Natural language processing
for
information retrieval

David D. Lewis
AT&T Bell Laboratories

Karen Sparck Jones
Computer Laboratory, University of Cambridge

July 1993

1 Abstract

The paper summarizes the essential properties of document retrieval and reviews both conventional practice and research findings, the latter suggesting that simple statistical techniques can be effective. It then considers the new opportunities and challenges presented by the ability to search full text directly (rather than e.g. titles and abstracts), and suggests appropriate approaches to doing this, with a focus on the role of natural language processing. The paper also comments on possible connections with data and knowledge retrieval, and concludes by emphasizing the importance of rigorous performance testing.

This paper will appear in *Communications of the ACM.*

2 Introduction

Automatic text, or document, retrieval has recently become a topic of interest for those working in natural language processing (NLP). The aim of this article is to indicate the key properties of document retrieval, distinguishing it from both data retrieval and question answering; to summarize past experience in the field; to review the external developments that are stimulating new interest in text and document retrieval; and to consider specific strategies for NLP research aimed at this form of information processing.

For the purposes of this paper we will generally treat the older term *document retrieval* and the newer term *text retrieval* as synonymous. Both are aimed at retrieving texts for humans to read, of anything between paragraph and book length. In the past, document retrieval was in practice concerned with pointing the reader to an offline document, typically a journal article or report. Technological advances have now made it practical to store, search and retrieve all or part of the full document text online. However, the essential requirements of the two are the same, and we will therefore use document retrieval (DR) as the general term, except where the distinction just made is relevant and it is necessary to refer to text retrieval (TR) as supplying the user directly with 'end' text. We will however use *information retrieval* (IR), sometimes taken to mean document retrieval only, as a global term covering everything from data retrieval to knowledge retrieval. We will throughout concentrate on IR, and more specifically DR, as an NLP task.

3 Document retrieval

Within the whole area of IR, DR is a proper and important task with its own distinctive properties, not to be confused with either data or knowledge retrieval.

DR is for the user who wants to find out about something by reading about it; that is where the user is generally ignorant, as opposed to wanting a specific data item or question answered. For example, take a user who wants to read about

`cheap production methods for simple prefabricated housing`.

This does not imply the user has any specific questions in mind, e.g.

`What are cheap production methods ...`

or

`How do cheap and expensive methods ... differ?`.

Moreover, even if the user has some questions in mind, the aim is to get overall information such that not only these questions but others that reading the documents themselves suggest can also be answered. This means that DR must find relationships between the information needs of users and the information held within documents, both considered in a very general sense, and neither directly available to the computing system.

Further, and equally importantly, the relation between the user's *need* and what meets it is not necessarily obvious. For instance our example need may be met by

`J. Kirk: Reed mat huts of Madagascar: design and construction`.

Retrieval thus depends on *indexing*, i.e. on some means of indicating what documents are about. Indexing in turn requires an *indexing language* with a *term* vocabulary and a method for constructing request and document *descriptions*. Indexing is the base for retrieving documents that are *relevant* to the user's need. It has to be supported by a search apparatus that specifies conditions for a match between request and document descriptions, and modulation methods to alter these descriptions if no match is initially forthcoming.

The fundamental aim of indexing is to increase *precision*, i.e. the proportion of retrieved documents which are relevant, and *recall*, the proportion of relevant documents which are retrieved. It has to achieve these in the face of two kinds of problems.

First, there are problems posed by the external context within which searching is done, for instance that there are typically few relevant documents and many non-relevant ones. Second, there are problems imposed by the internal constraints of the task itself, which are responsible for the characteristic uncertainty that the retrieval system has to overcome. The first constraint is the *variability* in ways that a concept may be expressed [11]. This is partly a matter of language, e.g.

`prefabricated` *vs* `unit construction`,

where the underlying notion of prefabrication is the same, and partly one of perspective, e.g.

`prefabricated` *vs* `factory made`,

where there are different views of how prefabrication is done.

The second constraint is request *underspecification*, whether because the request is vague, e.g.

`cheap` *as* `economical production` *vs* `cheap` *as* `low quality`,

or because it is incomplete, e.g.

`housing` *vs* `temporary housing`,

the difference between these two being that in the first case the user may not realize the ambiguity and in the second has failed to give sufficient detail. This is a less obvious, but nevertheless characteristic DR problem: it follows from the user's ignorance before reading.

The third constraint is the *reduction* of documents in their descriptions, so descriptions are indirect, e.g.

`building` *for* `reed mat hut`

or partial, e.g.

