Technical Report

UCAM-CL-TR-234
ISSN 1476-2986

Number 234

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Two tutorial papers: Information retrieval & Thesaurus

Karen Spärck Jones

August 1991

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/TechReports/

ISSN 1476-2986

TWO TUTORIAL PAPERS:

INFORMATION RETRIEVAL

&

THESAURUS

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, England

August 1991

Abstract

The first paper describes the characteristics of information retrieval from documents or texts, the development and status of automatic indexing and retrieval, and the actual and potential relations between information retrieval and artificial intelligence. The second paper discusses the properties, construction, and actual and potential uses of thesauri, as semantic classifications or terminological knowledge bases, in information retrieval and natural language processing.

These papers will appear in the Encyclopedia of Artificial Intelligence (ed Shapiro), 2nd ed, New York: Wiley.

INFORMATION RETRIEVAL

Karen Sparck Jones
July 1991

Information retrieval (IR) has not generally attracted the attention of workers in AI. But there has recently been a growth of interest in IR, and AI may be able to make some specific contributions, thought the general problem is highly intractable and the automated intelligent information assistant is a distant goal.

BASIC ELEMENTS

"Information retrieval" has conventionally been taken to refer to the retrieval of documents like scientific papers as opposed to books, from large but specialised bodies of material, through the specification of document content rather than via keys like author names, with the help of automation. It has been distinguished from library retrieval based on traditional, limited, book catalogue information through differences in the nature of the file items, which call for detailed and specialised descriptions and correspondingly refined search procedures.

Thus as a simple model, an IR system consists of a file of documents with brief content characterisations in the form of **index descriptions in an **indexing language; a procedure for **searching the file given the index description of a user's **request for documents derived from his **need for information about something; and a **matching criterion for evaluating comparisons between request and document descriptions. The system also includes procedures for deriving the index descriptions of documents and requests. It is driven by the user's need and the consequent assessment of retrieved documents for their **relevance to this need.

In an automated system the basic search and match operations on descriptions will be carried out automatically, though modifications of the request description, i.e. the search specification, may be done manually. The construction of descriptions may or may not be automatic. The determination of the user's need and its expression as a request may be automatically assisted but is essentially manual; the assessment of relevance is also a human, and specifically end-user, operation.

In practice this simple model is most often modified to take account of the fact that the machine files do not contain the actual document texts themselves but only **surrogate, notably abstract, texts, so indexing and even relevance assessment are based on these surrogates.

Database Context

Modern IR techniques emerged to meet the challenge of the vast growth of literature in the last forty years by taking advantage of the storage and processing resources of computers. The subsequent development of communications and networking has allowed an enormous development of on-line search services handling thousands of bibliographic and related databases, with individual bases containing millions of items (Gerrie, 1983). Full end-document text bases, first established for legal and news material, are becoming more common, and there are more and more varied quasi-text or "factographic" bases verging on conventional non-bibliographic databases in form and content. The range of resources also includes traditional book catalogues with many millions of entries. Improvements in terminal technology as well as networking have encouraged end-user searching without the help of professional information officers or **intermediaries.

User Need

The core operations in a retrieval system are clearly clearly document and request description, or indexing, on the one hand, and searching and matching, or retrieving, on the other. But the performance of a retrieval system also depends heavily on the extent to which the user's information need has been identified, and an appropriate request for documents to meet it has been expressed. Need and request have to be distinguished as retrieved documents are eventually and properly assessed for relevance to the user's need, not for relevance to the request, which can only be a substitute predictor of relevance to need. In identifying the user's need it is also necessary to take account of the wider problem context supplied by background components of the user's situation like educational status, work goals and so forth, which explain the need and may influence, but do not constitute, the request.

Document Relevance

Relevance refers to the user's perception of the value of document content in relation to his information need. This judgement is strictly ad occasionem, and as need and fulfilment are mutually determining, relevance has to be taken as an undefined primitive notion, with important consequences for system evaluation (see below). It is a fundamental fact of IR life that what makes a document relevant is not obvious. This places one set of constraints, emphasising hospitality, on system design. But it is also a fundamental fact of IR life that the absolute number of documents and, even more, proportion of the collection that is relevant to a request will be very small. This places a second conflicting set of constraints, emphasising selectivity, on system design.

Performance Measurement

System performance is therefore normally measured primarily in terms of the system's ability to retrieval all and only relevant documents. If a search partitions a collection of size N into four sets, r = relevant retrieved, $M-r$ = matching but not relevant, $R-r-$ = relevant but not retrieved, and $N-M-R+r$ = non-relevant and not retrieved, then **precision is defined as r/M and **recall as r/R, and performance can be defined as some combined function of precision and recall. With an all-or-

nothing matching function of the kind commonly used in commercial services, where search specifications are Boolean functions of terms, performance can be simply indicated as a precision-recall pair, and performance for many requests can be derived by straightforward averaging. With functions ranking output, performance can be obtained by computing precision for standard recall values (say .1, .2 ...), with averaging. Empirical observation has consistently shown that there is an inverse relationship between precision and recall, and also that is is in general very hard to attain, let alone improve on, values of .5 for both at the same time, especially for average performance (Lancaster, 1979; van Rijsbergen 1979; Sparck Jones, 1981; Salton and McGill, 1983).

Document Description

Descriptions, or representations, of documents are needed for retrieval. They are not simply regrettable substitutes for full texts associated with a lack of the space to store and time to search the texts themselves. Document index descriptions are necessary filters for the end-user who cannot be expected to read many whole documents directly. This information reduction, which titles classically offer, is essential with human searching; and it meets a human need even when searching is done automatically, and the end user has only to assess a relatively or even absolutely small output document set. But document descriptions have a much more important role as a means of optimising document-request matches. Document descriptions, like summaries, pick out the key information in a document and make it explicit. So if the user's request is for documents about X, matching the description "X" should indeed return documents that are about X.

Document descriptions are normally provided when documents enter the collection, though in practice they may be constructed at run time only for those documents extracted from the file using some first-pass coarse filter. But as descriptions are logical objects they be constructed a posteriori, at search time, rather than a priori, since a request can be taken as an index description for the documents it matches. Thus if the index description for a request consists of a set of three words, any document (or surrogate) texts which also contain these three have this set as an index description. The advantage of independent, as opposed to request-dependent, document descriptions is that it is easier to ensure the description reflects key document content. The disadvantage is that other content is lost. With request-dependent descriptions access to documents is open, but is difficult to control so as to ensure matches on significant rather than incidental document content.

Request-dependent description has been a natural consequence of the ability to store and search masses of running text, most often in the form of abstracts, but also, and increasingly, as full document texts. With abstracts, as these are deliberately designed to be condensed representations of full documents, request-based text searching can be very effective. It may also be useful, as with legal, news, or intelligence material, where any occurrence of the specified words may be valuable regardless of overall document topic. But in general, the flexibility and sensitivity to the user's perspective given by request-dependent indexing has a cost in mismatches.

Flexibility can be achieved with independent document descriptions by the way matching is defined, and by procedures for modifying given descriptions, say by substituting more general for specific terms, which are allowed in searching. But there is still the problem of failure to capture document content in the initial description; this can be particularly unfortunate when there are topic and perspective shifts in a subject field, as older documents may become inaccessible.

Key Problems

The key problems of IR, bearing directly on indexing and indirectly on retrieving, are of normalisation and discrimination, given a context dominated by uncertainty. There is uncertainty in working with descriptions of documents rather than documents themselves; with topic characterisations stemming from different perceptions, and with the linguistic expressions reflecting these; with requests from users who are not yet informed; and with relevance as an idiosyncratic relationship. A balance has to be struck between descriptive normalisation and descriptive discrimination. Normalisation is intended to reduce uncertainty by limiting variety, and is therefore determined by the units and relations of the indexing language and also by indexing policies affecting e.g. description detail. Discrimination is designed to separate the few documents relevant to a request from the many non-relevant ones, and is equally determined by indexing resources and policies. Indexing a document with a few general terms will blur differences, promoting recall at the expense of precision; indexing a document with a complex, non-modifiable and specific description will advance precision if any matches occur.

AUTOMATED SYSTEMS

The essential processes of IR as defined earlier are those of information retrieval in a much broader sense, and hence are also those found in libraries. Their particular properties stem from the distinctive features of the task context. Thus the nature of the documents and requests involved, and power of the machine, have led to real and important differences in indexing and retrieving, seen in their simplest form in the ease with which permutation and selection can be done on a set of **keys constituting an index description (van Rijsbergen, 1979; Salton and McGill, 1983; Willett, 1988).

