IV S E M A N T I C S

IV . 1 Statistical semantics

One distinction applies at all stages of semantic processing. This is whether the information and procedures involved are or are not statistically based. It has been claimed by some linguists that verbal meaning is distributionally determined, and hence is identifiable by statistical means. However few attempts have been made to pursue these ideas in general semantic studies, because of the enormous amount of data which in principle needs processing to obtain sufficiently rich and reliable information. Some limited attempts were made, for example, by Sparck Jones 1964. However it was early suggested that automatic indexing could make use of distributional information. This was in part faut de mieux. Since ways of identifying concepts or conceptual relations by non-statistical means were not obviously available, statistical techniques were proposed instead, particularly since computers might well be able to conduct the tedious counting operations required with less effort than human beings. But it was also argued, following linguistic advocates, that distributional information has positive merits, since it necessarily reflects the use of words in the texts from which it is derived. The assumption is that a posteriori semantic characterisation of documents is more reliable than a priori. It was also argued that statistical approaches could be more realistic in retrieval contexts than in general, because a very refined treatment of semantic information might not be required. Thus selecting some frequent words from a document might provide good enough content keys, and generating loose topic groupings of keywords from text cooccurrences, as opposed to pure synonym sets, might be acceptable.

Under each heading in this section, therefore, approaches will be distinguished according to whether they are statistical or not. For example, in document analysis, words may be selected as content clues by statistical or non-statistical means.

IV . 2 Input semantics

The object of this stage of document processing is to identify the content bearing words of the document or surrogate text.

Clearly, procedures storing the full texts of documents or abstracts as in, say, legal text searching systems, represent the lowest or null level of input analysis. Titles or abstracts are used by some operational mechanised systems, as indicated in Section Vl.

Syntactic criteria for selecting words have already been discussed. Non-syntactic methods involve either reference to a dictionary or the application of statistical criteria. Reference to a positive dictionary of terms to be selected is more properly considered in the next section. But it is reasonable to include reference to a negative dictionary of words to be excluded under the heading of input processing. A negative dictionary or 'stop list' is primarily a device for excluding words like

prepositions or conjunctions which are not directly semantically informative; but it may also include 'fluff' content words like "description", "method" and the like, or adverbs like "really". Referring to a stop list when documents are input for processing is practically advantageous, since it can reduce the bulk of abstract material, say, by 30%. Further selection of content words may of course be intended.

IV.2.1 Statistical extraction

If syntactic criteria are not used to select words, the only obvious alternative is to use statistical ones. Statistical extraction techniques are most obviously suited to full texts where sufficient information about word frequencies is available. Such techniques were first advocated for automatic extracting by Luhn (Schultz 1968), and their application to indexing naturally follows. Early work in this area is summarised in Stevens 1965a, and also Borko 1967 and Salton 1970a, and is represented by papers in Stevens 1965b.

The basic idea is quite simple: words are ranked by frequency in some way, and a threshold is applied to select more frequent words. It may not be appropriate to select the most frequent words: if an entire document text is input, stop list type words like articles are most frequent. These are clearly not wanted as content indicators. In this case two thresholds are required, an upper one to exclude the most frequent words, and a lower one to eliminate the least frequent. In practice, it may be convenient to remove stop words by list reference, so the main problem is choosing the lower threshold. The difficulty is that words above this threshold may not be very informative, for example if they are general words like "method", while those below may be valuable, but occur only rarely because they are synonyms of other more frequent words, for instance. It seems in any case to be desirable that frequency counts should be based on stems rather than actual word forms, as noted by Carroll 1970.

The essential problems of statistical extraction are thus a) the definition of frequency, and b) the selection of the threshold. The simplest definition of frequency is just the occurrence count of a term in a document. But this is not discriminating for two reasons: it does not take account of variable document length, or of variable gross language (or collection) frequency. What is really wanted is a formula which picks up the distinctive frequency of a term in a document.

Let f_{ij} be the occurrence frequency of term i in document j, f_i the total number of occurrences of term i in the collection, p_j the total number of term occurrences in document j, and N the total number of term occurrences in the collection. We then define

$$F_{ij} = f_{ij}/p_j \text{ and } R_i = f_i/N.$$

Edmundson and Wyllys early work (1961) studied three frequency criteria defining the value v_{ij} of a word i for a document j:

1. $v_{ij} = F_{ij} - R_i$

2. $v_{ij} = F_{ij}/(F_{ij} + R_i)$

3. $v_{ij} = F_{ij}/R_i$

More complex criteria can clearly be suggested.

A number of experiments were carried out in the mid sixties both to compare statistically and manually extracted term lists, and to study alternative bases for automatic selection. It should be noted that similar criteria may be used to select individual words from an input text, and to select an indexing vocabulary for a collection. The latter will be discussed later.

Damereau's 1965 experiments compared the three functions just given and a fourth one related to the Poisson distribution for word stems in 7 articles, using information derived from a separate sample of 1 million words to provide standard frequencies. Manually compiled lists were used for evaluation. Cutoff was determined in a rather complicated way not obviously extensible for ordinary use. The results showed that lists produced using the fourth formula agreed best with the manual ones. However, the agreement was not very good. There was no attempt to evaluate the lists for retrieval.

Carroll 1969 describes a similar study with 19 documents containing over 66000 words altogether. He compares a number of criteria like Edmunson and Wyllys and Damereau's, again evaluating by reference to manual lists. An arbitrary cutoff point seems to have been chosen. Calculating average rank correlation coefficients showed that in this case the simple word count f_{ij} worked best. There was again no performance evaluation.

IV.2.2 Conclusion on input semantic analysis

It is difficult to comment on these very limited experiments in selecting words as content indicators for documents. Evaluation by comparison with manual lists is of limited value since comparable variation between human indexers occurs and it is not evident that this influences retrieval performance materially. There is, however, an intrinsic difficulty associated with the use of some normalising coefficients which seek to determine word selection by overall collection frequency. If a collection changed character markedly, words rejected for early documents might have been selected later. It is worth noticing that the techniques suggested by Jones 1969 for obtaining an indexing vocabulary statistically take account of this, by having a backup store of all text words, so documents may be reindexed. However this is not realistic for large collections. Some term weighting schemes offer a solution to the comparable problem which arises when smaller texts like abstracts are retained.

