VII RELATED AREAS

VII . 1 Automatic abstracting and extracting

Not much work is being done in this area. In particular there is none on abstracting proper, which is hardly surprising given its manifest difficulty. Work on automatic extracting may attempt to identify key sentences either by structural analysis of the text, or by applying statistical criteria, following the early suggestions of Luhn (Schultz 1968).

VII.1.1 Parsing approaches

A variety of techniques have been investigated by Earl 1970, 1972, 1973. She reports attempts to identify sentences with particular syntactic forms, on the assumption that sentences with these forms are important content carriers. Sentence characterisation was achieved by partial parsing techniques designed to identify phrases as well as the parts of speech of individual words. Unfortunately the results with test texts showed far too many different sentence patterns for any to be selected as critical ones. A subsequent experiment attempted to combine this syntactic information with statistical information to select sentences. The parser was used to identify noun phrases. The non-function words in these were selected and frequency counts obtained for them. Sentences with high concentrations of high frequency words were then selected to form extracts. Earl 1970 describes the results as "mildly encouraging".

Earl 1972, 1973 describe further work on the parser. This is intended to operate at progressively higher levels, yielding a complete structure for a sentence at the highest. Ambiguities at any level should be resolved at that level: thus the lowest level is designed to resolve noun phrase ambiguities. To resolve ambiguity word government information is exploited: this takes account of the fact that words are habitually used with specific syntactic constructions they can be said to govern: for example the verb "believe" may govern a prepositional phrase of the form 'in+Substantive'.

This work on parsing is part of a general program on automatic analysis; the phrase structure approach might be adequate for extracting, for example, but the tests have been very limited.

A second project in extracting is that reported by Rush 1971, 1973. Rush 1971 discusses automatic extracting in all its aspects, and places particular emphasis on criteria for rejecting sentences as opposed to those for selecting them.

The project seeks to exploit contextual inference and syntactic coherence criteria for sentence selection. The first depends on both the location of elements in sentences, and the presence of cue words. For example the presence of the phrase "our work" in the main clause of a sentence probably indicates that the sentence and specific clause are important. To maintain syntactic coherence in an extract, a sentence may be rejected if it refers to another. The system developed is designed to produce extracts satisfying general requirements, for example that extract length should be about 10% of the source text length, that methodological remarks are excluded, etc. It depends on a word control list containing words or strings with (positive or negative) semantic weights, and syntactic values, and a set of rules for applying the weights. These rules depend on some syntactic information, so a partial, crude parsing of the text sentence is carried out, using the syntactic values. The assigned semantic weights determine the sentences to be selected. The procedure is illustrated with sample abstracts, of a quite plausible character.

Rush 1973 describes improvements to the procedure designed to produce smoother abstracts with better related sentences. These require a more elaborate syntactic analysis of sentences. A closed class dictionary is used, and the parser identifies clauses and functional components like subjects. This information is exploited for various sentence modification procedures, for example combination by coordinating or subordinating conjunctions. Performance in this case is not illustrated.

VII.1.2 Statistical approaches

These are perhaps more appropriately called verbal approaches, since some of the techniques, without being syntactic, are not statistical either.

In his survey Edmundson 1969 distinguishes four non-syntactic methods of extracting:

1) the cue method involving a comparison dictionary defining good, bad and indifferent words; "significant" is a good word, for example. The classes have different weights, and sentences are rated by their combined word values.

2) the key method depending on the frequency of words in a text; sentences are rated by their combined frequency weights.

3) the title method, in which the combination of title words in sentences determines their selection.

4) the location method, in which sentences are rated by location.

Methods 3 and 4 in fact imply a gross structural analysis of documents, and so could be described as syntactic. Edmundson describes comparative experiments with 200 chemical documents in some detail. The four methods were evaluated by comparison with target manual extracts, using the percentage of sentences co-selected. The best averages were for location and cue, with key worst. Combining cue, title and location improved the average slightly, and reduced variation. Keys were clearly shown to be detrimental, since including them as well reduced the average. Edmundson comments that this is practically a useful result, since frequency counting is an effort. On the other hand, a cue dictionary has to be provided, though this might not be large. The difficulty about the project as a whole is that no attempt was made to determine the value of the target extracts with which the comparisons were made.