Distinguishing items in, and selecting them from, a large mass of similar ones, typically without any leverage from relatively unequivocal keys like author names, calls for refined and variable document descriptions. These descriptions have also often to be provided within the context of rapidly developing and changing subjects and disciplines, as in many areas of science and technology. At the same time, there is no practical limit to the number of access routes to a single document topic description,as there used to be with hand-generated catalogues, let alone any need to choose the best single subject location for the physical documents, as there often is with books on library shelves. There is therefore no good basis, or real need, for global description schemes intended to characterise entire or large areas of knowledge in a systematic way, and to place all the documents in a collection within a single analytic framework, like those of traditional library classifications.

Indexing Forms

In IR therefore, it is customary to provide document descriptions consisting of one or more, perhaps many, concept labels drawn from an indexing vocabulary which is not in itself systematically organised into a single whole depending, for example, on hierarchical inclusion. But within this general framework for indexing, covering both the language used and descriptions produced, there are many different possibilities under the major headings of language type, description form, and processing mode. Broadly speaking, the indexing language may be natural language or an artificial **controlled language (normally using ordinary words in restricted or non-standard ways). The description form may be **precoordinate, where individual words are combined to give fixed complex wholes, or **postcoordinate, where individual words can be freely combined to give ad hoc topic characterisations. The processing mode may be by **derivation, with description material extracted from the source document text, or by **assignment, with document elements motivating the assignment of items from an independent indexing language.

It is customary to refer to the basic units of document and request description as **terms. These are normally word-like, and if extracted may be called **keywords, and while (fixed) compounds may be allowed, it is more usual to refer to complex multi-word descriptive units as **subject headings. The indexing **vocabulary may then consist of the basic terms, which if controlled may be called **descriptors, or of whole subject headings. The vocabulary itself may be given an organised relational structure as a **thesaurus, list of subject headings, or classification.

In practice, many different specific choices and combinations of language, description and mode are to be found, with many variants in particular in the treatment of terms, especially in relation to vocabulary control; of relations between terms, which may be implicit or explicit and natural or artificial; and of description integrity, as this interacts with the use of simple or complex descriptive units (Foskett, 1977; Lancaster, 1979; Chan, Svenonius and Richmond, 1985).

Retrieval Operations

Retrieving covers searching and matching. Searching in the broadest sense can cover the identification of the user's need and the expression of the user's document request, for example during browsing of a subject heading list, and at this point overlaps with request indexing. But it is often interpreted more narrowly to refer to the modification of a given request index description, i.e. search specification, to change the volume or nature of the matching document set. This may be done manually or automatically. With automated systems, it can also cover filtering operations designed to limit detailed matching to part files, analogous to a human search downwards in a conventional library classification like the UDC. In the narrowest sense, searching refers to the operations required to access items in the database file; with large databases these may be very complex and may depend on sophisticated file structures, or exploit novel architectures (Stanfil, Thau and Waltz, 1989). Matching refers to the individual comparison between request and document

descriptions. Matching can range from simple exact matching, as with Boolean expressions, to elaborate calculation to compute complex scoring functions (Salton and McGill, 1983).

In general, it is important to recognise that retrieval is usually an interactive process, and that searching is iterative (Belkin and Vickery, 1985).

Manual Indexing

With automated systems the main distinction between systems hitherto has been whether indexing is essentially manual or automatic. (Initial request formulation is a primarily human process, though it may be be indirect, as in the use of a reference document example, and also machine-aided; subsequent reformulation, normally directly of the search specification, may be manual, automatic, or some mix of the two.) In manual indexing concept labels may be extracted from the document or surrogate (or request) text, or assigned, either freely or from a pre-specified vocabulary; and the relations within or between labels may be similarly reflective, free, or constrained. Manual indexing has been particularly associated with the use of a controlled artifical indexing vocabulary and precoordinate subjects, since these normally require some understanding of the document, though with simple controlled terms automatic assignment from text clues is feasible. Controlled vocabularies have also to be constructed and maintained manually, implying substantial effort for large subject areas. Automatic indexing is normally associated with the use of natural language, applying criteria for the selection of key text elements for use in their own right or as leads to other or additional labels (Lancaster, 1979; Willett, 1988).

AUTOMATIC INDEXING

The automatic indexing and retrieval techniques established by IR research have been in part a response to the fundamental constraints stemming from uncertainty, and in part determined by the lack of natural language understanding (NLU) or even natural language processing (NLP) resources. But they have at the same time been influenced by experiments suggesting that the elaborate approaches, involving deep analysis and complex description, of manual indexing do not pay off in retrieval performance, perhaps because uncertainty implies a corresponding simplicity rather than a complementary refinement. These automatic techniques have also been motivated by a belief that the actual words used in natural language text are of crucial importance as direct content indicators. In addition, the statistical approaches adopted have reflected the importance of scale in IR as opposed to many other NLP applications: for example, if a term occurs in many documents it will be a poor selector even if it is an accurate indicator of individual document content.

Basic Model

The basic model on which automatic indexing and retrieval has been built exploits natural language by using simple extracted terms which are linked by coordination, i.e. simple conjunction, and are weighted by their distribution in and

across documents (or their surrogates). The terms will normally be **stems or fragments, possibly crudely defined from a linguistic point of view. All the terms occurring in a document are candidate description terms. The candidate indexing vocabulary for a collection is the union of these document terms. The actual vocabulary can in principle be manipulated purely statistically to eliminate non-discriminating terms, but common function words are in practice removed via a **stop list. The operational vocabulary is therefore confined to content words and function words are not retained for their relational utility.

Some content words may also appear unhelpful, but it is better to retain the entire content vocabulary and indicate the status of terms by explicit weights, as these can vary for different documents and can also be manipulated to reflect the properties of requests and relevance assessments, and the behaviour of term cooccurrences. A simple, demonstrably useful weighting formula balances term frequency within documents against frequency across documents: good terms are those with a high within-document frequency but a low across-document frequency, where a term's across-document frequency is just the number of documents in which it occurs. Specifically, if Dk is the number of documents in which term k occurs and N is the size of the collection, we can capture the germane properties of across-document frequency by defining k's **inverse document frequency Ik = log2N - log2Dk. Then if Fik is k's frequency within document i, we cover both aspects of term behaviour by defining k's weight for i as wik = Fk . Ik.

Requests are indexed in the same way, and request document matching is then determined by some suitable formula. In the simplest case, where there is little within-document variation (as with abstracts), weighting can simply be by inverse document frequency, and for the request terms alone. So for term k in request j, wik = Ik. Matching a request against the collection will give a set of scores, each summing the weights of terms occurring in a document, which can be used to rank documents for presentation to the user. Matching is thus more flexible, and more complex in its effects, than with the Boolean formulae found in most operational services, where the complete search specification has to match for a document to be retrieved.

Model Extensions

The basic model may be elaborated, still on a statistical basis, in various directions. One is to group terms to provide alternative or additional matching terms, either directly, or indirectly via the use of class names as derived descriptors; statistical constraints may also be applied to the formation of compound terms, or term phrases. A second is to seek document groups based on shared terms, both to reduce search effort by limiting the initial steps of search matching to group descriptions, and also to concentrate like (and hence hopefully co-relevant) documents. A third is to apply **relevance feedback, exploiting relevance assessments for initially retrieved documents to modify requests, say by adding new terms from relevant documents and/or reweighting the given terms to reflect their relevant document incidence. Thus using r, R, n = Dk and N to construct a four-way collection partition like that given earlier for searching but now defining the

way an individual term occurs in relevant or retrieved documents, for each term k in request j we can obtain a weight

$$wjk = \log2 \left[(r / (R-r)) / ((n-r) / (N-n-R+r)) \right]$$

with some appropriate adjustment, say adding .5 to the value for each set, to reflect the fact that the term's future relevance utility is being estimated (van Rijsbergen, 1979; Salton and McGill, 1983; Willett, 1988).

Model Performance

The basic model has been shown to compete effectively with conventional 'higher-quality' manual indexing in many, very different tests; and while grouping strategies have not proved generally helpful, relevance weighting is particularly valuable, especially in the first iteration in searching. But operational systems using natural language have until recently normally been very crude, without ranking or weighting, and while the necessary conclusive experiments on a really large scale have not been done, there is little doubt that respectable natural language systems could be provided (Sparck Jones, 1981).

Operational systems rely heavily on sensible request formulation by the user. This is always important, and while there may be no real substitute for the user's initiative, the basic model can be developed to support automatic request modification as just indicated, or applied in conjunction with the user's own changes. This has, however to be done with care, as the user's perceptions of what is useful may not fit the statistical realities. Automation also makes it easy to use other types of search key, like citations, to which statistical techniques may equally apply, and developments in computing and especially in interactive and display technology have encouraged a more varied and cheerful use of all the available information, e.g. searching on controlled language terms as if they were ordinary natural language ones, and have assisted search formulation through e.g. related term displays.