Carroll and Damereau's experiments do not, however, suggest that any one selection criterion is of special merit, so frequency criteria which did not refer to an overall collection might be adequate.

The general conclusion for statistical extraction from full text must be that we cannot reject it, but that there is no good evidence for adopting it in the face of its cost. If it could be established that indexing from full text gave a much better performance than indexing from abstracts, it might be worth looking again at automatic techniques for extracting words from text. The experiments reported here throw no light on this question. Those by Salton 1968a mentioned in Section II are not strictly comparable: they involved 82 short document texts, and terms were not absolutely selected or rejected, only relatively up or downgraded by weighting.

A rather different statistical approach is adopted by Cagan 1970. Most of those working with statistical extraction techniques have been concerned to pick up relatively frequent terms. Even if the most frequent terms are unhelpful, terms with middling frequencies seem of most general value in retrieval. Cagan emphasises the importance of rare terms for medical literature. His choice of index terms for a document is based on statistical associations. The linkage between pairs of terms in the collection is calculated using a coefficient designed to favour infrequent terms. (It should be mentioned that association between two terms is not defined by direct cooccurrence, but by cooccurrence with other words.) All terms in a document with a linkage above a threshold to the other terms in a document are flagged as index terms. The threshold was in fact set so high, at .8, that something like 90% of the index terms are words occurring only once in the collection. The experiments were carried out with 250 tropical disease abstracts, and 31 requests, apparently single word ones, were used for retrieval evaluation. It is difficult to determine how good the index term selection really was, since the account of the evaluation is obscure. 88.5% recall and 93.2% precision were apparently achieved for highly relevant documents, but these striking figures may be due to the interpretation of relevance and type of request.

IV . 3 Description semantics

Under input semantics I considered only the selection of content indicators from input texts. In description these are replaced by index terms. In some cases the replacement is trivial: the input text words are simply adopted as index terms without change. However the logical distinction is useful because more material substitution usually takes place. This may range from substituting stems for full words, to substituting thesaurus terms or classificatory headings. In such cases the input words are entry words to the indexing vocabulary.

IV.3.1 Derivative and assignment indexing

Stevens 1965a distinguishes derivative and assignment indexing: in derivative indexing the descriptors for a document are taken from the document itself, or come from a set of descriptors drawn from a set of documents including or like the given one. Using extracted keywords as descriptors represents the first, and class labels standing for sets of keywords grouped by statistical association techniques illustrates the second. In assignment indexing the descriptor set is essentially independent of the documents, and words extracted from the document do no more than lead to the assignment of descriptors. In general, we would expect that fully automatic indexing systems would be derivative in character. The point of interest is the extent to which the provision of descriptors imposes constraints on the initial extracted information.

In the actual formation of descriptions, a distinction can be made according to whether any reference is made to a descriptor list like a dictionary, or classification, or not. The object may be either to vet content indicators for indexing utility, or to replace them by items of a different form. This may perfectly well occur in derivative indexing, say if a vocabulary has been formed by statistical operations on all the words from a set of documents, or a classification generated from keywords.

The formation of descriptions is very simple if there is no descriptor list (except post facto, represented by the set of extracted keywords). In some cases no modification of the extracted words is undertaken, in others word forms are replaced by stems. (Indirect stemming imposed by request stemming is a separate matter).

IV.3.2 Keyword normalisation

It is generally accepted that some form of word normalisation is convenient or helpful, that is that words extracted from documents should be subjected to affix, and especially suffix, stripping. This seems appropriate whether the resulting stems are adopted as index terms as they stand or are subjected to further processing to generate, say, a classification. A good deal of work has been put into automatic suffix stripping in the general area of computational linguistics, and the appropriate techniques and their limitations are fairly well understood.

Proper suffix stripping requires a dictionary both of stems and of suffixes: this ensures the formally correct division of a word into a valid stem and valid affix, and prevents, for example, the selection of "twic-" as a stem for "twice". But it does not prevent inappropriate divisions, as of "many" into "man-y". If a full word dictionary was provided with instructions for the correct division of each word, some such problems would be avoided. But quite apart from the problem of providing this for highly inflected languages (not to mention those permitting free combination), homonyms with different internal structures like the German "wachtraum" ("wacht-raum"/"wach-traum") can occur, for which contextual checking is needed. In documentary applications this may not be feasible because the relevant information is not available (for example if titles or manual keyword lists are being processed), and it is in any case difficult and expensive. A suffix dictionary is typically quite small, but a stem dictionary may be large, and an effort to provide.

A suffix dictionary alone will inevitably generate some mistaken stems, but the proportion of errors may not be large. For example, for a sample of 5500 words Andrew's 1971 procedure generated about 5% errors. In any case, in retrieval, the important point is the consequences of mistakes. Stripping means that different words will be conflated through having common stems, so in some cases they may be wrongly conflated. (Improper stems may be generated, say for proper names, which do not result in false conflation: this does not matter.) False conflations may, however, have little effect on retrieval performance, since requirements for joint matches on several terms may eliminate them.

Lovins 1968, 1971 considers stemming errors in connection with Project Intrex. She divides them into two classes: understemming and overstemming errors. In the first case, a rather restricted suffix list containing, for example "-y" but not "-ily" is likely to lead to failure to identify common stems, and hence to missed matches in searching. In the second, a more generous list leads to pseudo-stems and hence false matches in searching. It is unfortunately a matter of swings and roundabouts, and particular choices may be appropriate in particular circumstances: this applies, for instance, to the use of "-or" as a suffix. (Chemical nomenclature must be treated as a special case.) Lovins attempts to estimate the amount of error in a test word list mathematically, and finds a 4% understemming error and 1% overstemming, which is encouraging.

A number of experiments designed to compare the retrieval effects of using unprocessed words, or word forms, and stems have been carried out, with both manually and automatically processed vocabularies. Manual form grouping is usually more discriminating than automatic, but this may not influence retrieval performance much. Cleverdon 1966 found no noticeable difference in performance for 221x1400 aerodynamics documents. Salton 1968a,c investigated automatic stemming, in fact using a stem as well as suffix dictionary. The word forms with which the stems were compared were themselves trivially processed to remove final "-s". The comparisons, for the 42x200 Cranfield, 34x780 IRE and 35x82 ADI collections showed no noticeable difference in performance.