`construction` *for* `design and construction`,

losing information in different ways. Since full texts of documents are increasingly available online, the degree of reduction is more under the control of the indexing method than in the past. But reduction can never be completely avoided—the author of a document always leaves much unsaid on a subject—nor is it always pernicious. Forming compact descriptions of the significant content of documents may increase both the efficiency of matching and its effectiveness in classifying textual material as relevant or nonrelevant, just as feature selection is critical in other classification tasks.

DR thus imposes conflicting demands on text descriptions, asking that they be both normalizing and accurate, both discriminating and summarizing. The result is that variations in indexing that raise precision more often than not lower recall, and vice versa. Beating this tradeoff and raising both recall and precision is the fundamental goal in constructing an index language.

There are many possibilities for indexing languages [19]. Terms may be any that appear in the text to be indexed (*natural* language), or may be limited to those from an artificial or *controlled* language, the design of which involves many of the same concerns as in treating meaning representation for NLP.[1] Languages vary in the form of, and emphasis placed on, terms and term relations; implicit and explicit relations; and *syntagmatic* (document or request specific) and *paradigmatic* (universally asserted) relations. Natural languages are perhaps the most widely used, but hybrids are common, such as natural terms combined with artificial relations, e.g. (with natural language elements in lower case and controlled language ones in upper case):

```
(hut MATERIAL (mat MATERIAL reed)) LOCATION madagascar
```

or

```
(reed mat hut) OF (madagascar)
```

as are wholly controlled forms e.g.

```
(UNIT CONSTRUCTION HOUSING)(MADAGASCAR).
```

4 Past research

Tests of a wide range of indexing languages over the last three decades have shown fairly consistent (if not wholly expected) results ([22], ch. 3; [29], pt. 3; [34]). These tests have shown that indexing of documents by individual terms corresponding to words or word stems produces results as good as those obtained when indexing by controlled vocabularies, whether simple or complex, and whether produced by manual effort or automatic language processing. Further, automatically combining single indexing terms into multi-word indexing phrases or more complex structures has yielded only small and inconsistent improvements over the simple use of multiple terms in a query.

In contrast, statistical DR methods, which ease and enhance the use of representations based on single terms, have provided significant improvements over alternative approaches, such as boolean querying [23]. Statistical DR methods rank documents based on their similarity to the query, or on an estimate of their probability of relevance to the query, where both query and document are treated as collections of

[1] We will use 'natural language' in this sense of drawing indexing terms from the document itself, and use 'NLP' when referring to natural language processing.

numerically weighted terms. The query can be an arbitrary textual statement of the user's information need, or can even be a sample document.

Statistical DR methods assign higher numeric weights to terms which show evidence of being good content indicators, thus causing them to have more impact on the ranking of documents. The number of occurrences of a term in a document, in the query, and in the set of documents as a whole may all be taken into account in computing the influence the term should have on a document's score. In addition, if the user indicates that certain retrieved documents are relevant, this information can be used to reweight and alter the set of query terms, in a process called *relevance feedback* [24, 28].

The focus in this baseline statistical DR strategy is on tuning the representation to the current user request, rather than on anticipating user requests in the document descriptions. The strategy has three major benefits. First, it allows for *late binding.* Complex concepts need not be anticipated during indexing, but are under the control of each user at query time. Second, *redundancy* is supported by drawing indexing terms from the document text, rather than using a limited vocabulary which may not support a particular user's needs. Finally, the representation is *derived* from the documents themselves, so that differences and similarities among the document texts are given the best chance to survive into the document representations.

Consider an example query presented to a statistical DR system:

```
A cheap method for prefabricated housing.
    25    5               30       20
```

The term weights shown, `25`, `5`, `30` and `20`, would be assigned automatically to the highlighted stems based on their statistics of occurrence in the set of documents. A document matching the query on the stems `cheap` and `prefabricat` would score highly. If the user indicated to the system that this document was relevant, then relevance feedback would increase the weights on `cheap` and `prefabricat`. In addition, highly weighted terms from the relevant document, say `unit` and `construct`, might be added to the request, with their own weights. They could then promote a hitherto uninspected document through a joint match on, say, `prefabric` and `unit`.

The research results showing the effectiveness of statistical DR methods are solid in that many tests have been done in different environments, for instance of subject domain; under ranges of system parameters, say for weighting; and using alternative evaluation procedures with, for example, distinct performance measures. The methods also generally apply to document routing, against standing rather than one-off needs and perhaps for coarser document categories. However, these studies have used small numbers of documents (at most 30,000, and usually much fewer) compared to operational DR systems, have mostly neglected non-European languages, and have been *surrogate*-based, i.e. using titles and abstracts, which are distillations of full document content with a high loading for what is especially important in the source. The approach also depends on users entering requests that are sensible topic specifications and provide several terms for alternative matches.