On the theoretical side, considerable effort has been put into, and some progress has been made in, developing a coherent formal model to which both the strategies adopted for the different components of an IR system and the alternative approaches to a particular operation, like term weighting, can be related. Much of the research done in IR has been in terms either of the vector model associated with Salton and Yu or of van Rijsbergen and Robertson's Bayesian model (Salton and McGill, 1983; van Rijsbergen, 1979). For some it is evident that the required overall model has to be a probabilistic one, but alternative claims can be made for e.g. a fuzzy logic model. One major problem with general theories proposed so far is that they have usually been so abstract it is difficult to know how to apply them practically in system design and operation. At a lower level a good deal of attention has been paid to combining weighting with Boolean requests, as a way of applying research results within the conceptual and data management framework of operational systems. But there are problems with this as the two approaches impose very different constraints on search specifications.

SYSTEM TESTING AND EVALUATION

IR systems are very complex, with many data variables and system parameters. Their behaviour is not well understood, which leads to conceptual problems that make it very hard to design proper tests. (These also tend to be expensive.) To determine performance for any particular combination of indexing and retrieving methods, or to compare alternative combinations, it is necessary to average across requests, and it may be very difficult to find a sound way of doing this (is a match on only 2 terms, but out of 3, better than one on 4 out of 7?). It is also generally necessary to determine what relevant documents there are to be retrieved, but it is clearly impossible to assess every document in a large collection for relevance to a user need. Sampling for assessment is, however, problematic when there are only a few relevant documents, as is usually the case. The normal strategy is therefore to search for a request using many different methods and to pool the outputs for assessment. This allows a measure of **relative recall for any individual method, but the relevance set is biassed towards those documents that are easy to retrieve. It may be helpful to use a measure which can be oriented towards recall or precision, to reflect user preferences. Many specific measures have been proposed, and it is important with operational systems to take account of factors like search effort as well as e.g. money costs; but user satisfaction, though important, has to be treated with caution as perceived and real performance, say for recall, can be very different. There is a further difficulty in establishing whether performance differences are significant, given the lack of suitable non-parametric significance tests. The sign test is of some, but only limited, use (Lancaster, 1979; Sparck Jones, 1981).

With modern interactive technology testing becomes much more difficult, because searching is determined by assessment. The user cannot be asked to apply alternative devices to the same starting request, for parallel searches before any output assessment, so larger request samples are needed. Testing to allow general claims for particular methods, as opposed to collection-specific claims, is very costly because evaluation across different collections representing different data contexts is needed. In this case, the notion of **test collection may be taken to refer not only to the document file, but to a particular request and relevance assessment file representing some particular user community.

ARTIFICIAL INTELLIGENCE

AI, defined here as computational reasoning over world knowledge, has not figured significantly in IR so far, but has several possible roles which are beginning to be explored (Croft, 1987; Jacobs, 1990). The most important potential roles are in the central processes of indexing and retrieving referring explicitly to subject or topic information. There are also other tasks for which expert system (ES) technology might be useful. Finally, AI may have a part to play in integrating the increasingly varied information bases and information management tasks that technology is offering the user at his desk.

Intelligent Indexing

The claim for AI in relation to indexing can be made in a very strong form. This is that as the user is really interested in the information supplied by

documents, this information should be extracted to form a single knowledge base replacing the documents themselves, and also of course their index descriptions. Authority data could be preserved as appropriate in the base, but information is no longer scattered and remote. Searching in consequence becomes reasoning over the knowledge base, and IR is thus assimilated to question answering.

The main problems with this proposal (apart from its practical feasibility) are that the actual expression of information in a natural language document text itself conveys information, and that IR is not necessarily, or even mostly, question answering (Sparck Jones, 1990). Even if it is formally question answering, it is not substantively question answering under the tacitly assumed fact retrieval paradigm, involving knowledge representation in some logical formalism.

A more moderate version of the claim is for a knowledge base superstructure which embodies only the essential collection content, but which allows initial inferential searching of a well-organised kind on this, bottoming out in pointers to documents supplying more detail. The assumption here is that AI techniques would provide a more effective, because more powerful, means of access to the documents than that represented by devices like conventional library classification schemes. But there is now a problem in the precise nature of the links between knowledge base and documents, which does not arise with the less exigent classification case, and with the relation between question-answering operations on the knowledge base and any further topic searching operations on the document base.

A yet weaker proposal is for AI-type individual document descriptions e.g. in some frame form, as providing a more organised characterisation of the documents that allows inferential description searching, both on individual descriptions and through systematic links between multiple descriptions (Lewis, Croft and Bhandaru, 1989). With this approach there would be a knowledge base only in an emergent and weak sense. It is however a fact that both arguments and often examples here are often unwitting reruns of the familar case for elaborate indexing of the kind hitherto provided manually, which has not been proven significantly better than much simpler automatic procedures. The specific issue is whether descriptions of this type are too constraining, even allowing for some modulation and relaxation.

The final approach is to operate within the generic natural language term model, but to seek to use AI rather than statistics to extract key terms for documents, and more importantly, to identify well-founded compound terms (for instance embodying case relations). Full NLU is beyond the present state of the art, but current NLP techniques might be useful (Lewis, Croft and Bhandaru, 1989). The particular issues here are first, whether purely syntactic processing (local or global) gives sufficiently better compounds than pure proximity or association methods to be worth the effort; and secondly and more materially, whether term identification can be improved if analysis is semantic as well as syntactic, assuming the necessary semantic resources (lexicon, patterns) can be supplied without too much effort for the wide-ranging material encountered in ordinary document collections. Some operational systems already use syntactic processing. But whether this or semantic analysis, on which only some limited research has yet been done, significantly

improve performance has still to be experimentally established (Sparck Jones and Tait, 1984). The plausible argument that sophisticated analysis is needed to identify real as opposed to false coordinations is unfortunately irrelevant if false coordinations do not actually match documents.

The view that strong claims for AI approaches to indexing may be misconceived refers to the general case. The situation may be quite different in specialised contexts, with particular types of material or use, where question answering on a knowledge base has a part to play (Jacobs and Rau, 1988).

Intelligent Searching

In searching, AI would be exploited to 'automate the intermediary', i.e. to replace the expert who currently helps users to formulate their requests, indexes these, and conducts searches, frequently through on-line services, on the user's behalf. This requires extensive knowledge of generic user properties, available information resources, indexing languages and practices, search strategies and so forth, as well as considerable general and domain knowledge (Belkin et al, 1987). An analysis of the intermediary's or reference librarian's practical knowledge and skills suggests that automating them is a task well beyond the state of the art. Whether less ambitious support, requiring more complementary initiative from the end user, would be useful, and could be provided, needs and deserves investigation. Expert system techniques have already been applied to request formulation and search specification (Vickery et al, 1987), and more limited facilities e.g. for translating requests into controlled languages, have been studied and can work (Pollitt, 1987). At the same time, AI inference methods have been applied to searching and matching using (manually constructed) search concept specifications (Tong and Applebaum, 1988). It has also been suggested that AI techniques could be used not for individual searches, but to allow longer-term system tailoring to the individual. Whether this could be more effectively done than with the non-AI techniques currently available for modifying individual searches or standing interest profiles also needs investigation.

Intelligent Service

It is possible that expert system technology could be used for other less exigent information management tasks still referring to subjects or topics, like categorisation, routing or database selection, though again much simpler term-based techniques might suffice, and also for other support functions like cataloguing. Some work has already been successfully done which combines term extraction without any NLP with rule-based exploitation of the extracted term data for categorisation (Hayes and Weinstein, 1990).

Intelligent Systems

The most ambitious proposed role for AI is in sustaining the integrated information system of the future. The aim is not just integration so that the user does not have to work explicitly within different subsystems, say for text preparation

and document retrieval, as he typically does now. The goal is a system providing a positive response exploiting different information resources, shifting initiative from the user to the system in the way adaptation also does. This integration presupposes an infrastructure, in the form of a common knowledge base and reasoning apparatus, to make content links between different types of information and function bearing on the user's inferred need. However a cool look at the range of information types and management functions, both private and public, involving not only a single user but many users, that any serious system would have to support, suggests that this is a remote dream (Sparck Jones, 1990). Much weaker mechanisms relying on associations between the words used in different subsystems may, on the other hand, be not only attainable but also the conceptually correct ones, given the real heterogeneity of these systems.