In general, however, quite apart from practical convenience, we should expect suffix stripping to be of increasing value as collection size grows. This is suggested by the results in Salton 1968d where stem performance for 48x1268 documentation abstracts is materially, and indeed strikingly, superior to that for word forms.

IV.3.3 Dictionary reference

When reference is made to a term dictionary, the process is usually quite simple: the extracted words form an entry vocabulary and the associated descriptors are assigned to the document. This is a common manual indexing procedure (see Lancaster 1968b), and appears in mechanised systems (see, for example, Clough 1971). Mechanised lookup of this sort does not constitute automatic indexing. However attempts have been made to automate a more sophisticated assignment procedure, involving some controls: an example would be assigning descriptor D for word w_1 if word w_2 was also present. Fangmeyer 1969, 1970 has studied techniques for controlling assignment from the Euratom thesaurus, involving both relations

between different text words, like morphological ones, and between text words and thesaurus descriptors, reflecting the past correlation of a word and a descriptor. Chains linking words and descriptors can therefore be established, and descriptors are assigned, for example, according to the number of linked word elements present in an abstract. Retrieval tests examining assignment criteria of different strengths for 20x529 nuclear science abstracts showed that when moderately restrictive criteria were applied, performance was as good as for manual assignment.

IV.3.4 Indexing for substitution and addition

Dictionary reference may lead to the assignment of one descriptor for several entry words, or vice versa. More generally, the resources of an indexing language may be exploited to provide index terms for a document description in two different ways. The language may be intended to allow verbal substitution for matching: thus different input words may be replaced by the same class name. Descriptors used this way will have a normalising function, or act as recall devices. Alternatively, the language may be intended to allow verbal elaboration for matching: input words may be replaced by a set of index words. Descriptors used in this way have a specifying function, that is, act as precision devices. The two approaches to providing a description are easily illustrated with (automatic) keyword classifications. In the first case, any keyword in the class is replaced by the class name, allowing indirect keyword matching; in the second, any keyword is replaced by all the others, allowing joint keyword matching. Unfortunately the literature is very confusing here, since the word "expansion" is used to cover both enterprises. I shall use "substitution" and "addition" respectively.

Substitution is of course a familiar search device, but it is useful to consider both it and addition as applying to document indexing as well. It should be pointed out that indexing substitution of the type just mentioned must be applied to both document and request, for input words are replaced by descriptors of a different type. Substitution of one descriptor for another in a request is slightly different. Addition may be applied to either requests or documents or both.

In general the choice between substitution and addition in indexing is forced by the character of the index language itself. But some types of language, like keyword classifications, offer either alternative, so a proper comparison with respect to their retrieval value, either against one another or against unmodified keyword descriptions, can be attempted. Lesk 1969 comments that statistically associated terms when added to document descriptions functioned more as precision aids than as the alternative substitution possibilities they were originally intended to represent. Some limited experiments by Sparck Jones 1971b attempted to compare substitution and addition for the 42x200 Cranfield collection. Simple addition appeared to give as good performance as substitution, with much less effort. But the results are difficult to evaluate as only exhaustive analysis of individual requests would show whether performance was really determined by combined or substitute matches as

opposed to the mixture which is allowed by both expansion and substitution through a heavily overlapping classification. Vaswani 1970 compared a variety of substitution and addition strategies with 93x11571 mixed subject abstracts and found the best strategies of each kind and gave the same performance.

Other experiments with statistical associations have investigated the effects of addition on either documents or requests or both. Sparck Jones 1971b, Vaswani 1970, and Dennis 1967 have made various comparisons, the latter with two collections of 61x5121 and 6x556 legal documents. The results do not support any solid conclusions: for example Dennis' findings with the two collections, the smaller a subset of the larger, differed. Other experiments of my own suggest that enlarging both requests and documents may be too undiscriminating.

IV.3.5 Description exhaustivity

This has already been mentioned in connection with the form of input to indexing. It also obviously matters in indexing for addition. Some experiments have been carried out to examine the effects of different levels of exhaustivity in indexing derived from a constant source. In the Cranfield experiments and subsequently, Cleverdon 1966, 1970 investigated different levels for manual indexing. The earlier tests with three quite different levels of indexing showed no performance difference for either the 221x1400 or 42x200 collections. The later experiments with 14x237 defence abstracts showed very exhaustive indexing of both requests and documents was not very competitive. These tests emphasised the fact that exhaustivity of document indexing needs to be related to request exhaustivity, which is normally treated as independent, in searching. Sparck Jones 1973d compared different levels of exhaustivity for both requests and documents for the 42x200 Cranfield collection and 47x407 documentation abstracts. The tests suggested that different levels of exhaustivity in document indexing could be counteracted by compensatory treatment of requests (with useful economic consequences). They also suggest that while performance may be affected if very short or very long document descriptions are provided, the optimum level must be very broadly defined.

Some of these results, and those obtained by, for example Aitchison 1970, suggest that the uncritical adoption of entire abstracts (i.e. all the non-function words) leads to very exhaustive indexing which must be consciously counteracted in the treatment of requests. From this point of view Salton's work which is usually based on abstracts, is of interest. His experiments typically seem to show a plausible general level of performance, and it must be presumed that this is due to the use of term weights, and a matching coefficient which takes account of document length. The experiments just described used no weighting and simple coordination level matching.

IV.3.6 Term weighting

In manual indexing, usually on a limited scale, weights may be assigned on an intuitive basis, particularly to request terms. Weights may also be computed automatically, and a variety of experiments in the use of such weights have been carried out. The obvious base is term frequency information. The rationale for different

types of weighting is discussed in Sparck Jones 1973e. Specifically, information about any or all of the within-document frequencies of terms, the length of documents, and the number of term postings, may be used; and actual weights may be generated from this information with a variety of different functions. (It should be noted that the consequences of the three types of information differ: in the first case, the more frequent a term, the higher its weight. The other two operate inversely, so that the longer the document description, or the higher the postings, the lower the weight.)