In addition to these caveats to the success of statistical DR, the question also remains of why intuitively plausible improvements in document representation have had so little impact on effectiveness. Why is there no gain from linguistic sophistication, e.g. from the use of syntactic role relations between terms? Is it that NLP intended to produce sophisticated indexing has been inadequately done? Is it that our transformations of natural language, even when done well by humans, have been misdirected? Or is it that so much leverage was gotten by searching surrogates in previous experiments that little room for improvement was left. Still, with typical effectiveness results in the range of 30 to 60% recall or precision [23], there is considerable room for improvement. Further, the research results just described must be considered in the context of current operational practice and of the new TR situation where full source texts, and not just their surrogates, are available for direct searching.

5 The current state of DR

There are thousands of bibliographic databases now accessible, mainly in surrogate form, through a variety of services. The longstanding debate on the merits of controlled vs. natural language indexing has become less important, since many commercial databases now use both. Most searches of these databases are conducted for end users by professional intermediaries who know about database coverage as well as about the controlled language and indexing practice with it. These intermediaries generally believe a controlled language is superior to natural language, though the controlled languages used illustrate many different design options, with no clear winners [19].

The searching of well-cared-for bibliographic databases is no longer all DR must deal with, however. A DR session today may involve a personal computer user scanning their hard disk for a missing file or a student searching thousands of Internet servers for an archived Usenet posting. End-user, natural language searching becomes inevitable, because there are neither opportunities nor resources to use intermediaries and indexers, so when full text is available it seems natural to search it directly.

The fruits of IR research have been brought to bear against this flood of both traditional and nontraditional data, with some success [26]. Statistical text retrieval systems of the sort suggested by DR research now span the range from personal computers to 100-gigabyte service databases [21]. Still, the situation is far from satisfactory, with at least three classes of problems.

First, the penetration of the best methods into operational practice is uneven. Many systems still require Boolean logic or other user-befuddling query syntax. When natural language querying is available, weighting may be unavailable or poorly chosen, and relevance feedback is rarely supported. Word stemming operations may also be unsatisfactory or ill-understood.

Secondly, there is much that is unknown about the proper application of statistical DR methods to large, heterogeneous databases, particularly of full-text documents. Test collections of this sort have only very recently become available and experiments with them, while verifying a reasonable level of efficacy for standard techniques, have revealed many surprises and problems [13].

Thirdly and most important, many end users have little skill or experience in formulating initial search requests, or in modifying their requests after observing failures. Even when relevance feedback is available, it still needs to be leveraged from a sensible starting point [5].

Thus, while established research results show that natural language indexing and searching is effective to a degree, it is natural to ask first, whether it is possible to improve on the very simple strategies described earlier without increasing the load on the user, and second, whether it is necessary to look for more sophisticated approaches to handle full text, where the conceptual detail is much greater. Thus more discriminating methods may be needed to separate the sheep from the goats in large files of full texts, as well as desired because with full text more focusing on particular content is possible.

There are therefore two issues. One is whether natural language indexing, perhaps of a more refined kind than statistically controlled use of single word terms, is wanted, or whether controlled language indexing is really what is needed. Both controlled language indexing and more sophisticated natural language indexing imply non-trivial NLP, so the other issue is whether the required NLP capabilities are available or in prospect, since large-scale human full-text processing is not a practical proposition.

These issues will be addressed in the context of NLP research which is itself in an exciting and rapidly changing state. An increase of interest in robust processing, in processing large amounts of real-world text, and in statistical methods in NLP make this an opportune time to consider interactions between DR, and more specifically TR, and NLP.

6 A TR research agenda

All the evidence suggests that for end-user searching, the indexing language should be natural language rather than controlled language oriented. Indexing, i.e. selective text content characterization, is needed, but it should be derived from the text, with redundancy and late binding to compensate for uncertainty. The indexing language should moreover, for interactive searching, be directly accessible to the user for request formulation: the user should not be required to express their needs in a heavily controlled and highly artificial language. This does not, however, mean that the system cannot enhance the user's indication of what they want, for instance with statistical data or concept definitions they may not be able to interpret in detail.

There is also some evidence to the effect that combining single terms into *compound terms*, representing relatively fixed complex concepts, may be useful. Many controlled languages allow this, and it has been found effective to a degree when done 'statistically', i.e. on a simple co-location basis within a text window [9]. While linguistically grounded compounds have not been found more effective that statistical ones in past studies, this may change in a TR context.