However this does not imply that AI may not contribute to the individual subsystems within a multi-purpose system, say to create a specialised information base (Young and Hayes, 1985). Linking with other systems would still rely on the language of the subsystem.

CONCLUSION

Conventional operational retrieval systems are well-established and useful. But they are institutionalised and their size makes them difficult to modify, so change stimulated by research findings has been very slow, particularly since helpful mundane improvements, like better document delivery, appear of more value to the consumer. Developments in information technology are having an obvious impact on information retrieval, especially for the end user, but it is not yet clear how these will interact with the approaches suggested by IR or AI research. AI research in the area is itself only beginning and, while promising particularly for some specialised cases like message interpretation, has not yet demonstrated solid, generally applicable results. AI techniques may be relevant for some purposes, like request formulation, which are outside the scope of IR theories. But there is a more interesting intellectual challenge in whether knowledge-based AI and statistically-based IR can be legitimately or effectively combined (Salton and McGill, 1983; Lewis, Croft and Bhandaru, 1989). Modest operations like weighting terms delivered by NLP are clearly feasible, but the issue is what more there may be to do.

BIBLIOGRAPHY

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

N.J. Belkin et al "Distributed Expert-Based Information Systems: An Interdisciplinary Approach", Information Processing and Management 23, 395-409, 1987.

L.M. Chan, P.A. Richmond and E. Svenonius (Eds), Theory of Subject Analysis: A Sourcebook, Libraries Unlimited, Littleton, Colorado, 1985.

W.B. Croft (Ed) Special Issue: Artificial Intelligence and Information Retrieval, Information Processing and Management 23 (4), 249-366, 1987.

A.C. Foskett, The Subject Approach to Information, Bingley, London, 1977.

B. Gerrie, Online Information Systems - Use and Operating Characteristics, Limitations, and Design Alternatives, Information Resources Press, Washington DC, 1983.

P.J. Hayes and S.P. Weinstein, "CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories", Proceedings of the Second Annual Conference on Innovative Applications of Artificial Intelligence, Washington DC, American Association for Artificial Intelligence, 1990.

P.S. Jacobs and L.F. Rau, "Natural Language Techniques for Intelligent Information Retrieval", Proceedings of the 11th International Conference on Research and Development in Information Retrieval, (Grenoble, France), Association for Computing Machinery Special Interest Group on Information Retrieval, 85-99, 1988.

P.S. Jacobs (Ed), Text-Based Intelligent Systems: Current Research in Text Analyisis, Information Extraction, and Retrieval, Report X, General Electric, Schenectdady, New York, 1990.

F.W. Lancaster, Information Retrieval Systems: Characteristics, Testing and Evaluation, 2nd ed., Wiley, New York, 1979.

D.D. Lewis, W.B. Croft and N. Bhandaru, "Language-Oriented Information Retrieval", International Journal of Intelligent Systems 4, 285-318, 1989.

A.S. Pollitt, "CANSEARCH: An Expert Systems Approach to Document Retrieval", Information Processing and Management 23, 119-138, 1987.

C.J. van Rijsbergen, Information Retrieval, 2nd ed., Butterworths, London, 1979.

G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

K. Sparck Jones (Ed), Information Retrieval Experiment, Butterworths, London, 1981.

K. Sparck Jones and J.I. Tait, "Automatic Search Term Variant Generation", Journal of Documentation 40, 50-66, 1984.

K. Sparck Jones, "Retrieving Information or Answering Questions?", The British Library Annual Research Lecture, The British Library, London, 1990.

C. Stanfil, R. Thau and D. Waltz, "A Parallel Indexed Algorithm for Information Retrieval", Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, (Cambridge, Massachusetts), Association for Computing Machinery Special Interest Group on Information Retrieval, 88-97, 1989.

R.M. Tong and L. Applebaum, "Conceptual Information Retrieval from Full Text", RIAO 88, Proceedings of the Conference on User-Oriented, Content-Based Text and Image Handling (Cambridge, Massachusetts), 899-909, 1988.

A. Vickery, H. Brooks, B. Robinson and B. Vickery, "A Reference and Referral System Using Expert System Techniques", Journal of Documentation 43, 1-23, 1987.

P. Willett (ed.), Document Retrieval Systems, Taylor Graham, London, 1988.

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.

THESAURUS

Karen Sparck Jones
August 1991

The term "thesaurus", meaning a treasury, may refer to collections of all kinds, but was early used to refer to collections of words. Its modern uses stem from Roget's Thesaurus of English Words and Phrases (Roget, 1852). These uses all treat a thesaurus as a classification of linguistically expressed, and especially lexical, information, which is the essential property of Roget's Thesaurus. But they interpret and exploit this in different ways. The three modern uses of a thesaurus are (1) as a vocabulary reference, familiar from printed forms; (2) as an information retrieval (IR) device; and (3) as a natural language processing (NLP) resource. This article focuses on (2) and (3), considering their common and distinct elements within the context of linguistic information processing as a proper concern of AI; but it starts with an account of Roget's Thesaurus, as this was a historically important stimulus to proposals for IR and NLP and also illustrates key features of a thesaurus exceptionally well.

FOUNDATION

Roget's Thesaurus

The Thesaurus, as Roget says in his Introduction, was a classification of concepts: "the words and phrases of the language are here classed ... strictly according to their signification". The basic concepts, or **heads of classification, approximately 1000 in number, were organised in contrasting pairs, e.g. 'Expansion' and 'Contraction', and were grouped hierarchically in the structure given in the Table of Contents. Thus 'Expansion' and 'Contraction' were grouped under 'Dimensions', which was in turn combined with 'Form' and 'Motion' under 'Space', and so forth, within a single all-embracing hierarchical scheme. Each concept had a naming label, representing the head, and subsumed a class of words having similar meanings or topic relationships justifying their being placed together under the same head, e.g. "augmentation", "swelling" and "knob" under 'Expansion'. Each class was subdivided by the major parts of speech, and within each such syntactic category there were progressively finer subdivisions signalled by paragraphing and other punctuation, down to the finest classes consisting of close synonyms, subsequently called **rows (Sparck Jones, 1986). Cross references from one head to another, indeed between specific within-head locations, reflected the fact that concepts of the kind represented by the heads could not really be viewed as exclusive. Thus e.g. "knob" leads from 'Expansion' to 'Rotundity'.

A word could appear in as many heads as were justified by its meanings or senses (regardless of whether these distinctions were etymologically motivated or

not). The display of the thesaurus information in the form of words grouped under heads, with the heads ordered according to the Table of Contents, was complemented by an alphabetical listing linking each word to the set of heads in which it occurred.

All of the essential properties of a thesaurus as a **semantic classification appeared in Roget's Thesaurus. Each class can be seen in two complementary ways: as showing what words are similar because they stand for a common, autonomously defined notion taken as represented by the head name; or alternatively, as defining a common notion through their recognised similarity of meaning. This duality is most clearly seen in the basic heads, but also applies, though in the weaker form of shared elements of meaning, to the superclasses which combine lower-level classes under given labels; it clearly applies in principle to the subclasses as well, since these could be labelled though they are not in fact. Thus the Thesaurus at the same time distinguishes word senses through their thesaurus class allocations and defines or at least characterises these senses in terms of the explicitly or implicitly given class concepts. However as the head names are just ambiguous natural language words, it is better to view class concepts as implicitly defined via similarity than as explicitly defined by reference to an independent concept, since this is arbitrarily labelled and inaccessible. At the same time, describing the members of a class as similar in meaning is sometimes too strong, since other types of associative relation figure which are more properly viewed as determining topic classes; but the basic notion of a thesaurus as founded in similarity stands.

Roget offered his Thesaurus as a practical aid to thinking and writing, but at the same time related this function to serious scientific need by referring to the idea of a Universal Character, i.e. universal language, which was a significant feature of 17th Century scientific thought, notably among members of the Royal Society, of which Roget was later Secretary. A Universal Character would supply a comprehensive set of distinct, basic concepts which, when applied with an appropriate combinatory grammar, would constitute at least a transparent and unambiguous, interlingual means of describing scientific phenomena or further, as envisaged by Leibniz, provide an apparatus for scientific reasoning (Sparck Jones, 1972, 1986 Appendix 1; Knowlson, 1975; Slaughter, 1982; Large, 1985). This view of a thesaurus as a component of an **interlingua has also played an important role in modern uses of a thesaurus.