A fairly simple approach deriving weights on three levels from within-document frequencies was tested by Artandi 1969a,b on 15 drug documents. Comparison with manual weighting showed fairly good agreement, and the highest weighted terms derived from full text tended to be those terms occurring in the document abstracts. However much more substantial experiments by Dennis 1967 with 6x556 legal documents showed that such weights performed noticeably less well in retrieval than simple within-document frequency weights (scaled to allow for document length). Term weights representing within (abstract) text frequencies are normally used in Smart experiments, with normalisation for description length via the search matching coefficient. Comparative experiments reported in Salton 1968a,c with the 34x780 IRE, 42x200 Cranfield and 35x82 ADI collections show that weighted descriptions never perform worse than unweighted ones, though they do not always perform noticeably better. Sparck Jones 1973e, working with 47x407 documentation abstracts, found within-document frequencies gave no improvement, but attributed this to the fact that highly weighted terms probably occurred too rarely to influence retrieval performance. In tests with the 63x797 Keen, 42x200 Cranfield and 97x541 Inspec collections, she did not find that weighting to take account of description length was particularly useful, though it did not degrade performance, presumably because both long and variable descriptions are required before it is of value. On the other hand Sparck Jones 1972, 1973e has found that simple collection frequency based weighting, assigning high weights to rare terms, leads to a material improvement in performance for these three collections. A very similar idea was implemented by J.Williams 1968b in a patent search system involving 17000 abstracts; unfortunately the rather limited evaluation reported does not explicitly show its value. Salton has recently studied, 1972a,c, 1973c, a more complex collection distribution based weighting scheme, as well as that proposed by Sparck Jones, in the context of a general discussion of the values of terms for retrieval. In Salton's scheme terms are weighted by their 'Q-value', which represents the extent to which, by their posting frequency, they relate and distinguish documents. A good term is one whose removal from the collection reduces document separation. This means that both common and rare terms tend to be less useful than medium frequency terms. Comparative experiments with three collections, all approximately 25x450, reported in Salton 1973c, show that the Q function is useful, but not more than simple collection frequency based weighting, which still performs noticeably better than terms not thus weighted.

(p. 4.9a follows)

Weights may also be based on term associations. In Cagan 1970 terms have different weights, or "relevance values" for documents according to the strength of their linkage with the other terms in an abstract. However there is no comparison between weighted and unweighted descriptions. Further possibilities in exploiting term associations have been studied by Vaswani 1970, Sparck Jones 1971b and Lesk 1969. For example when descriptions are expanded using a classification, source terms may be weighted more heavily than added ones. Alternatively, if words are replaced by class names, weights can be assigned to classes according to the number of their sources, and so on. There are many complicated possibilities. But the results of experiments along these lines do not suggest that they have any particular merit.

The weighting techniques described so far are all statistical ones. Hillman's 1968, 1969, 1973 very different approach is syntax based. As noted in Section III, parsing reduces input sentences to canonical components, that is expressions consisting of relationally linked noun phrases. Noun phrases are selected as terms and weighted according to the number of times they appear as relational arguments in the set of components for a text. Unfortunately, as mentioned, there is no performance evidence for Hillman's methods.

The general conclusion from these studies of weighting is that weights can be computed automatically, using relatively simple techniques, which may be quite effective in retrieval, though there are no tests showing how valuable they are for large collections. A point of interest is the effort involved in computing them on different bases. Within-document frequency weights are most conveniently computed on input; collection frequency based weights are naturally computed at the search stage, and very economically.

IV . 4 Index language semantics

The nature of the index language used for describing documents and requests is usually regarded as the most important determinant of retrieval system performance, though the experimental evidence of the last decade suggests that this may not be the case: see Saracevic 1968, 1971, for example. It is, however, true that a substantial volume of literature is concerned with language design. In particular, articles continue to appear on thesaurus construction, as illustrated by J.Aitchison 1972. Lancaster 1972a is a useful recent survey, and for a discussion of language properties see Vickery 1971, for example.

It is convenient to distinguish the indexing language vocabulary from any relational or classificatory structure the language may have. The main controversy of the last ten years has concerned the degree of control embodied in the language: that is, first, whether the vocabulary should consist essentially of keywords, or of terms or subject headings with particular interpretations; and secondly, whether any relations between descriptors should be indicated. The first imposes direct control in indexing, the second indirect control in searching. In general, a controlled vocabulary is associated with some structure, but this need not be the case. Some structures are relatively simple, consisting merely of notes of some relations like BT, NT and RT where appropriate; in other cases all the terms or headings are embodied in a complete classification. Equally, an uncontrolled vocabulary may be given a classificatory structure, so the indexing language incorporates an element of control.

In the present context two points are important. The first is what general evidence there is for the belief that some degree of control is of value for retrieval (in mechanised searching: I am not concerned with visual user aids etc. here). The second is how automatic indexing techniques allow for control if this is desirable.

It is difficult to comment on the first point. A variety of experiments have been carried out comparing manual indexing using uncontrolled or natural language vocabularies and controlled ones consisting either of precoordinate subject headings or postcoordinate thesaurus terms, and varying degrees of classificatory structure. The results obtained by, for example, Aitchison 1970, Cleverdon 1966, the Comparative Systems Laboratory (Saracevic 1968, 1971), and Keen 1972, 1973 suggest that elaborate controls do not pay off, and further that even moderately controlled languages with some vocabulary restrictions or some descriptor relationships, though they may perform better than uncontrolled keyword vocabularies, do not perform strikingly better. The paradox is that the use of relatively controlled languages continues in operating systems, though this may be justified by their much greater scale.

If we accept that moderate control may be useful, how can this be achieved automatically, given the derivative character of automatic indexing? We must consider first control of the indexing vocabulary, and secondly control through the classificatory structure of the language.

In derivative indexing the situation is complicated because vocabulary control may be applied at two levels. If a system involves only keywords, vocabulary control occurs in the obvious sense when the set of keywords which may be used to index documents is restricted in some way. However if the system incorporates a keyword classification, in which class names are used as descriptors, the set of descriptors constitutes a controlled language which may be imposed on top of an initially selected keyword vocabulary. Unfortunately keyword classifications are somewhat confusing because this classificatory information may be exploited in controlling searching, with initial descriptions in keyword form. For convenience I shall restrict the word "vocabulary" to the keyword vocabulary, and treat class descriptors under the discussion of classification.