The proposals which follow develop these themes, as an approach that might give better results than the simple baseline described earlier. They address first the 'words', 'phrases' and 'sentences' that form individual document descriptions and express the combinatory, syntagmatic relations between single terms that are captured by the system's NLP-based text-processing apparatus; second the classificatory structure over the document file as a whole that indicates the paradigmatic relations between terms which allow controlled term substitution in NLP-based indexing and searching; and third the system's NLP-based mechanisms for searching and matching.

6.1 Indexing descriptions

What should the linguistic units of indexing descriptions be like? That is, what should the size and depth of text forms sought, and of representation forms delivered, be? For example, should one go for any words, or only for nominal group heads; for concatenated or case-labelled phrases? Our proposal is for well-founded simplicity both for the natural language units taken from the text as inputs to the indexing process, and for the natural language or near-natural language units in the indexing language descriptions output by the indexing process. So as units, taken as or made up from elementary terms, one would use linguistically solid compounds e.g.

`prefabricated housing`

or basic propositions e.g.

`produce(factory, house).`

The success of such a proposal is in the details, and several we consider crucial differ from what might be assumed from traditional NLP practice. First, given the proven value of statistical weighting, any units that NLP produces should be filtered and weighted by the statistics of their occurrences in the database searched and perhaps in other textbases as well [8, 18]. The evidence is that with compounds this must be done particularly carefully [9].

Secondly, we have stressed the importance of late binding and sensitivity to the uncertainty of evidence. Compound terms will not be identified as definitely occurring or not occurring in a document. Rather, each document will provide some amount of evidence for the presence of each known concept. An occurrence of the syntactically-checked noun phrase `prefabricated units` in a document would provide very good evidence for the corresponding concept's presence. An occurrence of the verb phrase `(they) prefabricated units` in a document would provide only slightly less evidence for the noun phrase concept. The occurrence of the two words in separate paragraphs would provide much less evidence, but more than the amount given by the presence of just one of the words, or of a related word [6, 30].

Thirdly, basic compound units of the type described above would not typically be further combined into frames, templates, or other structured units. (Though there might be exceptions, as discussed in Section 9.) The description of a document would be an unordered set of 'phrase' units and individual words. This applies whether compound terms are formed at document file time or introduced by requests at search time. The rationale here is that more complex structures are labor-intensive to design, difficult to fill accurately, and that matches on even basic propositions are so unusual that finer-grained distinctions are unlikely to be provide additional information.

Simply applying the appropriate NL procedure to extract all instances of compound terms should produce a reasonable representation for moderate sized documents. For very large full text documents, further reduction may be needed to get a reasonable summary representation of content which is not swamped by the idiosyncrasies of large numbers of subparts. One could restrict terms to being drawn from particular portions of the text or, better, take into account both the global and local structure of the document in matching [25]. In either case, statistical control in unit choice and weighting is again required. Only experiment can show what forms of reduction are useful and not too costly.

Thus for processing individual texts what is proposed is more refined compound terms representing complex concepts, but with loose coupling between these to allow for flexible matching. Many experiments are needed on the precise form of these compound terms and on how they should be selected and weighted, for instance relative to their constituents, where the issues are clearly more complex than for single terms.

6.2 Indexing and searching resources

If recall as well as precision is to be increased, some mechanism which allows non-identical terms to match is required. The traditional approach to this is through *normalization*, i.e. replacing several forms by a single canonical form. Stemming is a normalization based on morphology, e.g.

`prefabricated`, `prefabrication` → `prefabricat`.

Semantic normalizations are possible as well, both ones based on manually-defined classes,

`house`, `apartment`, `hut` → `DWELLING`

and ones based on, say, automatically detected but hitherto unrecognized, statistical associations in a document file,

`house`, `lawn`, `gasoline` → `CLUSTER-1738`.

However, any normalization applicable to indexing can also be used more flexibly during matching. Retaining original document descriptions has important advantages—notably fidelity—and relational knowledge can be invoked in a context-sensitive and adaptive way during searching. Relationships can be adjusted to suit the individual query either directly (say via user browsing in an graphical display of associations), or indirectly (say by inference from the user's relevance judgements). This strategy also avoids costly reindexing of the entire document file when alterations or additions are made to the system's paradigmatic knowledge.

In Section 7 we say more about the kinds of paradigmatic information that NLP might provide. Under a model where term relationships are suggestive of, rather than demanding of, normalization, a wide variety of resources indicating term relationships or term classes can be used. Non-obvious candidates include any form of symbolic data (computer programs, knowledge bases, expert system rulebases, data dictionaries) since symbols are almost always given names which are natural language words or compounds. These resources may be more useful in their particular domain than general purpose lexical resources would be.

