What might be in a summary ?

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK
sparckjones@uk.ac.cam.cl

June 1993

From 'Information Retrieval 93: Von der Modellierung zur Anwendung'
(ed Knorz, Krause and Womser-Hacker), Universitatsverlag Konstanz, 1993, 9-26

1 Abstract

The paper presents a framework for, and strategies adopted in, an investigation of summarising designed to place future work on automatic summarising on solid foundations. The work reported has been focused on the role of large-scale text structure, and the paper describes comparative studies of different approaches to the characterisation of source text structure and to the use of this structure in summary formation.

2 Introduction

In this paper I shall describe some foundationally-motivated work we have been doing in Cambridge, establishing and exploiting a framework for automatic summarising. I shall introduce this account by noting some relations with indexing, and in conclusion return to the connection between the two forms of information capture and use.

We can all, and do, summarise; and we use other people's summaries. Summarising, especially in the guise of extracting, was an early task for automation (Luhn, 1958). But not much progress has in fact been made with automatic summarising, other than for very restricted or relatively undemanding applications. Research in natural language processing (NLP) has become increasingly concerned in the last decade with processing discourse, but mostly as dialogue rather than as extended single-source text. Text processing has recently become an important area of work, stimulated by the (D)ARPA initiative in message processing (MUC-4, 1992), but this is aimed rather at extracting than at summarising proper. Progress in NLP in general nevertheless means that it is now worth addressing summarising. The project I shall describe, joint work with Philip Gladwin and John Beaven, has been intended to stake out the ground for automatic summarising, in particular through initial research on the nature and use of the large-scale text structure on which summarising may depend.

2.1 Summarising and indexing

Summarising is one form of information management. It is closely related to indexing, since a very brief summary may alternatively be viewed as a complex, structured index description. Summaries (i.e. abstracts) have also been taken as the surrogate search base in document retrieval, and may be viewed in text searching as multi-purpose, open index descriptions that the individual retrieval request sharpens through the specific match it makes. However though document retrieval may benefit from searching on such a distilled concentrate of the full text, as opposed to searching on the ultimate text itself, this is a use of existing summaries and we are concerned with summary formation itself. From this point of view the two major distinctions between summaries and index descriptions are that summaries are texts themselves, and further, that they are intended to be taken as substitutes for the source documents and not just as keys to these. These are normal properties of summaries as opposed to index descriptions, even if the boundary between the two is not a sharp one, and index descriptions may, for instance, be used as substitutes in search output scanning.

The other important point of similarity, and difference, between summarising and indexing is in their evaluation. Both are intrinsically predictive, from linguistic elements to content pertinence. In both cases this makes evaluation an intractable problem, not so much because needs may not be accessible as because pertinence is a 'one-off', i.e. is situational to a unique occasion (Sparck Jones, 1990). The difference between summarising and indexing is that the indexing role is coarser, so prediction is easier and evaluation can be more robust. A summary is an intrinsically more complicated object, standing in a more complicated relation to its source text, than an indexing description for that text; and this implies more opacity, and variability, in how and therefore whether it meets the user's need.

3 Summarising framework

Our work has been against a background that can be only briefly noted here. Thus we have had to make some ground assumptions about what summarising is, and to place it in its context.

On the first point, we take a summary to be a condensed derivative of a source text, i.e. a content reduction through either selection or generalisation on what is important in the source. In a summary, source information is abridged and its expression is shortened, since we take the way information is expressed to be one aspect of text content. Thus if we think of condensation as operating on two axes, we can contrast summaries with other forms of reduction that are equally concerned with key content, by the relative emphasis they place on the information and expression axes. A synopsis, for instance, is more concerned with content information than its expression, where the reverse holds for a precis. The implication is that concentrating on the one rather than the other can carry with it more regard for semantic coherence or for syntactic cohesion in the output summary text. However all of these notions are still informal: one object of the project has been to explicate them, and in particular the common core interest in conveying what is important in the source text. (Though for many "abstract" and "summary" are synonymous, practical concerns have tended to associate the former with technical journal literature and thus, perhaps, to place slightly greater emphasis on information than expression. We have therefore adopted "summary" as the most general, and balanced, term.)