IV.4.1 Vocabulary

The requirements an indexing vocabulary must satisfy have been considered, for example by Lancaster 1972a. Many of these requirements can be given a statistical interpretation: for example avoiding excessively general words may be equated with avoiding unduly frequent words. The specificity of index language terms can be given a straightforward interpretation in terms of posting frequency. Ideally an indexing vocabulary should consist of terms with comparable frequencies, and enough of them to provide sufficient information about the differences and similarities between documents. Vocabulary requirements are discussed in this way by Salton 1968a, 1973c. Selecting an indexing vocabulary may therefore be treated as a problem of eliminating both very frequent and very rare words from an initial collection of extracted words.

IV.4.1a Statistically extracted vocabularies

As noted earlier, the type of statistical measure which may be used for providing index lists for individual documents may also be used to select a vocabulary for a collection. In the first case the indexing vocabulary for a set of documents is essentially the union of the lists for individual documents. In the second the procedures are used to select a subset of an initial collection of words extracted from documents, say by taking all the non-function words in abstracts, so the documents are eventually indexed using the subset only. There may be an element of assignment in this procedure, if the subset is determined using a collection sample, with assignment of the selected terms to the remainder. Dennis 1965, 1967 indexed a set of 5121 documents with a vocabulary derived from a subset of 2649 documents. Inspection of her retrieval results showed no real performance difference between documents with derived and assigned indexing.

Dennis reports a variety of experiments with different selection criteria. She found the ratio of simple word stem occurrences to the reciprocal of the coefficient of variation of within-document frequencies most helpful, on an intuitive basis. This led to a stem ordering of over 15000 items which could be reduced to an "informing" vocabulary of about 7000 stems by cutting off the top and tail of the list in a fairly straightforward way. In these tests the top included both 'stop' words

and frequent content words, the bottom infrequent words. A variety of retrieval tests were carried out. The analysis of the results for 18x5121 and 6x556 with a rank cutoff of 35-40 documents shows recall of 25-50% and precision of less than 10%. Unfortunately there was no attempt to compare the performance of the automatically selected vocabulary with a manual one.

Stone 1967a,b, 1968 made some similar comparisons between different selection criteria, but using only a small collection of 217 computing reviews, representing 70000 words of text. Very high frequency words were in fact eliminated with a stop list; an ad hoc cutoff to delete low ranking words left a vocabulary of 984. The various criteria studied were evaluated by their ranking of words in relation to manual lists representing distinctions between different types of words of presumed relevance for indexing. Thus speciality words for the particular subject area were distinguished from non-speciality words, and general words from specific ones. Stone's conclusion is that there is no one measure which is of real value for both types of separation, though a Poisson-based function related to Dennis' preferred formula performed quite well.

Jones 1967a investigated an associative technique rather like Caqan's for individual documents. The hypothesis was that words restricted in their environments, i.e. occurring with few other words, are more discriminating than 'dispersed' words occurring with many others. The tests involved the 999 most frequent non-trivial words from 10749 abstracts. The value $v_i = C_i/f_i$, where C_i is the number of different words cooccurring with i, was calculated for all the words. Visual inspection of the results suggested that the most restricted words were good candidates for an indexing vocabulary.

A subsequent A.D.Little report (Jones 1969) mentions the use of statistical extraction techniques to obtain an indexing vocabulary for an operational on-line retrieval system for research reports. Stop words are removed, and there is some suffix stripping. The selection criteria are not specified in detail. It is worth noticing that word pairs as well as single words are considered. Also Jones comments on the problem of growing files: all text words are retained in a backing store, so when the vocabulary changes, documents may be re-indexed.

These studies are somewhat defective in retrieval testing; but they are of interest in showing general agreement in their views on the distributional properties required of good index terms; namely that words are likely to be of value if they are of middling frequency and exhibit restricted cooccurrence relations. This idea has been developed and tested recently by Salton 1972a,c, 1973b,c using the Q-function mentioned earlier in connection with term weighting. The function orders terms according to whether they are good or bad document discriminators, and bad discriminators, typically very frequent words, and null ones, typically very rare words, may be deleted from the indexing vocabulary. The experiments reported, particularly in 1973b, compared performance with simple stems and thesaurus for two Medlars abstract collections, 29x450 and 29x852 (ophthalmology). The results for the first collection, without searching involving feedback, are of more concern here. The indexing vocabulary was obtained by

eliminating first, quite crudely, words occurring only once or very frequently, and then poor discriminators. The results showed a noticeable improvement in average recall and precision over simple stems; and more significantly, that the automatic discriminator dictionary gave as good a performance as that obtained with a manual thesaurus (or the Medlars own system indexing).

The difficulty about all these procedures is in dealing with growing collections. There would clearly be problems about building a vocabulary from a sample of documents and then doing selective indexing for subsequent ones. But on the other hand, a good deal of effort could be involved in retaining all the source information for a really large collection, particularly involving full texts, and in reconstructing the vocabulary, using the more complex formulae, to achieve re-indexing.

IV.4.1b Specificity weighting

Extraction intended to obtain a fixed indexing vocabulary is not very satisfactory. More generally, it is difficult to satisfy a priori statistical criteria when actually indexing individual documents. It just turns out that some terms occur more frequently over a collection than others. It may, however, be possible to achieve vocabulary control post hoc, by weighting. Instead of making a choice of good terms, for subsequent assignment, the initial keywords extracted from a document are retained, and are given different values in searching at different times in the life of the collection. This is clearly much the same as reconstructing a vocabulary at intervals from full source information, particularly since weights can be zero. So the interesting question is whether simple weighting schemes can be effective.

The experiments by Sparck Jones 1972, 1973e mentioned earlier in connection with document descriptions suggest that simple techniques can be very effective. Weighting terms inversely by their straightforward collection frequency improved performance for the three test collections concerned materially. The generally Zipfian characteristics of indexing vocabularies (Krevett 1972) were exploited in the formula used. It is further of particular interest to find from Salton's 1973c tests with three 25x450 collections not merely that this method of weighting performs as well as his more elaborate one using Q-values, but that using the Q-function to determine a cutoff, as described above, gives the same results.