Reference thesauri

Roget thus regarded the classificatory structure of his Thesaurus, and particularly that supplied by the Table of Contents, as a central element of the whole. Many of his successors, however, failed to grasp or appreciate his intentions and, viewing the Thesaurus primarily as a synonym display, thought its value would be enhanced by reorganising its body alphabetically and by providing additional entries reflecting a stricter interpretation of synonymy (e.g. Lewis, 1961). The Table of Contents was jettisoned as unhelpful and the index as unnecessary. There are now many different works all labelled "Thesaurus" or even "Roget's Thesaurus". Roget (1962) is a legitimate descendant, but others are very different. The range shows at

once how variously the basic idea of concept characterisation through classes can be interpreted, and how robust this idea is. Thus while there may be no explicit relations between or ordering on classes, these are implicit in the presence of common words in different classes.

As noted earlier, the innovative modern uses of a thesaurus have been in information retrieval (IR) and in natural language processing (NLP). The Cambridge Language Research Unit (CLRU) recognised its common generic function for these purposes in the 1950s (Masterman, Needham and Sparck Jones, 1959), and sought to give this a more substantive interpretation within a shared formal model for the semantic information processing involved in each, and in a well-founded common procedure for constructing a thesaurus. But since the 1960s there has been a divergence, and while the thesaurus has flourished in IR, it has languished in NLP. The published general-purpose reference thesauri just mentioned have not contributed to either since the first laboratory 'proof of concept' done with the printed Roget. However there appears to be a revival of interest now in the use of thesauri for NLP purposes, and these ordinary thesauri may in particular have a part to play within the general move to exploit machine-readable dictionaries to provide lexical resources for NLP.

THE THESAURUS IN IR

Development

The thesaurus as an indexing and, more importantly, as a retrieval device was a response to the postwar challenge presented by the rapid growth of the specialised scientific non-book literature (Foskett, 1980; Roberts, 1984). The librarian's two traditional tools, **subject heading lists and **subject classifications (cf e.g. LC, 1986; BSI, 1985), were essentially aimed at fixed topic characterisations and fixed topic affiliations, and were implemented in schemes suited to single-unit 'thumbnail' descriptions for whole books. They appeared both too rigid and too coarse to meet the new needs for flexible and refined description. A thesaurus was seen as providing a set of irreducible unit **descriptors which could be freely combined to form complex concept or topic descriptions. The constituents of a subject heading or class specification were **precoordinated, i.e. were related in a permanently fixed way. The constituents of a **document description formed with a thesaurus would, in contrast, be **postcoordinated, i.e. would be freely combined in ad hoc document and **request descriptions. Postcoordinate indexing in particular allowed arbitrary topic specifications at search time, as combinations of thesaurus descriptors, or **terms, in requests could match subsets of those originally assigned to documents in a very flexible way. Conventional subject matching required exact request-document matching for the whole descriptions involved, or at least aimed at controlling partial matching in a systematic way, by restricting it to previously defined relationships like those explicitly embodied in any classificatory links, or implicitly permitted through the syntactic modification of a heading (Lancaster, 1972; Foskett, 1980).

It was recognised that to function effectively, thesaurus descriptors would have to be derived from, or at least be strongly motivated by, the particular

scientific literature for which they were to be used. The literature was regarded not as an incidental expression of the autonomous underlying real state of the world, but as itself embodying that state as defined by published scientific knowledge. The set of terms for a thesaurus would thus be obtained by an examination of the literature of a field. However the crucial point at which the thesaurus approach to indexing and retrieval differed from a simple **uniterm one (Lancaster, 1972) using text keywords, was in the provision of a lead-in vocabulary. Though the different publications in a scientific literature may share common basic concepts these are not always conveyed in the same words. As Luhn noted, descriptors would be represented by a group of similar or closely related words, as in a conventional thesaurus like Roget's (Luhn, 1957). The presence of any of these words in a document or request would justify the assignment of the corresponding descriptor. The thesaurus terms, viewed as descriptive labels, thus had an essential normalising purpose, namely to ensure conceptual matching regardless of the surface variety or ambiguity of natural language (Lancaster, 1972).

A retrieval thesaurus is thus at once an **indexing device designed to facilitate flexible topic characterisation, and a **searching device suited to mechanisation through full or partial matching on term specifications ranging from simple term lists to more complex Boolean expressions.

It was soon found, however, that the view of thesaurus terms as a band of brothers on an equal footing was too simple, especially for a large document collection. It was necessary to allow for hierarchical relations between terms, though these were primarily local and were not seen as a means for integrating all the terms into a single unified scheme. These relations are primarily supports for searching. Though indexing and searching are complementary, the way thesauri are used to specify topics by postcoordination rather than precoordination places more emphasis on their retrieval role. Thus while it is a general rule to index with the most specific terms available, the more general ones ensure matching when used in requests.

Characteristics

A modern thesaurus (cf Foskett, 1980) will therefore consist of a term vocabulary supported by a prescriptive apparatus and amplified by a relational one. The terms themselves may be words or phrases, but will usually have a normal form, e.g. singular rather than plural. The prescriptive apparatus includes scope notes (SN) indicating the meanings of terms, and leads from entry expressions to their appropriate indexing terms, of the form x USE y, e.g. 'teenager USE adolescent' (perhaps with reciprocal y USED FOR x). The relational structure is given by links which for a given term may be to **broader terms (BT), **narrower terms (NT), and **related terms (RT), for example 'boy BT adolescent, man', 'child NT infant', 'child RT pediatrics, pupil'. The first two of these relations cover set relationships, the latter, in practice, a miscellany including e.g. part-whole, cause-effect, action-instrument. The NT/BT relations may be exploited automatically in searching, but RTs are more likely to be offered as options to the searcher.

Current thesauri are typically large, elaborate, institutionalised structures intended for use by professional indexers and searchers and managed by large organisations, e.g. the INIS Thesaurus (IAEA, 1987) and Medical Subject Headings, which despite its name is a thesaurus (NLM, 1990). Both of these have tens of thousands of descriptors. These thesauri have been constructed and applied manually, and manuals and guidelines have been developed for this purpose (e.g. UNESCO, 1973; NLIAC, 1980; Aitchison and Gilchrist, 1987); they are more easy to revise than traditional classifications, but a great deal of effort goes into keeping them up to date as science and technology develop. Within the broad framework just described there are many variants, and a great deal of attention has been paid to presentation, to ensure that the thesaurus content and structure is clearly and fully shown within, for example, the confines of a printed book. Automation has made maintenance, revision and use much easier, but their sheer size and intrinsic complexity makes these thesauri difficult for end-users to understand and exploit.

Multilingual thesauri are very important: as thesaurus descriptors are motivated by the subject matter of a field they are in principle interlingual, so all the terms can have their associated sets of different language equivalents (e.g. Viet, 1973). Multilingual thesauri may be hard to manage in printed form, but automation makes selective operation within one language, or combined operations across languages, much easier.

Machine indexing

Since thesauri are usually supplied with entry words or phrases, it is natural to consider automatic term assignment as a means of reducing document indexing effort: entry words occurring in the full document text or its abstract would justify indexing with their associated terms. However the available entry sets may be too limited, and an entry word occurring in a document need not imply its term is legitimate, since the word may be being used in another sense, or if the term is legitimate, imply that it is important. It is better to start afresh by taking the entire document vocabulary and seeing how words in it are associated with term assignments in documents that have already been indexed. If reliable correlations can be established between words, or word sets, and terms, these can be used to justify future assignment. These text items can thus be taken as entry keywords for the thesaurus terms. Even if manual checking is desirable, assignment proposals can make indexing less effort; but automatic assignment has also been implemented.

Thesaurus construction

It is clear that in constructing a thesaurus, great care has to be taken in establishing both terms and their relations. In Foskett's view (Foskett, 1980), while the index descriptions of documents have many affiliations through their components, the component terms should be exclusive and univocal and should participate in only one hierarchical relation. In his view the best way of obtaining a proper classificatory structure for a thesaurus is to ground it in a **facet analysis of the subject field, so individual terms are all associated with specific facets, e.g. process and product, action and instrument. Aitchison's "thesaurofacet" (Aitchison,

1970) illustrates the advantages of this approach in obtaining a well-organised structure.

It is always difficult to control individual interpretations of language in manual thesaurus construction. However, as the automatic assignment strategies just mentioned suggest, and as the definition of a retrieval thesaurus implies, a thesaurus is essentially derivative, and it should be possible to identify its terms by considering the way natural language words behave in the texts of a subject literature. Thesaurus terms are class concepts representing, or represented by, text words, which are alternative expressions or indicators of the underlying concept. In the strongest case these classes will be synonym sets like those of a conventional thesaurus. This suggests that while the members of a retrieval thesaurus class may be viewed as a set of entry words leading to a preferred common label, the proper view is that the members of a class are directly intersubstitutable.