6.3 Searching procedures

Finally, for searching, what should the mechanisms used to set matching conditions and determine request modification be? For example, should matching be loose or tight, and modulation free or constrained? It again appears that natural simplicity is right, allowing straightforward element stripping or substitution in compound terms, e.g. replacing

```
cheap prefabricated housing
```

by

```
prefabricated building;
```

and permitting obvious relational relaxation or substitution, e.g. trying

```
cause (building)
```

instead of

```
produce (factory, house).
```

The assumption again is that statistics will be applied as a further guide or control, in iterative searching, through selection and weighting. Explicit probabilistic models may be favored over alternative matching schemes for their ability to combine a wide variety of evidence [33], but admittedly all current models find it difficult to deal appropriately with complex descriptions and their elements.

6.4 Comments

Thus drawing on the DR lessons of the past, we propose that the general flavor of future TR, as well as DR, systems is of simple flexible natural language indexing forms, support devices, and use strategies. These allow and encourage the user to concentrate on request development which, as opposed to document characterization, is what really matters; and they do it in a way that supports derivation, redundancy, and late binding. This approach is, moreover, potentially both economically viable even for large volumes of material, and practical from the user's point of view given modern interface technology like windows. It is also appropriate for two particular full text cases. One is retrieving subtexts, say paragraphs: even for short, focused pieces of text, it is still necessary to index on significant concepts. The other is two-level retrieval, first coarse then fine, whether because this allows motivated zooming or is operationally convenient.

Many indexing strategies can in principle be applied either at document file or request search time, with issues of space, speed, or portability dictating a particular choice. NLP might be completely restricted to queries, with evidence that the resulting compound terms apply to a document determined solely by testing for proximity of words [30]. Besides the obvious efficiency advantages of avoiding NL analysis of the document file, an interface using NLP query analysis could be installed on top of the existing file access structure of many conventional boolean query systems. Another tradeoff point between efficiency and preciseness in matching would be to apply NLP only to documents scoring high on a word-based query. Even when NLP is applied to the whole document file, there are tradeoffs between explicit indexing on compound terms (speeding up querying but increasing the size of access structures), and indexing only on their components or generalizations of their components (e.g. stems). In other cases both efficiency and effectiveness may dictate the same course, as in the use of reduction in indexing. Careful design of the system as a whole is required to optimize the many factors involved, given their interdependencies.

For end users, natural language indexing strongly related to actual texts is attractive, and while they are required to participate in search development, fast processing and multi-window displays make it easier for them to exploit available information sources. There are, however, challenges in ensuring that any user does understand what is happening and both can and does, for instance, exploit a store of paradigmatic knowledge. It may be difficult to convey the significance of statistical data; and while artificial description forms like predicate-argument structures can be applied in TR in a way that is hidden from the user, to avoid repelling their users or being misunderstood, it is necessary to motivate retrieval output for the user and hence to link the indexing descriptions actually used to comprehensible surface manifestations.

7 NLP implications

From the NLP point of view, the clear challenges are, first, the generic one of whether the necessary NLP can be done; and second the more specific ones both of whether non-statistical and statistical data can be appropriately combined, and of whether data about individual documents and whole files can be helpfully combined, since it is always necessary to treat a document in its file context.

The demands imposed on NLP by the above program differ from those in most NLP tasks. TR, even more than DR, is tolerant with respect to errors in document representations. In addition, probabilistic indexing [10] allows the NLP system to leave some ambiguities unresolved in its output. On the other hand, NLP applied to documents must cope with vast amounts of variable quality text from broad domains. User requests present smaller amounts of text, but even more variability in form and content. Each of the three main aspects of our strategy—forming text descriptions, providing and exploiting terminological resources, and ensuring matching in searching—poses different challenges for NLP, as we examine in this section.

We left open what syntagmatic relationships between terms in text would suffice for those terms to form a compound term. Strategies for traditional, if partial, syntactic analysis allowing processing of hundreds of megabytes of text have been tested for TR [32], but traditional semantic analysis on a large scale remains to be demonstrated. New approaches are also possible. Accurate and highly efficient syntactic taggers are available, and some compound terms, for instance head nouns and premodifiers, are easily extractable from tagged text [3]. A variety of strategies for finding important collocations in large corpora have been developed [27], and may provide an improvement over traditional IR methods for statistical phrase formation. Compound terms must not only be generated, but also selected among and weighted. Methods for exploiting the discourse structure of large texts may be useful in identifying which terms are central to the content of a text.