Taken together, these experiments suggest that inhibiting very frequent words is generally helpful, though perhaps not to the extent of eliminating them altogether if high recall is to be maintained, since request words are often frequent ones. Svenonius 1972 tried a very simple elimination of frequent words on the 42x200 Cranfield collection, without much effect on recall; but Sparck Jones 1973b tests with the 63x797 and 97x541 Keen and Inspec collections showed a rather disastrous lowering of the recall ceiling. However there seems to be a consensus that eliminating extremely rare terms is unlikely to degrade performance.

Cagan 1970 claims that they are very important, but their precise contribution to his retrieval experiments is not clear. Medium to low frequency words seem to be most important: for example both Svenonius and Sparck Jones found that deleting them was not helpful. It must be emphasised that since frequency is entirely collection dependent, frequent words may be technically specialised ones which are simply common in a particular set of documents and so are not discriminating for it.

IV.4.2 Classification

When index language classifications are constructed manually, words are grouped by explicit reference to their meanings. Unfortunately computers cannot recognise word meanings: their only resource is to make inferences about the meanings of words from their behaviour. Thus if two words tend to cooccur with the same third word, or set of words, we may infer that they stand in a synonymic, quasi-synonymic or paradigmatic relationship. Statistical association techniques are intended to pick up such relationships between one word and others, while statistical classification techniques are designed to identify classes of mutually cooccurring words, on the assumption that the lists or groupings found will be of value in retrieval. The product of an associative technique is an association list or 'semi-classification' indicating simply the relation of other words to a given word; for a vocabulary we obtain a series of these lists, one for each word. A classification technique produces groupings of mutually related words, ordinarily derived from association lists; in this case there may be no one-to-one correspondence between the number of words in the vocabulary and the number of classes. The problems of constructing classifications are clearly greater than those of forming lists, and are the main topic of what follows, though many remarks apply to both.

Constructing index language, and specifically keyword, classifications on a statistical base is probably the most distinctive line of work pursued in the general area of automatic indexing. It is reviewed in some detail in Sparck Jones 1973a. The main general points will be considered briefly here. Some of them are elaborated in Sparck Jones 1970a.

First, it must be recognised that many of the problems associated with the use of statistical classifications for retrieval are equally associated with manual classifications. In general, the fact that the object of a retrieval classification is to retrieve relevant documents no more determines the form of a manual classification than it does that of an automatic one. More specifically, if we think that a classification is required to act as a recall device, say, this does not constitute a very detailed specification of what the classification should be like. At the other end, evaluation of automatic classification is no more difficult than that of manual ones.

The particular difficulty of automatic classification comes in translating any specification of requirements we may have into automatic procedures. For example, if we want classes of synonyms,

how should we obtain these by purely numeric operations on initial data consisting simply of distributional information, i.e. records of the occurrences of keywords in documents? There are in fact two questions here: one concerns the classification base, that is what type of distributional information is taken as input; and the other the choice of formal criteria for grouping.

IV.4.2a Classification base

Early work on classification was typically based on the assumption that a classification should promote keyword substitution and therefore, by analogy with manual classification, that synonym classes (either general or subject orientated) were required. Synonyms typically occur in so-called complementary distribution in text, so they would be picked up via their cooccurrences with common other words. The difficulty is to find sufficiently significant cooccurrences to generate such groups. Further it is not necessarily true that this is the only meaning relation of value for retrieval. If two words tend simply to cooccur, they may be treated as substitutes and be just as usefully grouped for retrieval. There is clearly less difficulty about discovering such direct tendencies to cooccur. These direct tendencies to cooccur are sometimes referred to as first order associations between keywords, and indirect ones via other words, as second order associations.

Most work has been done with first order associations, an additional motivation being that much less computing effort is involved. Second order associations have been used by Cagan 1970 and by Vaswani 1970. The two forms have been compared by Jones 1968, Lesk 1969, and by the author. The comparisons suggest that there is no obvious gain from using second order associations, since the same term relationships tend to be picked up. In fact grouping with first order associations may easily group synonyms through their links with other class members.

IV.4.2b Formal criteria

Two choices have to be made: first, of measure of pairwise association between keywords; and second, of definition of class.

Association measures

The general literature on association, or similarity, measures is substantial. They have been extensively studied for biological taxonomy, for example (see Sokal 1963). Many of the more recondite problems occurring in such contexts do not concern us here. The important point is that the normal input for keyword classification is relatively simple, since it consists most often of simple presence/absence records for keywords in documents, or sometimes of simple weights (Lesk 1969, Minker 1973.) A variety of keyword association measures have been studied, for example by Jones 1967b, without reference to retrieval performance, and some limited comparisons were made by Sparck Jones 1971a. In principle

problems arise because high association values may not be significant, and equally that some low ones may: these are considered by Lesk 1969 and Vaswani 1970. We may generally expect most high associations to be significant, but difficulties arise with both very frequent and very rare words: associations between these may be of no benefit in retrieval, since they may not be particularly well correlated with relevance. Both Lesk and Sparck Jones removed both types of term. Applying a cutoff to retain only strong associations also seems desirable, as both authors show. The effort required to evaluate alternative similarity measures properly is considerable, and real tests have not been done. However there are good grounds for thinking that in this context, any robust simple coefficient like the Tanimoto (Jaccard) one is appropriate.

Class definitions

Ideally, anyone with a practical classification problem should be able to invoke standard, well-founded techniques. Unfortunately these do not exist, or at least exist only for some forms of classification. Ordered, and specifically hierarchial, classifications have been extensively studied and are relatively well understood. See for example Jardine 1971. But these do not seem, or at least have not been thought, to be obviously suited to keyword classification, though they have been used for document clustering. There is, however, an absence of any good mathematical theory of unordered classifications. In general, overlapping classes are accepted as appropriate to the different meanings or contexts of use of individual words, but exclusive classes are sometimes used.

