can of themselves make a large contribution to 'mainstream' document indexing and retrieval. They should certainly be tried for this, but better motivated in relation to exactly how they differ from conventional indexing and searching, as means or for ends, than they often are. They need, in particular, to be more fully considered from the point of view of request rather than document properties; and they need to be studied from the point of view of 1scale effects, not on processing, but on discrimination. One of the disconcerting findings of the past has been that quite different forms of indexing or retrieval have much the same effect in the little and the large. Thus it is necessary not merely to show difference of method but difference of outcome.

My second conclusion, however, is that even for the 'mainstream' case (and taking this as more homogeneous than it is), novel NLP techniques should be tried when they are to be

applied within the framework of multi-level processing, for example with coarse-grained and then fine-grained matching adopted as an intellectually rather than economically motivated search strategy. Though hybrid strategies are used in conventional systems, the particular forms which NLP would allow the system (rather than the user) to apply have not been a practical option in established systems.

My third conclusion is that modern NLP techniques call for trial within the working environment offered by current interface technology, where many different types of information object and information management operation can be conveniently combined. This will not be easy, as any attempt to automate the production of hypertext links suggests, and it may also not be easy to establish that any particular device, like parsing, is making any noticeable contribution to overall performance. But the opportunities here should certainly be investigated.

Finally, and most importantly, there is every good reason to experiment with substantive NLP and AI methods for information determination and retrieval for special types of application context or in individual, currently non-standard, retrieval environments. This clearly applies to the case where an explicit information or knowledge base wholly or partly replaces source text, but it could clearly also hold in the document case where the nature of the material and user requirements demanded it. The manifest need therefore is to obtain a better idea of what these conditions justifying more than only statistical language processing actually are, and exactly how they should be met. Thus if on the one hand, as Hayes (at the AAAI TBIS Symposium 1990) noted, effective routing may not call for syntactic text analysis, it would seem to be called for when an information request can be properly treated as a direct question for which an answer may be sought in the stored text. The pressing research need is thus to establish what the many data variables, from collection size or typical relevant/nonrelevant ratio to user experience and goal, imply not just for the feasibility but for the potential utility of NLP in new and different, as well as old and familiar, retrieval environments.

References

D. Austin and J.A. Digger, 'PRECIS: the preserved context index system', in *Theory of subject analysis: a sourcebook* (Ed L.M. Chan, P.A. Richmond and E. Svenonius), Littleton, CO: Libraries Unlimited, 1985.

N.J. Belkin and A. Vickery, *Interaction in information systems*, Library and Information Research Report 35, The British Library, London, 1985.

N. Bely. et al, *Procedures d'analyse semantique appliques a la documentation scientifique*, Paris: Gauthier-Villars, 1970.

D.C. Blair, 'Indeterminacy in the subject access to documents', *Information Processing and Management* 22, 1986, 229- 241.

D.C. Blair and M.E. Maron, 'An evaluation of retrieval effectiveness for a full-text document-retrieval system', *Communications of the ACM* 28, 1985, 289-299.

Borgman, C.L. 'Why are online catalogues hard to use? Lessons learned from information-retrieval studies', *Journal of the ASIS* 37, 1986, 387-400.

L.M. Chan, P.A. Richmond and E. Svenonius (Eds), *Theory of subject analysis: a sourcebook*, Littleton, CO: Libraries Unlimited, 1985.

C.W. Cleverdon, 'The Cranfield tests on index language devices', *Aslib Proceedings* 19, 1967, 173-194.

C.W. Cleverdon, *A comparative evaluation of searching by controlled language and controlled language in an experimental NASA database*, Report ESA 1/432, European Space Agency, Frascati, Italy, 1977.

P.R. Cohen, 'A survey of the Eighth National Conference on Artificial Intelligence: pulling together or pulling apart?', *AI Magazine* 12 (1), 1991, 16-41.

W.B. Croft, R. Krovetz and H. Turtle, 'Interactive retrieval of complex documents', *Information Processing and Management* 26, 1990, 593-613.

F. Debili, C. Fluhr, and P. Radasoa, 'About reformulation in full-text IRS', *Information Processing and Management* 25, 1989, 647-657.

G. DeJong, 'An overview of the FRUMP system' in *Strategies for natural language processing* (Ed W.A. Lehnert and M.D. Ringle), Hillsdale, NJ: Lawrence Erlbaum, 1982.

T.E. Doszkocs, 'CITE NLM: natural-language searching in an online catalogue', *Information Technology and Libraries* 2, 1983, 364-380; reprinted in Willett, 1988.

J.L. Fagan, *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods* (PhD thesis), Report 87-868, Department of Computer Science, Cornell University, 1987.