Another role for NLP is in automated and semi-automated acquisition of paradigmatic knowledge. Automated formation of clusters of related words is again attracting attention, despite the historical lack of success of this technique in DR. More linguistically motivated approaches, such as clustering based on syntactic context, may prove an improvement on traditional strategies [14, 15]. Leveraging of hand-coded resources, such as inducing semantic information from labelled training data or from machine-readable dictionaries, may be a more effective, if less general, approach.

Finally, the type of NLP that is done constrains what forms of matching are possible. For instance, element stripping might be restricted to just adverbs, or to words which do not appear in a domain-dependent vocabulary, but these restrictions can be implemented only if NLP has marked compound term elements with the necessary information. NLP need not be applied identically to queries and documents: thus in particular, one might do a very careful extraction of compound terms from the request, but use a quick and dirty approach to find compound terms in the vastly larger amount of document text. The resulting uncertainty in the document representation may be taken into account in the matching process. NLP applied to the user request might also focus on distinguishing between request words that should be matched against documents and those that convey other information about user needs (e.g. *Please retrieve journal articles published after 1987 about...*).

A general caution is needed about the prospects that simple NLP strategies will significantly improve TR effectiveness. Recent work in NLP has made heavy use of the context of a word as a clue toward its meaning. Methods very similar to request/document matching in IR have been used, for instance, for word sense disambiguation. It is not surprising, then, that when a document and request match on several words, it is likely that individual matching words have the same word sense [17]. The matching process itself has provided a kind of disambiguation. As another example, words tend to be accompanied by paradigmatically related words in documents, and relevance feedback may add these words to the request, much as a paradigmatic knowledge base would.

Thus, NLP techniques are faced with the challenge that the basic methods of statistical IR have picked some of the easy fruit off the tree. The result is that, to date, choices among alternate statistical retrieval methods have had much more impact than choices among alternate text representations [1]. This should not discourage research into NLP applications in IR, but does suggest careful examination of where NLP is likely to have the most impact.

8 Data retrieval

Within IR we distinguish DR from other forms of retrieval. We can only comment briefly on these forms here, concentrating on their relationship to DR (and TR) first to see what carry over there may be from one to another and how they may be combined, and second to see how NLP experience may be transferred.

We define *data retrieval* as the case where file information is precoded for specific properties and where the conceptual categories for queries have to be known in advance.

Natural language access to databases, replacing the use of formal query languages, has been investigated for three decades and there are well-established commercial systems [4, 7]. Natural language clearly offers advantages in convenience and flexibility, but there are correspondingly major challenges in query interpretation, precisely because query expression is decoupled from search formulation. Input queries can require extensive transformation to map onto file categories and this may have to be mediated by a rich and extensive domain model; and there are particular problems in dealing with 'ill-formed' input. Thus natural language front ends can be effective, but normally only after significant customization effort.

The specific difference between DR and data retrieval is that in data retrieval the set structure for the query is critical and has to be specified precisely. The quantificational structure of the input has therefore to be identified in natural language analysis. The user may in fact have a vague query; but it has still to be interpreted in one or more definite ways for searching, much like a DR Boolean query: post hoc set specification as in DR coordinate searching is not allowed.

It is not clear that data retrieval experience is directly applicable to DR, as there is a fundamental difference in the nature of the information base and type of need, though work on developing natural language analyzers capable of resolving predication structures for data query is relevant to compound term identification. DR techniques might on the other hand be applied in data retrieval to provide 'relaxed' queries automatically if initial ones do not provide an answer. They might also be used to generate substitute or 'partner' queries for searching accompanying text files [30, 35]. Finally, it is possible that DR and TR techniques of the kind we have described may be appropriate for databases with free text field values, and even more for what may be called *record bases*, as illustrated by e.g. museum catalogues, where there are often several free fields as well as coded or controlled ones and free fields may extend to paragraph-length text.

9 Knowledge retrieval

The relationship between DR and *knowledge retrieval* (or 'question-answering') is potentially more interesting, where we define knowledge retrieval as direct, like data retrieval, but as not depending on such rigorous precoding and thus requiring more powerful inference capabilities than either data or document retrieval.

It is sometimes supposed that replacing a document file by the knowledge base it embodies would obviate the need for DR, while allowing better IR. This could be useful in some contexts, and has been done, though with high start-up effort, for some very limited types of texts, for instance banking telexes [36]. But it is still desirable to be able to get at the writer's own presentation of information, which is one aspect of document content. Presentation becomes more important with longer text, and complete replacement by a knowledge base version is also much less feasible here.