A variety of class definitions are in common use, and some have been exploited for retrieval, but without any very solid justification for their adoption or understanding of their formal properties. Probably the most popular definitions are those of clique (maximal complete subgraph of the similarity graph), and connected component (of the graph). The fact that the former imposes very strong requirements on the connections between class members, while the latter imposes virtually none, has encouraged the use of various definitions intended to allow some weakening of connections in the first case, and to require some strengthening of connections in the second. The first course has been more popular. Such 'quasi-cliques' were studied by Vaswani 1970, and techniques for combining the very small, heavily overlapping cliques often found, like those adopted by Gotlieb 1968, have a similar object. The clump definitions used by Sparck Jones 1967, 1971a,c are equally intended to pick up relatively better connected parts of the association graph. Cliques themselves have been used by Minker 1970a,b, 1972, 1973 and Sparck Jones 1971a, and connected components by Minker and Hillman 1968, 1969, 1973. Other humble definitions have been used by Sparck Jones, including that of a 'star' which is derived from the set of keywords most strongly associated with a given word. It should be emphasised that if the association information is processed so that only very strong connections are retained, different class definitions may pick up the same sets of keywords. The choice of definition may therefore not be very important, and the computationally cheapest selected. Sparck Jones 1971a found performance for different classifications much the same, for this reason, and stars may therefore be exploited for test purposes. In general more sophisticated definitions like those of a clump have been designed to pick up significant but weak connections, and so may not be relevant to retrieval.

Computer algorithms

Forming a keyword association matrix and finding classes is much more strenuous than other automatic indexing procedures: the similarity matrix may be very large, and some class finding procedures, for example those for finding all the cliques in a given set of objects, are exigent. It should be pointed out that some definitions of class do not imply any particular class finding algorithm, so discovering a good algorithm may be very important. In general the grouping algorithms adopted have been rather rough and ready and designed, for example, to find some but not necessarily all of the classes satisfying a given definition.

IV.4.2c Classification experiments

The literature tends to be confusing because the same input information may be exploited in different ways, the same technique applied to different information, or the same output used for different retrieval purposes.

There have been two phases in work on automatic keyword classifications, either semi or full. In the first, from about 1960 to 1965, effort was concentrated primarily on showing that statistical classifications could be constructed and were prima facie plausible. A good deal of energy was expended in overcoming the non-negligible problems of handling realistic data samples with the limited computing facilities available. A number of association measures and classification techniques, like clumping, were studied. The results were evaluated for linguistic conviction: did the classes look like semantic groupings? There was no real attempt to exploit the classifications in retrieval. This early work is summarised in Stevens 1965a, and is represented by papers in Stevens 1965b.

In the second phase, a rather smaller number of workers concentrated on testing classifications in retrieval.

Semi-classifications, i.e. association lists, have been investigated at A.D.Little, by the Smart project, by Vaswani and by Dennis. Jones 1968 reports experiments with association matrices for 1000 items, in this case NASA thesaurus terms rather than keywords, indexing about 100000 documents (?abstracts). The association lists were to be used for request expansion, but the retrieval tests were so trivial that no evaluation is possible. Jones 1969 reports the use of lists with queries in an operating on-line retrieval system, but again with no performance evaluation. Much early effort was put into statistical association techniques at A.D.Little and this fragmentary literature is depressing. The lack of evaluation is not typical of recent work in the area in general.

The Smart work with statistical associations is reported in Salton 1968a,c, and fully by Lesk 1969. Lesk's experiments involved the 34x780, 42x200 and 35x82 IRE, Cranfield and ADI collections, with abstracts in the first two cases and full text in the third. He performed a range of experiments, for example with different cutoff levels. Performance comparisons were with simple stems on the one hand, and manual thesauri

on the other. In these tests both requests and documents were expanded. Statistical associations did not perform noticeably better than stems, and performed noticeably less well for the IRE and ADI collections than the manual thesauri. Lesk concludes that association lists may have a limited value as precision aids, but the thesauri have the advantage that they are recall aids as well. His view is that statistical information may be most valuable as an aid to manual thesaurus construction.

Vaswani 1970 included retrieval via association lists in his experiments with classifications on 93x11571 abstracts. The vocabulary was limited to 1000 stems. Among the various alternatives examined expanding both requests and documents was found effective, but the overall list performance, like that of the best classification, was no better than that for simple terms.

Dennis 1965, 1967 experiments were also quite substantial, involving a vocabulary of about 7000 keywords, statistically derived. A range of comparisons with expanded requests, or documents, or both was carried out for her 61x5121 and 6x556 legal texts. The results were evaluated against simple keyword matching. The performance of the various associative options was not consistent for the two collections. For the smaller collection the best association options worked better than simple keywords, but for the larger they were no better. Unfortunately, as noted earlier, the overall performance level was low.

Full classifications have been investigated by Vaswani, Minker, and Sparck Jones. Vaswani's 1970 experiments are distinguished by the large collection used, just described. As mentioned, he studied a number of different ways of using statistical information, as well as ways of generating classifications. His evaluation methods were slightly eccentric, depending, for example, on such criteria as the ability to retrieve specified numbers of documents (without ranking). Overall, the best cluster set and cluster using strategy did not perform noticeably better than stems (or indeed than the cheaper lists); though Vaswani's conclusion is that in some circumstances classificatory associative strategies would be useful: for example, an initial stem search with poor recall could be followed by an associative one.

Minker's experiments, 1972, were with connected components and cliques for two Smart collections of abstracts, 34x780 IRE and 18x275 Medlars, indexed with extracted word forms and stems, and with manual thesaurus terms. He has devoted a good deal of attention to such questions as the number of items per cluster, and so on, but their relevance to retrieval is not always clear. A distinguishing feature of his approach is the use of very high similarity thresholds, apparently without vocabulary precautions like those adopted by Lesk. In his tests query expansion had in general no effect on performance, degradation occurring only with large, low threshold connected components, or at high precision. This is not surprising since connected components are poorly motivated clusters, and high similarities are usually connected with rare terms with unpredictable retrieval behaviour. The use of these techniques for clustering well-organised thesaurus terms is somewhat dubious, and in general the justification for the approaches adopted is not obvious. Subsequent

experiments (1973) with the IRE and 35x82 ADI collections were designed to investigate the effects of using weighted input terms. This did not lead to any cluster performance improvements with the IRE collection, but did with ADI: the reason for the noticeable improvement with the ADI collection is not clear, but the collection is so small that the result must be treated with caution.