N. Fuhr, 'Models for retrieval with probabilistic indexing', *Information Processing and Management* 25, 1989, 55-72.

M. Hancock-Beaulieu, S. Robertson and C. Neilson, 'Evaluation of online catalogues: eliciting information from the user'. *Information Processing and Management* 27, 1991, 523-532.

D. Harman and G. Candela, 'Bringing natural language retrieval out of the closet', ms. National Institute of Standards and Technology, Gaithersburg MD, 1991.

E.T. Hewins, 'Information need and use studies' in *Annual Review of Information Science and Technology,* Vol 25, (Ed M.E. Williams), Amsterdam: Elsevier, 1990.

L.F. Rau and P.S. Jacobs, 'Creating segmented databases from free text for text retrieval', *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, New York: Association for Computing Machinery, 1991, 337-346.

E.M. Keen, 'The Aberystwyth index languages test', *Journal of Documentation* 29, 1973, 1-35.

E.M. Keen, 'The processing of printed subject index entries during searching', *Journal of Documentation* 33, 1977, 266-276.

E.M. Keen, 'The use of term position devices in ranked output experiments', *Journal of Documentation* 47, 1991, 1-22.

F.W. Lancaster, *Vocabulary control for information retrieval*, Washington DC: Information Resources Press, 1972.

F.W. Lancaster, C. Elliston and T.H. Connell, 'Subject analysis' in *Annual Review of Information Science and Technology*,

Vol 24 (Ed M.E. Williams), Amsterdam: Elsevier, 1989.

D.D. Lewis, W.B. Croft and N. Bhandaru, 'Language-oriented information retrieval', *International Journal of Intelligent Systems* 4, 285-318, 1989.

W.H. Mischa and J. Lee, 'End-user searching of bibliographic databases' in *Annual Review of Information Science and Technology*, Vol 22 (Ed M.E. Williams), Amsterdam: Elsevier, 1987.

A.S. Pollitt, 'CANSEARCH: an expert systems approach to document retrieval', *Information Processing and Management* 23, 119-138, 1987.

M.F. Porter and V. Galpin, 'Relevance feedback in a public access catalogue for a research library - MUSCAT at the Scott Polar Research Institute', *Program* 22, 1988, 1-20; reprinted in Willett, 1988.

C.J. van Rijsbergen, *Information retrieval*, 2nd ed., London: Butterworths, 1979.

S.E. Robertson and M.M. Hancock-Beaulieu, 'On the evaluation of IR systems', *Information Processing and Management*, in press??

N. Sager, 'Natural language information formatting: the automatic conversion of texts to a structured database' in *Advances in Computers*, Vol 17 (Ed M.C. Yovits), New York: Academic Press, 1978.

G. Salton, 'Another look at automatic text retrieval systems', *Communications of the ACM* 19, 1986, 648-656.

G. Salton, 'Developments in automatic text retrieval', *Science* 253, 1991, 974-980.

G. Salton and C. Buckley, 'Improving retrieval performance by relevance feedback', J*ournal of the ASIS*, 1990, 288-297.

G. Salton and M.J. McGill, *Introduction to modern information retrieval*, New York: McGraw-Hill, 1983.

M. Sanderson and C.J. van Rijsbergen, 'NRT (news retrieval tool)', Computing Science Department, University of Glasgow, 1991.

T. Saracevic et al, 'A study of information seeking and retrieving. Part I: Background and methodology. Part II: Users, questions, and effectiveness. Part III: Searchers, searches, and overlap', *Journal of the ASIS* 39, 1988, 161-216.

K. Sparck Jones, 'Search term relevance weighting - some recent results', *Journal of Information Science* 1, 1980, 325- 332.

K. Sparck Jones (Ed), *Information retrieval experiment* London: Butterworths, 1981.

K. Sparck Jones, 'The role of artificial intelligence in information retrieval', *Journal of the ASIS* 42, 1991, 558-565.

K. Sparck Jones and C.A. Webster, *Research on relevance weighting 1976-1979*, British Library R&D Report 5553, Computer Laboratory, University of Cambridge, 1980.

R.M. Stein, 'Browsing through terabytes', *Byte*, May 1991, 157- 164.

A. Vickery et al, "A reference and referral system using expert system techniques", *Journal of Documentation* 43, 1-23, 1987.

P. Willett (Ed), *Document Retrieval Systems*, London: Taylor Graham, 1988.

1 S.R. Young and P.J. Hayes, "Automatic classification of banking telexes", *Proceedings of the Second Conference on Artificial Intelligence Applications*, IEEE Computer Society, 402-408, 1985.

General Reference

Annual Review of Information Science and Technology, Vols 1 - 25 1966 - 1990; various editors and publishers: Vol 25 Ed M.E. Williams, Elsevier, Amsterdam.