3.1 Summary factors

In relation to context, there are three classes of factor affecting summarising. The first is the nature of the input source text: this subsumes are range of properties under the general headings of form and subject type, including scale and medium under form, and ordinary or specialised under type. Thus a scientific article and a management plan have very different characteristics as sources. The second class of factor is the intended purpose of the summary, covering different possibilities under function and audience. Thus summaries may be for informing or alerting, and equally may have narrowly targetted or quite comprehensive audiences, and ones with different degrees of subject knowledge. The third class refers to output factors, to do with material and format respectively. Material covers the degree of coverage of the input, format the way of organising it, for instance as running text or itemised points. Input and purpose together strongly influence output choices, but cannot usually be tightly enough defined to strictly determine them. Moreover as all of the context factors mentioned are hard to analyse and define, there is substantial work to be done in investigating them and in determining their implications for the process of summarising as this has to be specified for automation. Thus input factors relate to genre, a notoriously slippery notion.

There is however one immediate consequence even of a quite informal review of summary context factors. This is that we should not expect to have a single general technique for automatic summarising. What we are thus seeking, just as in document retrieval, are techniques which are general in being relevant to well-defined classes of situation. We are not concerned with trying to build a system suited only to one particular application. However even this modified generality is a long-term goal, and within the scope of our quite modest project we have had to work within narrower limits, though we should be able to get some pointers to broad strategies, as described below. In particular, with respect to input texts we have worked only some short, 'ordinary' texts, though of several presentational genres; in relation to function and audience we have assumed a need to inform the same sort of general readership as the sources; and for output we seek conventional running text, though as the sources are short, summaries are very short. But we can still, within these limits, explore the most important issues for summarising.

3.2 Process structure

We have also adopted a global processing model as follows. We assume that summarising requires some meaning representation for the source text, i.e. that in general it is not sufficient to work from the surface text alone even those this is, at the limit, one possible strategy. We further assume that the source representation is available as a whole. Thus while human summarising, especially as a normal human activity during discourse processing, may be dynamic and so lead to a continually changing summary (Kintsch and van Dijk, 1978), we are initially simplifying by having a stable representation for the complete discourse. This simplification can however be justified by the fact that summaries are often also prepared (by human abstracters) by review, and indeed re-review, of a text as a whole. We are further concentrating on what the source text representation is like, rather than on the detailed processing by which it is obtained. We are not, that is, addressing the specific operations required to derive the source representation, primarily because we want to focus on what use different kinds of representation may be for summarising, but also because no-one can yet automate the complex processes involved. We are thus seeking the pragmatic compromise of 'pseudo-automation' in source processing. We also assume that in general source text interpretation subsumes two subprocesses or stages, providing analyses for individual sentences, and integrating these to form the text representation as a whole.

Correspondingly, just as summarising requires a source text representation, we have initially to form a summary representation embodying the organised content of the summary, from which the actual output summary text is synthesised. The latter may in turn subsume two subprocesses, text planning and text realisation, though the need for and level of planning may depend on the nature of the summary representation. Within the overall process of summarising, the formation step where the summary representatiion is derived from the source representation is clearly the crucial one, since this is where the condensing transformation takes place. This stage has also, in our project, to be approached pseudo-automatically, and we are currently not investigating the synthesis step in detail, partly because it is less important and partly because for short outputs it is not problematic and is within the scope of current automation. The key questions we are thus focusing on in our project are first, what information can the source text representation make available for summarising, where this can include information supplied from existing knowledge sources to support that directly supplied by the input text; and second how this information is exploiting for summarising (again allowing for supplementation from knowledge sources independent of the text).

4 Relevant background

There are three areas of prior art relevant to our project. One is human summarising, the second is work on automatic summarising so far, and the third is research on the nature of discourse interpretation and representation. We take the current state of the NLP art in sentence processing, e.g. delivering logical forms as sentence meaning representations, along with some anaphor resolution capability, for granted. Relevant work on long text processing under e.g. the (D)ARPA message understanding conferences (MUC) (cf e.g. MUC-4, 1992) can be subsumed under the automatic summarising head.