Sparck Jones series of experiments, 1970b, 1971a,b, 1973c, have been concerned with a range of classification and class using procedures, typically with manually extracted keyword stems as input. Most experiments have been with the Cranfield 42x200 collection but the 63x797 Keen and 97x541 Inspec collections have also been used. Some recent tests have been with nearly 3000 terms automatically extracted from abstracts for the 47x407 Keen subcollection. Initial comparisons between a number of different class definitions, used to give descriptor specifications of requests and documents, showed that classification must be restricted to non-frequent terms, and to fairly strong similarity connections, but that differences of class definition do not affect retrieval performance. Request expansion via stars seems to work as well as anything, and is usefully cheap. Classification procedures led to material performance improvements for the Cranfield collection, but not for Inspec and Keen. Sparck Jones has perhaps made more attempt than other workers to discover the reasons for variations and failures in classification performance. These are studied in Sparck Jones 1973c. In particular, the suggestion is that if relevant and non-relevant documents are poorly separated, it may be difficult to improve on the base performance given by simple term matching. In general the influence of collection factors like initial description exhaustivity is not clear.

In Hillman's operational on-line retrieval system (1968, 1969, 1973) connected component term classes, or genera, are used as an aid in formulating requests. The approach is similar to that of Jones 1969: terms related to those initially proposed are displayed to allow request modification. Devices of this sort are presumably useful, though they are not explicitly evaluated.

Automatic classification as an aid

It has sometimes been suggested that automatic classification procedures should be used not to generate index languages, but as maintenance aids. This was tried by Gotlieb 1968, with LC headings. In general, though most experiments have been with keywords, there is no priori reason why thesaurus terms should not be clustered, provided that the characteristics of the initial data, and objectives of the exercise, are well understood. For example, if thesaurus terms are designed to be exclusive, and very few are assigned to documents, there is not much ground for classification. The A.D.Little tests reported by Jones 1968 were with NASA thesaurus heads, and recently Jaccquesson 1973 discussed statistical associations for the ILO indexing vocabulary. Thesaurus improvement procedures have also been proposed by Wolff-Terroine 1971.

IV.4.2d Classification value

The main defect of most of the automatic classification experiments described is their relatively small scale. Even where substantial numbers of terms have been involved, there have been few documents. Or there have been few terms for many documents. Performance comparisons have to be made in two ways: first, against simple unclassified terms; and second, against manual classifications. Most comparisons have been against simple terms only, and as noted, some improvement has been obtained in some experiments, none in others. On the other hand, the best representatives among the automatic classifications and association lists do not perform less well than simple terms. Unfortunately very few comparisons have been made with manual classifications: Lesk's are the most notable. At best some tentative inferences can be made: the most obvious, though not necessarily encouraging, is that since manual thesauri do not ordinarily perform much better than simple terms, they cannot be expected to be very superior to an automatic thesaurus.

It is difficult to compare the various automatic experiments themselves. There is great variation, particularly in the source and number of the initial keywords, and typically in the initial level of document description exhaustivity. In particular, while the Smart experiments and Minker's used keywords extracted from abstracts, Sparck Jones mainly used manual keywords, A.D.Little and Dennis a statistically extracted vocabulary, and Vaswani a manually extracted vocabulary.

The conclusion must thus be that while automatic classification techniques have not been disproven, neither have they been proven. More systematic experiments using such information as we have about classification characteristics, the criteria for their success, and the effect of different collection factors on them, are needed. It should at the same time be emphasised that the computational and economic side is not a real problem: compared with manual thesaurus construction, making and updating a simple automatic classification for use, say, in query expansion, is not likely to be a major enterprise.

IV.4.3 Automatic index language maintenance and improvement

The use of computers for index language maintenance and in assisting thesaurus construction (other than by classification) deserves a mention, though it is on the humblest level of automatic indexing. It is of interest mainly because programs designed to provide statistical information, or to spell out relationships, and so on, presumably contribute to the good organisation and value of the language itself. Illustrative references are Clough 1971, Hines 1971 and Rolling 1970. Lefever 1972 discusses computer aids for rationalising a large natural language vocabulary for BIOSIS.

IV . 5 Search semantics

Boolean requests have already been dealt with under syntax. The main semantic devices used in searching are keyword or term truncation and weighting. Truncation is applied (usually manually, but in principle automatically) in automatic systems in which natural language titles, abstracts, or extracted keyword lists are available as document descriptions. It is, for instance, applied by IITRI (Williams 1972). Left truncation may be adopted, for example for chemical data. There seems little doubt that truncation is useful.

Automatic weighting has been partly considered under document description. Some term weights, like within-document frequencies, are necessarily specified for documents. But collection frequency based weighting, for instance, may be associated with request rather than document terms (the distinction being immaterial), though the ease with which this can be done may vary. The simple scheme used by Sparck Jones 1973e is readily applied to request terms. Miller 1971b suggests a modification of this to take account of the user's estimate of the relative frequency of terms in relevant documents. Experiments with a 25x(6x35000) Medlars collection compared these probabilistic request formulations with standard Boolean ones, the former showing a slight performance gain. The live user in this sort of scheme might be replaced for profiles by information based on the success of past searches, Intuitive user weights are allowed in some systems: see, for example, Hisinger 1971. Complicated scoring systems based on them are discussed by Sommar 1969, Matthews 1970 and Clough 1971.

In exploiting index language structure in searching, related or higher level descriptors may be invoked from a thesaurus or classification. No automatic classifications have been tested for this. Request expansion is a slightly different technique, as is Cagan's 1970 use of request-document term linkages for matching.

IV . 6 Conclusion on semantic indexing

It is possible to be rather more definite about the value of automatic techniques in relation to semantics than it was about syntax. In initial document analysis statistical extraction methods do not obviously pay their rent. In description computers may be used in a useful way for such purposes as suffix stripping. More importantly, statistical approaches to the treatment of keywords in descriptions, and to vocabulary selection and control, all essentially through weighting, appear to have positive merits. In particular, they may improve simple keyword performance so that it reaches a level competitive with manual controlled language indexing. Statistical techniques for structuring an indexing vocabulary to generate some degree of classification cannot, on the other hand, be unequivocally supported.