In general, it is difficult to evaluate the work on extracting since the results presented are somewhat anecdotal. Rush is clearly pursuing a fairly sophisticated approach, so it would be interesting to have some concrete results. It is perhaps noteworthy that there has been so little follow-up to the statistical approaches advocated by Luhn, though they are clearly quite crude. Experiments in extracting are in any case very expensive. A substantial corpus of full texts is required for proper experiments; and there is perhaps little need for automatic extracting or abstracting when document authors tend to provide their own. Even if these are not always satisfactory, it is unlikely that significantly better abstracts could be provided by the kind of technique described here.

VII . 2 Question answering (fact retrieval)

Research on question answering systems is relevant to document retrieval at two levels. At the general level it is of interest if it is assumed that the real need is for direct retrieval of information, rather than indirect retrieval via documents. At a lower level the techniques developed for analysing input texts could in principle be exploited for obtaining document descriptions.

Work on question answering, or fact retrieval, systems derives in part from early work on computational linguistics (and hence, ultimately, machine translation), and in part from research in artificial intelligence. It is characterised by a hands-on, get-with-it attitude to applying linguistic ideas not typical of theoretical linguists. Question answering systems usually involve procedures for analysing input natural language text, an information store containing both permanent dictionary information and temporary data acquired during a question and answer dialogue, and rules to make the inferences necessary to provide answers. The essential feature of a question answering system worthy of the name is some 'understanding' of its input. In document retrieval input analysis is confined to identifying key items which are available for recognition during searching. There is no real attempt to spell out what the key information implies.

Question answering research differs from that done by theoretical linguists in two important respects. The requirement for proper input analysis means that some attempt must be made to tackle semantics as well as syntax. Second, the interest in analysis has led to procedures reflecting the language user's recognition processes rather than his abstract linguistic capacities.

Most question answering projects concentrate on one particular aspect of the whole activity: for example on operations on the information store rather than on input analysis, so that fairly crude syntactic and semantic analysis procedures are accepted as adequate, or input forms are restricted. But more ambitious systems attempt a deep syntactic and serious semantic analysis.

An overview of the state of the art in question answering can be obtained from the papers in Minsky 1968 and IJCAI 1969, 1971, 1973, and from the surveys by Simmons 1970 and Walker 1973. Four projects may be selected as representative of the range of work in the area, chiefly from the analysis point of view.

The Linguistic String Project (Sager 1972, 1973a,b) has been concerned with the identification of information in scientific texts and its presentation in regular 'formats'. The approach is based on the view that particular scientific fields have characteristic word class relations. So parsing involves a specific sublanguage grammar with appropriate classes and structures. In the field investigated, pharmacology, different noun classes define words acting as, for example, pharmacological agents, which are associated with distinct classes of verbs. The parsing procedure is designed to identify the string structure of a sentence, which consists of one or more elementary sentences with operators applied to them. The elementary sentences will typically represent characteristic propositions of the science field, such as 'digitalis inhibits', which are modified by such verbs as "promote". Each input sentence is therefore replaced by a formatted version exhibiting the string operator structure. The parser is context free with restrictions, allowing identification of structural relations underlying the surface of a sentence. Sager 1973a notes that the parser is relatively successful, dealing with about 75% of the input sentences tested. Sager 1972 mentions that it has been applied to a corpus of articles, but there is no information so far about exploitation of the results.

The other three parsers I shall describe are intended to be general for English.

Wood's 1970, 1972, 1973 augmented transition network parser is derived from simple state transition networks, but is much more powerful, first in allowing recursion, and second in associating tests and structure building actions with transitions. The result is equivalent to transformational grammar, but the parser, while providing deep structure information, has the efficiency of phrase structure operation. The main application of interest here has been in the Lunar Sciences Natural Language Information System, where the parser is used in seeking information from a data bank and set of documents about moon rocks. The parsed input sentences are therefore translated into convenient forms for searching either the systematically organised data about mineral analyses, etc., or document keyword sets. In the analysis procedure the treatment of syntax and semantics are separated, though Woods notes that syntactic and semantic analysis could be done concurrently.