Automatic techniques

The essential strategy for deriving thesaurus classes from information about the way words behave in text is then as follows. Behaviour is defined in terms of occurrences and coocurrences. Thus two words may be said to behave in the same way if they cooccur frequently with a common partner, for example a and b behave in the same way if they each cooccur with p. This implies they are strongly related, in the limiting case as synonyms in complementary distribution. The resemblance between a and b is strengthened if they also cooccur frequently with, say, q and r as well, and equally other words, say c and d may also cooccur with p, q and r. All of a, b, c and d may thus be treated, as their behaviour is similar, as forming a class. For retrieval purposes they may then be taken as intersubstitutable so that if, for example, a document uses a and a request b, this can be taken as a match just as if both were indexed explicitly with the class concept X via its label "X". The concept label "X" is thus simply a convenience: the concept itself is implicit in, or emergent from, the word class. Taking this line further, moreover, suggests that as a and p tend to cooccur, they too can be used intersubstitutably for retrieval purposes, though the relation between them is syntagmatic rather than paradigmatic: habitual collocates are equally legitimate representatives of the underlying concept.

Thesaurus construction from cooccurrence information is formalised, and is thus potentially mechanised, through the application of statistical distribution measures. In general these are used first to define the similarity between pairs of words and then the similarities required among a set of words for these to form a class, but there are many specific ways of doing this. For example, the similarity Sij between a pair of words i,j may be defined as Cij/Oi + Oj - Cij, where Cij is the sum of i and j's cooccurrences and Oi and Oj are the sums of their respective occurrences. An acceptable class of words may then, for instance, be defined as a set of words with high internal and low external similarities, or at least stronger internal than external similarities, which can be established by minimising the cohesion across the boundary between a class and the rest of the objects. Thus if Saa and Sab are respectively the sums of similarities between the members of a putative class a and between the members and the non-members, and Na is the number of objects in

a, we may seek to minimise the cohesion function

$$(Sab/Saa) \cdot ((Na2 - Na)/Saa)$$

(Sparck Jones, 1971b).

The precise nature of a thesaurus is determined on the one hand by the formal definitions on which it is based (van Rijsbergen, 1979), and also by the algorithms actually used to find classes since, for example, it is almost always impracticable to seek all legitimate classes by testing all subsets of the universe of objects. It is determined on the other hand by the nature of the units being grouped, and by that of the text context acting as the frame within which occurrences and cooccurrences are counted.

Thus in relation to the definitions used, similarity as just defined will give a high value to directly rather than indirectly cooccurring words. But the class definition just given is an undemanding one which does not require that every member is strongly related to every other, and so can pick up indirectly cooccurring words; and it is also a definition which allows words to appear in different classes. These relatively loose and overlapping classes, reflecting text variety as well as text regularity, again appear suited to retrieval, which has to be tolerant. In relation to the data elements, these may be word forms, or stems, or multiword strings, and they may include all or only some of the available vocabulary: stems are often used in practice, to increase matching, and very frequent words are excluded, to inhibit matching. The occurrence context may be explicitly syntactically or simply locationally defined, and narrowly or broadly, e.g. as adjective(s) plus noun or as an entire abstract. Contexts are usually defined locationally, but not just because automatic parsers are lacking, and also broadly, to obtain enough statistically reliable data.

Thus taking method and data definitions together, the view in initial work in automatic classification was that what was required were topic classes which would be much broader than synonym sets and could include both paradigmatically and syntagmatically related words. They would thus encompass both the BT/NT and the RT relationships of manual thesauri, reflecting the view that the primary function of a thesaurus is to promote recall, the retrieval of relevant documents. Sense selection and, more generally, discrimination in matching designed to promote precision, by excluding non-relevant documents, would be achieved by postcoordination.

Retrieval performance

Thesaurus construction was an early application of general-purpose automatic classification methods, and research in the area was concerned with many matters of detail within the broad framework just outlined (van Rijsbergen, 1979; Sparck Jones, 1971a). Unfortunately, the experiments done, while they served to emphasise the complexity of IR, never established significant performance gains for automatic classification (Sparck Jones, 1981 Chapter 12; Salton, 1975). It was possible to show that some specific approaches worked better than others, notably those confining classes to strongly related keywords, and these findings have motivated strategies for

enhancing simple keyword retrieval by e.g. working with statistically related term pairs as compound terms (Salton and McGill, 1983). But if automatic classification could not be shown to be of material value, there was equally no hard evidence in favour of manually constructed and applied thesauri as opposed simply to postcoordinate keyword or phrase matching (Cleverdon, 1977; Salton, 1986). Though some investigations have been made, notably by Cleverdon (1977), operations on the scale of major IR services have never been rigorously evaluated through properly controlled comparisons, designed to establish the relative merits of different indexing languages irrespective of the many other factors involved in indexing and searching. Thus it is not clear, for example, whether a conventional thesaurus makes other factors more or less critical, especially as collection scale increases and in the context of on-line searching. The use of a retrieval thesaurus as an indexing device thus remains an act of faith, and while modern technology may make displays of conventional, or unconventional associative, thesaurus structures available for search formulation, how valuable this would be still needs serious investigation.

THE THESAURUS IN NLP

Semantic processing

Thesaurus classes in IR define generic indexing descriptors, and the thesaurus has been envisaged as having a similar function in NLP, that of providing **semantic primitives. This connection was explicitly recognised in early work on the use of a thesaurus for NLP, and specifically for machine translation (MT). Early workers on MT, attempting general text translation, say of research papers, were immediately faced with the problem of word sense identification and, wherever direct links between input senses and their output equivalents were not, or could not be, provided, with the problem of output word (sense) selection as well.

A thesaurus was seen as providing a set of general-purpose, domain-independent semantic primitives allowing the specification, or at least an indication, of the essential semantic concepts expressed by, or relating to, a text, and hence allowing sense determination.

The initial simple model proposed by the CLRU treated thesaurus classes (and for experimental purposes those in Roget's Thesaurus) as characterising, and hence distinguishing, the senses of words; and it treated text as necessarily semantically repetitive, since it was only by the repetition of the generic concepts represented by the heads of classification that the relevant specific senses of the words in a text could be identified. As natural language words normally have multiple senses, context, in this case established via head repetition, served to achieve correct lexical determination in text interpretation. An analogous model would work for text generation: the words shared by the heads selected for an input text would be the right output lexical items. The model in particular allowed for the fact that concept repetition is needed for sense determination, not only because lexical repetition cannot be guaranteed even for the straightforward case, since there may be synonymic variation, but because it cannot be expected for syntactically different items.

Some very rudimentary MT experiments applying these ideas were successfully performed (Masterman and others, 1986, originally 1957). These exploited Roget's Thesaurus supplemented for other languages, and thus used the Thesaurus as an interlingua. More particularly, though concept repetitions to resolve ambiguity could not always be found when searching was limited to Roget's main head set, the hierarchical Table of Contents could be used to extend searching: shared higher-level, broad concepts could be found. Roget's Thesaurus was thus functioning not only as an interlingua but as a terminological knowledge base defined by class inclusion, with the subclasses within heads available if necessary as defining narrower concepts. This approach to the semantic operations required for NLP was seductively simple, and could also, it was claimed, be given a formal underpinning by lattice theory: a lattice model would represent concepts defined by equivalence classes of word senses as nodes, and semantic processes manipulating these would be implemented as lattice operations. Moreover, as with the IR thesaurus, an NLP thesaurus could in principle be built up using distributional information, thus guaranteeing objectivity for the language in question; and as with an IR thesaurus, it would be possible to form a (not necessarily exclusive) hierarchy by grouping classes.

Thesaurus formation

Indeed in practice this derivation, while notionally directly based on a corpus of running text, could be more conveniently and realistically based directly on existing conventional dictionaries. These could be viewed as indirectly based on text, but as condensing the relevant information and providing it either in the already processed form represented by the synonym and quasi-synonym definitions which are frequently encountered, or in the readily exploitable form, suggesting related terms and supplying testing contexts, represented by descriptive definitions and their amplifying examples. Existing dictionaries, assuming they were well-founded, could thus be taken as giving a helpful manual starting boost to automatic classification. The work reported in Sparck Jones (1986, originally 1964) represented a serious attempt to place automatic thesaurus construction for NLP purposes on a firm foundation, bootstrapping classification from minimal information about contextually subsitutable word senses. The actual experiments carried out took lexical data obtained from a dictionary and constituting rows as a starting point, and then applied the clustering techniques already used for IR thesaurus construction. These tests were limited in scope, but successfully obtained Roget-style classes.