Thus a more potentially useful strategy would be to provide DR with more depth and integration through an organized superstructure over the file, which would be exploited as a knowledge base in initial searching. Document frames or templates, for instance, supported by inference capabilities of the kind developed in AI would give detailed, consistent, linkable document characterizations which would allow regulated matching, coherent modulation, and focused browsing. Going further and using a propositional knowledge base would give a unified, high-level collection model which would allow more intensive inference. Many conventional approaches to DR, for example using facetted indexing, and also hypertext [20], can be seen as gestures in this direction: the putative difference would be in the explicit and thorough provision of automatic inference.

EP-X, for instance [20], can be seen as an advance along these lines, but it is based on a controlled language and its knowledge base is still manually constructed. It is clearly very hard to build such a base automatically from documents in a way which maximizes the derivation of information from the documents themselves, successfully selects what is important in documents, and manages backup from base to individual documents correctly. Some beginnings have been made in this area [16, 12], but chiefly in limited domains, and by taxing current NLP capabilities to their limits. Processing itself is also knowledge-heavy so for wider and larger files bootstrapping the lexicon, for instance, is needed.

Thus though the function of the knowledge base is to encourage query development, and this could include question-answering on the base itself, the conditions as well as practicalities of DR suggest that the right approach to knowledge base design is to try for a simple structure embedding natural language, with rich text pointers, e.g.

```
BUILDING
  ( TYPE      : hut          ---> text 1
    UNIT      : mat          ---> text 1
    MATERIAL  : reed         ---> text 1
    USE          -
    PLACE     : madagascar   ---> text 1
    ...)
...
```

A structure like this would be hospitable and not too constraining. A good case can be made for the use of the same type of structure as a means of linking different bases and types of base within global systems: different bases within such hybrid systems would all be treated as if they were document (i.e. text) collections and tied together, to support 'travels in information space', through associative lexical indexing [31].

10 Conclusion

We have seen that while conventional DR services continue to make heavy use of strongly-controlled indexing languages (like the National Library of Medicine's *Medical Subject Headings*), increasing use is being made of indexing where terms are drawn from the natural language of documents. These simple natural language indexing techniques have been shown to be adequate in a wide range of experiments, though not on a really large scale. These techniques are also beginning to be used for TR.

However the greater information detail embodied in full text appears to call for more sophisticated NLP-based approaches to indexing and retrieval. We have suggested that appropriate strategies for this new situation are those building on the simple DR methods, but extending these to allow for well-motivated compound terms and similar descriptive units. The required NLP technology is now being established, and work on applying it to TR is beginning. However there are major challenges first in making this technology operate efficiently and effectively on the necessary scale, and second in conducting the evaluation tests that are essential to discover whether the whole approach, and what specific form of it, works [29], especially when these tests must be for interactive searching and with large files. Thus it is in particular necessary to show whether NLP-derived compound terms are significantly, and usefully, better than e.g. simple collocational compounds.

From this point of view, the present surge of activity in TR, stimulated by the ARPA-sponsored Text Retrieval Conferences (TREC) [13], is to be welcomed. This is a major evaluation study, with much more data than previous experiments and comparing many different strategies, with and without NLP. It is far too soon to draw conclusions on relative merits, especially since tailoring to the particular retrieval application must be discounted. The retrieval needs in TREC are by no means typical of many, or indeed most, DR or TR contexts, so care is needed in transferring results, especially since interactive searching is not a primary object of study. So while these tests are on a gratifyingly larger scale than earlier ones, they have their limitations. More importantly it is far too easy in DR, and hence in TR, to intuit wrongly that things do or will work well, whether these are old approaches, old approaches dressed up in shiny modern technological guises, or truly new approaches. It is essential to test, test, and test again.

References

[1] Buckley, C. The importance of proper weighting methods. To appear in *ARPA Workshop on Human Language Technology* (March 21–24, Plainsboro, NJ). Morgan Kaufmann, San Mateo, CA, 1993.

[2] Callan, J. P. and Croft, W. B. An evaluation of query processing strategies using the TIPSTER collection. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (June 27 - July 1, Pittsbrugh, PA). ACM/SIGIR, New York, 1993, pp. 347–355.

[3] Church, K. W. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing* (Feb. 9–12, Austin, TX). ACL, Morristown, NJ, 1988, pp. 136–143.

[4] Copestake, A. and Sparck Jones, K. Natural language interfaces to databases. *The Knowledge Engineering Review* 5, 4 (1990), 225–249.

[5] Croft, W. B. and Das, R. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the Thirteenth Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval (Sep. 5–7, Brussels). ACM/SIGIR, New York, 1990, pp. 349–365.

[6] Croft, W. B., Turtle, H. R., and Lewis, D. D. The use of phrases and structured queries in information retrieval. In *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Oct. 13–16, Chicago, IL). ACM/SIGIR, New York, 1991, pp. 32–45.