The output of the syntax analyser is checked for semantic propriety using constraints associated in this case with the subject area of lunar science. For example the correct interpretation of the word "sample" depends on the presence of rock names. Woods 1972 reports 78% success in treating requests in a public test.

Winograd's 1972 system integrates syntactic and semantic analysis procedures during sentence processing. Unlike the preceding workers, he is specifically concerned with dialogue between a human and a (simulated) robot manipulating blocks. The parser seeks to identify the underlying structure of a sentence, characterised in this case in a very different way from Chomsky's, carrying out semantic operations for resolving structural ambiguities and selecting word senses in the process. The distinctive feature of his work is the translation of input sentences into procedures for establishing their meaning rather than static representations of meaning. Thus the sentence "Is there a green block in the box?" is represented by a series of instructions designed to discover whether there is a block in the box and it is indeed green. The system is very impressive, and very complex sentences are successfully handled to generate continuously developing dialogue.

The semantically most ambitious of these projects is Wilks' 1971, 1973. In this there is no attempt at conventional syntactic analysis. The procedure is designed to map semantic message forms, or 'templates' onto text to provide an interlingual characterisation of the text (ostensibly for machine translation). Thus sentences are identified at the highest level as instances of templates such as MAN+HAVE+THING, which would apply, for example, to the sentence "The fat man owned a shiny new car", while lower levels of analysis are reflected in dependent templates: thus MAN+BE+MUCH might characterise "the fat man". The parsing procedure first segments input texts using such indicators as punctuation and prepositions, and then maps templates using an inventory of templates for sentence and phrase forms and a dictionary with appropriate entries in the same conceptual language for individual words. The idea is a sort of backdoor approach to the real semantic structure of the sentence. An interesting feature of Wilks' work is that his procedure is specifically intended to analyse paragraphs as wholes, rather than individual sentences separately. This allows for the fact that semantic resolution is often not possible within sentence boundaries.

From the document retrieval point of view, the main defect of this work to date has been the very limited scope or size of its test input. Sager has been concerned only with pharmacological literature, Woods with questions about lunar samples, Winograd with questions and commands about toy blocks, and Wilks with an assortment of test paragraphs.

The form of organisation of the information store referred to to answer questions, and operations on it, have been studied by a number of question answering projects. This question is also in principle of interest to workers in document retrieval. An information store in a question answering system will naturally have permanent information, represented chiefly by a dictionary but also, for example, by axioms about the world of discourse involved; it may also build up a temporary store during a dialogue. Four forms of store organisation can be distinguished: simple lists, where individual entries can be quite complicated but there is no overall structure of any complexity; networks, where different items in the store may be linked in a variety of ways; predicate calculus representations; and procedures. Operations may be divided into those allowing only simple matching and those involving some degree of inference. The latter may be represented by path following in networks; by the application of strict logical deduction; or heuristic mixtures.

Simmons 1970 comments on the progress evident since an earlier review in 1966. It is clear that more comprehensive, effective and rapid analysis procedures have been developed, that there is more sophistication in the organisation of stored information, and that more realistic, non-trivial inference procedures have been implemented. These developments have been materially assisted by advances in computer hardware and software. Nevertheless, from the point of view of workers in document retrieval, the most conspicuous feature of experimental question answering systems is their limited universe of discourse. The information stores, numbers of works and numbers of sentences involved are all typically small. This means that the effort of full scale semantic analysis is avoided. The difference between document retrieval systems operating on tens or hundreds of thousands of documents, but relying on quite simple linguistic processes, and question answering systems operating on tens or hundreds of documents, but applying quite elaborate linguistic techniques, is very striking. The possible benefits of research on question answering and fact retrieval for document processing are therefore only likely to appear some time in the future, rather than in the present.