Crucial issues

The basic model for thesaurus use in semantic processing underlying this early research was nevertheless far too simple. This became apparent when more serious tests were attempted. The model assumed a one-one correspondence between word senses and heads, but with heads representing broad concepts this might not be discriminating enough on senses; at the same time it assumed that at the head level, the single concept fully characterised whatever the dependent senses had in common. This raises the fundamental question as to whether a thesaurus, and hence the set of semantic primitives it embodies, is a descriptive, indeed definitional, device or is just a selective processing device. The information required for the latter may not be the

same as, and may be less than, that required for the former. The problem is very clearly seen when, in an attempt to ensure repetition, more general concepts are used, since with increasing generalisation the differences between word senses are reduced.

But this is not satisfactory from the processing, as well as from the descriptive, point of view: the variety of text contexts in which words are actually used places them in very different lights and correspondingly implies different selectional perspectives, so effective processing seems to require the richer characterisation that the definitional view also seems to need. However it is in turn evident that it is difficult to capture the meaning of a word sense simply by giving an unordered list of relevant concepts. This is most clearly seen with verbs when their meaning is unpacked, since this invokes **case relationships. It then follows that the notion of sense characterisation by a simple set of concept labels has to be replaced by that of characterisation by a syntactically structured expression in a primitive language. This approach, which may be viewed as a fully developed application for NLP of the uses of defining basic vocabularies and entry formats made in some conventional dictionaries (e.g. the Longman Dictionary of Contemporary English (LDOCE), Procter, 1978), has been advocated and investigated by Wilks (1975, 1977), and by Schank in his Conceptual Dependency Theory (Schank, 1975).

The second major problem area is that conceptual repetition is too crude, and that is therefore necessary to use semantic patterns of the kind exemplified in minimal form by conventional selection restrictions or embodied in semantic grammars. This has immediate implications for the definition of the context for semantic processing. The original simple model did not use syntactic structure to constrain matching, but the more complex requirements represented by patterns, notably those focused on verbs, imply the use of syntactic structure constraints. But these cannot be too rigid, and more generally, as Wilks' notion of preference recognises, semantic pattern matching cannot be made absolute, or the novelty which is as characteristic of discourse as its regularity will ensure they do not work (Wilks, 1975).

The issues raised by initial work with thesauri are thus fundamental to NLP: they concern the nature of text meaning representations, and the nature of the knowledge bases required to obtain and exploit these. If semantic primitives have a key role in both, what is the status of these primitives: are they concept labels in some autonomous mental language, or are they necessarily natural language objects (cf Wierzbicka, 1972; Wilks, 1977; Jackendoff, 1983)? Again, how far do the functions of a thesaurus in discourse interpretation extend beyond sense selection to, say, anaphor resolution, or in response to discourse, as in question answering, to the support of inference?

THESAURUS PROPERTIES

The essential feature of a thesaurus is that it is a classification. It may thus be distinguished from a lexicon and a dictionary: from a semantic point of view a

thesaurus organises terms where a dictionary defines them (Scott, 1988), though it is evident that the relationship between these two is an intimate one. A thesaurus as a semantic classification deals with word meanings, which refer to the world, but it does this from a primarily linguistic point of view as a **terminological knowledge base, complementing the direct description of the world embodied in the **assertional knowledge base (Brachman, Fikes and Levesque, 1985). A thesaurus thus characterises word senses through their relations with others.

The elements of the thesaurus are words (or rather word senses), which characterise one another relationally at different levels of granularity, but without any difference of kind in the levels (Wilks, 1978). Semantic primitives, though they may be thought of as undefined ground elements serving as the decompositional basis for the whole structure (e.g. Mel'cuk and Polguere, 1987), are more properly thought of as words which de facto play a dominant role in the characterisation of others, as with the basic vocabulary of a dictionary like LDOCE. But insofar as a thesaurus is a taxonomy based on lexical similarity, therefore, saying that all its classes define semantic primitives implies both that the thesaurus has a rich classificatory structure, and that this can be interpreted in the sophisticated way required: it cannot simply be assumed that grouping finer classes into coarser ones will suffice to give decompositional generic primitives like 'cause'. The alternative is to identify words with a specific classificatory behaviour, for example ones with many senses figuring in many classes, as primitives, e.g. 'do': this maintains the view of primitives as natural language objects but allows them a special character reflecting a perceived distinctive functional role.

A thesaurus classification based only on similarity relations has structure; but a thesaurus may have a more complex structure. Conventional manually-constructed IR thesauri, though dominated by set relationships, include others as well, like part-whole or action-instrument, and in particular make these relatively explicit where they were only implicit in the experimental automatically-constructed IR thesauri. They also embody a much wider range of relations than the synonymy-based thesauri used in early NLP research. The NLP model proposed by Sparck Jones (1986) was strictly grounded in synonymy, which was essentially broadened only to likeness. Roget's Thesaurus and other conventional printed thesauri include both subclasses and hence class members related to one another by other relations (for example agency), but these other relations are not treated systematically or fully, and these thesauri are dominated by sense synonymy and similarity. At the same time the hierarchical relationship is rather loosely interpreted, with set relations much more obvious in some cases (e.g. as when "scabbard" and "knapsack" appear under 'Receptacle') than in others (e.g. as when 'Receptacle' appears under 'Existence in space'). In constrast, in the application of the 'grammar' rules for lexical definition used by Wilks (1977) and also those used by Schank (1975), other relations are explicitly involved in the form of cases and qualifiers, but there is no significant hierarchical classification of the primitives. Wilks (1978) however recognises the need for hierarchy and allows for multiple levels of specificity by combining Roget's levels of classification with structured definitions.

It is thus clear that a thesaurus as a serious terminological knowledge base will have to embody a case relational structure as well as a categorial taxonomic one, i.e. allow for relational as well as categorial primitives. These case relationships should cover both the types of facet relationship found in conventional IR thesauri and also, insofar as these are different, linguistic case or predicate-argument type relations. These syntagmatic relations together with the paradigmatic hierarchical ones should allow for fine-grained predication constraints. Thus one may, for example, envisage an associative semantic network structure encompassing frames and embedding similarity relationships. But the base must be implemented using a well-defined formalism allowing discriminating inheritance with, ideally, a proper semantics. A thesaurus with a structure like this can then meet the requirements imposed by its dual role in supporting semantic discourse processing, in sense and structure disambiguation, and in providing information about word meanings which can contribute to the further use of a text meaning representation. The scale and richness of both printed and IR thesauri suggests, however, that this computational thesaurus will not be easy to provide. In particular, it will not be easy to ensure that the thesaurus has the required openness to the lexical variety and subtlety of natural language, and to ensure that it does not make inappropriate ontological commitments.

CURRENT STATE

Thesauri are standard features of operational information, and especially document, retrieval systems. They figure in every form, embodying varying degrees of vocabulary control, and merging on the one hand with subject headings and on the other into subject classification. They remain primarily manual indexing and searching devices. They are in fact substantial instances of terminological knowledge bases, though their underlying semantics may not be fully developed, formally expressed, or explicitly indicated in a manner which would allow direct machine use: much is left to the human user's ability to interpret the natural language expressions they contain.

Some beginning have been made, however, on applying approaches relying on NLP and AI techniques to the design and implementation of this kind of thesaurus, particularly for the purposes of text-based IR. An AI approach to knowledge representation using frames has been (manually) applied to the construction of a medical thesaurus intended to be used in conjunction with NLP applied to request and document text (Evans, 1987); and Fox and others (1988) have exploited an existing machine-readable dictionary to build a thesaurus based on a well-defined semantic network formalism. This was tested with conventional indexing, but could in principle also be used to serve more sophisticated request and document text processing. The CYC work (Guha and Lenat, 1990) also has thesaural elements within its more comprehensive scope.

In recent and current NLP work, the role claimed for the thesaurus is typically filled by some combination of selection restrictions and a **sort, i.e. set, hierarchy, often of a limited and/or domain-specific kind. However more challenging applications, and particularly those involving unrestricted, large-scale text

processing, have led to a new interest in the type of linguistic resource represented by thesauri. This interest has also been stimulated by the revival of MT research, which may also involve wide-ranging text material, by the growth of interest in large-scale text skimming, and, from a rather different direction, by the need to improve transportability and reduce startup costs in implementing application systems, like database front ends, for individual domains.