[7] Engelien, B. and McBryde, R. *Natural Language Markets: Commercial Strategies.* Ovum Ltd, 7 Rathbone Street, London, 1991.

[8] Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G., and Monarch, I. A. Automatic indexing using selective NLP and first-order thesauri. In *RIAO 91 Conference Proceedings: Intelligent Text and Image Handling* (Apr. 2–5, Barcelona), 1991, pp. 624–643.

[9] Fagan, J. L. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* PhD dissertation, Department of Computer Science, Cornell University, Sept. 1987.

[10] Fuhr, N. Models for retrieval with probabilistic indexing. *Inf. Process. Manage.*, 25, 1 (1989), 55–72.

[11] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. The vocabulary problem in human-system communication. *Commun. ACM*, 30, 11 (1987), 964–971.

[12] Hahn, U. Topic parsing: accounting for text macro structures in full-text analysis. *Inf. Process. Manage.*, 26, 1 (1990), 135-170.

[13] Harman, D. *The First Text REtrieval Conference (TREC-1).* National Institute of Standards and Technology Special Publication 500-207, Gaithersburg, MD 20899, 1993.

[14] Hindle, D. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics* (June 6–9, Pittsburgh, PA). ACL, Morristown, NJ, 1990, pp. 268–275.

[15] Hirschman, L., Grishman, R., and Sager, N. Grammatically-based automatic word class formation. *Inf. Process. Manage.*, 11 (1975), 39–57.

[16] Jacobs, P. S. and Rau, L. F. SCISOR: Extracting information from on-line news. *Commun. ACM*, 33, 11 (1990), 88–97.

[17] Krovetz, R. and Croft, W. B. Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10, 2 (1992), 115–141.

[18] Lewis, D. D. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (June 21–24, Copenhagen). ACM/SIGIR, New York, 1992, pp. 37–50.

[19] Milstead, J. L. *Subject Access Systems.* Academic Press, Orlando, 1984.

[20] Parsaye, K., Chignell, M., Khoshafian, S., and Wong, H. *Intelligent Databases.* Wiley, New York, 1989.

[21] Pritchard-Schoch, T. Natural language comes of age. *Online*, 17, 3, (1993), 33–43.

[22] Salton, G. and McGill, M. *Introduction to Modern Information Retrieval* McGraw-Hill, New York, 1983.

[23] Salton, G. Another look at automatic text-retrieval systems. *Commun. ACM*, 29, 7 (1986), 648–656.

[24] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. *J. American Society for Information Science*, 41, 4 (1990), 288–297.

[25] Salton, G. and Buckley, C. Global text matching for information retrieval. *Science*, 253 (1991), 1012–1015.

[26] Schwartz, M. F., Emtage, A., Kahle, B., and Neuman, B. C. A comparison of Internet resource discovery approaches. *Computing Systems*, 5, 4 (1992).

[27] Smadja, F. A. From n-grams to collocations: An evaluation of Xtract. In *29th Annual Meeting of the Association for Computational Linguistics* (June 18–21, Berkeley, CA). ACL, Morristown, NJ, 1991, pp. 279–284.

[28] Sparck Jones, K. Search term relevance weighting—some recent results. *J. Information Science* 1 (1980), 325–332.

[29] Sparck Jones, K. *Information Retrieval Experiment.* Butterworths, London, 1981.

[30] Sparck Jones, K. and Tait, J. I. Automatic search term variant generation. *J. Documentation*, 40, 1 (1984), 50–66.

[31] Sparck Jones, K. Fashionable trends and feasible strategies in information management. *Inf. Process. Manage.*, 24, 6 (1988), 703–711.

[32] Strzalkowski, T. TTP: a fast and robust parser for natural language. In *Proceedings of the Fifteenth International Conference on Computational Linguistics* (Aug. 23–28, Nantes). ACL, Morristown, NJ, 1992, pp. 198–204.

[33] Turtle, H. R. and Croft, W. B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9, 3 (1991), 187–222.

[34] Willett, P. *Document Retrieval Systems* Taylor Graham, London, 1988.

[35] Woods, W. A. Progress in natural language understanding — an application in lunar geology. *Proceedings of the 1973 National Computer Conference, AFIPS Conference Proceedings*, Vol 42, 441–450, 1973.

[36] Young, S. R. and Hayes, P. J. Automatic classification and summarization of banking telexes. In *Second Conference on Artificial Intelligence Applications* (Dec. 11–13, Miami Beach, FL). IEEE Computer Society Press, Los Alamitos, CA, 1985, pp. 402–408.