In all of these cases, the concern is with broad-ranging general-purpose semantic resources and, as these are difficult and expensive to develop, with the use of existing lexical resources like conventional dictionaries, on which a great deal of work has already been done, as starting points. This is one element in the recent growth of research on exploiting machine-readable dictionaries (Boguraev and Briscoe, 1989; Evens, 1989). Some of this work is aimed at building terminological knowledge bases with the varied relational structure represented by IR thesauri, but with a well-founded and fully explicit semantics, and some is restricted to identifying simpler taxonomic and similarity relationships; some involves parsing, and some simply statistical operations (e.g. Fox and others, 1988; Alshawi, 1989; Wilks and others, 1989). Thesauric information of the narrower synonymic kind has also been seen as a necessary tool to underpin and extend dictionary analysis (Byrd and others, 1987). At the same time, proposals for exploiting corpora have been made, for example in sophisticated analysis to generate a complex base (Anick and Pustejovsky, 1990) or, by applying statistical measures to identify semantic associations, to provide an objective underpinning for lexicon development (Church and Hanks, 1990). This statistical work has not, however, so far progressed from the use of pairwise similarity measures to the full-scale automatic grouping attempted in earlier IR research.

Much of this work is based on ideas about thesauri which are not new; but it can exploit NLP techniques for text parsing and AI techniques for knowledge representation which are now available, and can take advantage of machine resources for heavy data processing which did not exist when these ideas were first put forward. There is, however, a manifest need for more work on applying automatic classification methods that already exist to this field, and on developing relevant new methods. It is also the case that significant advances in NLP are required before detailed specifications for broad-ranging or general-purpose thesauri can be provided, as these depend in turn on detailed specifications for their modes of use which have themselves to be provided, and which can only be operationally established.

BIBLIOGRAPHY

J. Aitchison, 'The Thesaurofacet', Journal of Documentation 26, 1970, 187-202.

J. Aitchison and A. Gilchrist, Thesaurus Construction: A Practical Manual, 2nd ed, Aslib, London, 1987.

H. Alshawi, 'Analysing the Dictionary Definitions', in Boguraev and Briscoe, 1989.

P. Anick and J. Pustejovsky, 'An Application of Lexical Semantics to Knowledge Acquisition from Corpora', COLING-90: 13th International Conference on Computational Linguistics, 7-12, 1990.

B. Boguraev and E. Briscoe (Eds), Computational Lexicography for Natural Language Processing, Longman, Harlow, Essex, 1989.

R.J. Brachman, R.E. Fikes and H.J. Levesque, 'KRYPTON: A Functional Approach to Knowledge Representation', in Readings in Knowledge Representation, Ed R.J. Brachman and H.J. Levesque, Morgan Kaufmann, Los Altos CA, 1985.

BSI: British Standards Institute, Universal Decimal Classification, FID 571, British Standards Institute, London, 1985.

R.J. Byrd and others, 'Tools and Methods for Computational Linguistics, Computational Linguistics 13, 219-240, 1987.

K.W. Church and P. Hanks, 'Word Association Norms, Mutual information, and Lexicography', Computational Linguistics 16, 22-29, 1990.

C.W. Cleverdon, "A Computer Evaluation of Searching by Controlled Language and Natural Language in an Experimental NASA Data Base", Report ESA 1/432, European Space Agency, Frascati, Italy, 1977.

E.D. Dym (Ed), Subject and Information Analysis, Marcel Dekker, New York, 1985.

D.A. Evans, 'Final Report on the MEDSORT-II Project: Developing and Managing Medical Thesauri', Report CMU-LCL-87-3, Department of Philosophy, Carnegie Mellon University, 1987.

M. Evens, 'Computer-Readable Dictionaries', Annual Review of Information Science and Technology, Vol 24, Ed M.E. Williams, Elsevier, New York, 85-117, 1989.

D.J. Foskett, "Thesaurus", Encyclopedia of Library and Information Science, Vol 30, Ed A. Kent, H. Lancour and J.E. Daily, Marcel Dekker, New York, c1980; reprinted in Dym, 1985.

E.A. Fox and others, 'Building a Large Thesaurus for Information Retrieval', Proceedings of the Second Conference on Applied Natural Language Processing, Association for Computational Linguistics, 101-108, 1988.

R.V. Guha and D.B. Lenat, Building Large Knowledge Based Systems, Addison-Wesley, Reading MA, 1990.

IAEA: International Atomic Energy Authority, INIS: Thesaurus, International Atomic Energy Agency, Vienna, 1987.

R. Jackendoff, Semantics and Cognition, MIT Press, Cambridge MA, 1983.

J. Knowlson, Universal Language Schemes in England and France, 1600-1800, University of Toronto Press, Toronto, 1975.

F.W. Lancaster, Vocabulary Control for Information Retrieval, Information Resources Press, Washington DC, 1972.

A. Large, The Artificial Language Movement, Blackwell, Oxford, 1985.

LC: Library of Congress, Subject Headings, 10th ed, Library of Congress, Washington DC, 1986.

N. Lewis (Ed), The New Pocket Roget's Thesaurus in Dictionary Form, Washington Square Press, New York, 1961.

H.P. Luhn, "A Statistical Approach to Mechanised Encoding and Searching of Literary Information", IBM Journal of Research and Development 1, 309-317, 1957.

M. Masterman and others, "Agricola Incurvo Terram Dimovit Aratro", Cambridge Language Research Unit, 1957; reprinted with Introduction by K. Sparck Jones, Computer Laboratory, University of Cambridge, 1986.

M. Masterman, R.M. Needham and K. Sparck Jones, "The Analogy between Mechanical Translation and Library Retrieval", Proceedings of the International Conference on Scientific Information, National Academy of Sciences, Washington DC, 917-935, 1959.

I. Mel'cuk and A. Polguere, 'A Formal Lexicon in Meaning-Text Theory (Or How to Do Lexica with Words)', Computational Linguistics 13, 261-275, 1987.

NLIAC: National Library and Information Associations Council, Guidelines for Thesaurus Structure, Construction, and Use American National Standards Institute, New York, 1990.

NLM: National Library of Medicine, Medical Subject Headings 1991, 3 vols, National Library of Medicine, Bethesda MD, 1990.

P. Procter (Ed), Longman Dictionary of Contemporary English, Longman, Harlow, England, 1978.

C.J. van Rijsbergen, Information Retrieval, 2nd ed, Butterworths, London, 1979.

N. Roberts, "The Prehistory of the Information Retrieval Thesaurus", Journal of Documentation 40, 271-285, 1984.

P.M. Roget, A Thesaurus of English Words and Phrases, Longman, London, 1852; revised edition, Roget's Thesaurus of English Words and Phrases, Ed R.A. Dutch, Longman, London, 1962.

G. Salton, Dynamic Information and Library Processing, Prentice-Hall, Englewood Cliffs NJ, 1975.

G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

G. Salton, "Another Look at Automatic Text Retrieval Systems", Communications of the ACM 19, 648-656, 1986.

R.C. Schank, Conceptual Information Processing, North-Holland, Amsterdam, 1975.

D.S. Scott, 'Capturing Knowledge with Data Structures', in Data and Knowledge (DS-2), Ed R.A. Meersman and A.C. Sernadas, North-Holland, Amsterdam, 1988.

M.M. Slaughter, Universal Languages and Scientific Taxonomy, Cambridge University Press, Cambridge, 1982.

K. Sparck Jones, Automatic Keyword Classification for Information Retrieval, Butterworths, London, 1971. (1971a)

K. Sparck Jones, "The Theory of Clumps", Encyclopedia of Library and Information Science, Vol 5, Ed A. Kent and H. Lancour, Marcel Dekker, New York, 1971; reprinted in Dym, 1985. (1971b)

K. Sparck Jones, "Some Thesauric History", Aslib Proceedings 24, 408-411, 1972.

K. Sparck Jones (Ed), Information Retrieval Experiment, Butterworths, London, 1981.

K. Sparck Jones, Synonymy and Semantic Classification (thesis 1964), Edinburgh University Press, Edinburgh, 1986.

UNESCO, UNISIST Guidelines for the Establishment and Development of Monolougal Thesauri, SC/WS/555, UNESCO, Paris, 1973.

J. Viet, EUDISED Multilingual Thesaurus, Mouton, Paris, 1973.

A. Wierzbicka, Semantic Primitives, Athenaeum Verlag, Frankfurt, 1972.

Y.A. Wilks, "An Intelligent Analyser and Understander of English", Communications of the ACM 18, 264-274, 1975.

Y.A. Wilks, "Good and Bad Arguments about Semantic Primitives", Communication and Cognition 10, 181-221, 1977.

Y.A. Wilks, "Making Preferences More Active", Artificial Intelligence 11, 197-223, 1978.

Y.A. Wilks and others, 'A Tractable Machine Dictionary as a Resource for Computational Semantics', in Boguraev and Briscoe, 1989.