In relation to work on human summarising, and specifically that on professional practice (e.g. Rowley, 1982; Cremmins, 1982) rather than in the psycholinguistic area, the main points are as follows. The first is that there is often a strong emphasis on particular purposes, e.g. abstracting scientific journal articles. The second is that the account is strongly processual, with cycles over the source text drafting, writing, and revising the summary. The third is that, essentially as a consequence of the first two, the summarising can be guided by checklists. The fourth, however, is that from the automation point of view the instructions are too abstract (e.g. Rowley, 1982, pp 24-5), even if human abstracters taake advantage of surface cues. Finally, and particularly importantly from our point of view, there is no discussion of source text representation as such, especially in detail, as opposed to what may be deemed the summary representation given by e.g. notes of the main points of the source or (more shallowly) a list of key terms. Human practice thus does not offer much concrete input to future automation, other than via the use of source expressions defined on quasi-statistical or locational grounds as key content cues, and of structure indicators like subheadings.

Work on automatic summarising has in general been concerned with a very specific application (Young and Hayes, 1985; Rau, 1988), or has applied relatively shallow techniques, for instance those more properly characterised as aimed at extracting rather than abstracting, exploiting statistical and locational cues (cf Paice, 1990). Work within the MUC framework (cf MUC-4, 1992), though sometimes applying deeper analysis, is explicitly directed towards fact identification and extraction for a predefined and narrowly defined class of facts, and not towards condensation, though some of the methods have potential for summarising. Perhaps the most general approach to summarising so far significantly tested computationally is that adopted by DeJong (1979), using scripts, but these were applied in too rigid a way, and later proposals for a more flexible version of this approach (Tait, 1983), were not extensively tested. Hahn (1990)'s work exploiting text thematic structure in conjunction with domain knowledge is intended to be more responsive to general linguistic principles of text organisation, but still relies very heavily on specific domain resources embodied in lexical experts and frames. Thus research on automatic summarising to date has not provided us with much on which to build. The most relevant idea is that of using frames or scripts, which we have included in our investigations as offering one form of deep source representation to be compared with others. But the shallow approaches relying on surface cues that have been investigated also need to be considered in our comparative evaluation, as they are important for theoretical as well as practical reasons.

Relevant work from discourse studies includes that on text coherence and cohesion (which can be simplistically taken here as dealing respectively with text semantics and text syntax) (e.g. Halliday and Hasan, 1976; Hobbs, 1985); on the role of rhetorical relations in discourse organisation (e.g. Mann and Thompson, 1987; McKeown, 1985); on the discourse properties of particular genres, as illustrated by story grammars (e.g. Rumelhart, 1977) and by Liddy et al (1993); and on forms of discourse organisation, notably those giving text content a hierarchical structure, that appear to be general to any kind of extended discourse (e.g. Kintsch and van Dijk, 1978). It seems evident that any kind of discourse organisation above the sentence must play a part in marking what is important and thus be relevant to summarising. Indeed Kintsch and van Dijk see their hierarchical structure as a summarising device that is exploited during discourse interpretation to impose control on the mass of incoming material while preserving its most important content. However as the literature on discourse structure above the sentence level is extremely heterogeneous, and many notions like rhetorical relation are far from adequately defined, it is very difficult to apply the ideas stemming from this area of research.

5 Project strategy

It is evident that as we are interested in general summarising principles rather than in any particular application, the range of possibilities to be considered is very large. We have therefore had to limit our investigations, but have tried to do this in an appropriately motivated way. Thus on the basis of our analysis of context factors and global processing, the strategy we have adopted is as follows. All the evidence is that large-scale text structure plays an important role in summarising. We have therefore focused on the nature of large-scale text structure, as reflected in the source text representation, and on its use in forming summary representations. We treat interpretation as independent of summarising, both because we want to see what the presumably rich source representation required to support discourse interpretation and response in general offer for summarising, and because we do not want to allow any premature bias towards summarising to lead to over-restrictive source representations. If it appears that any particular kind of representation is especially suited to summarising, it would then be reasonable to see whether tailoring exclusively towards summarising in constructing the source representation during interpretation could be efficient. We also wish to treat source representations are neutral to allow for different approaches to exploiting any one kind of representation for summarising.

5.1 Information types

Then in order, further, to control the range of possibilities to be examined, we have found it useful, at least as an initial heuristic, to view discourse as involving three *types* of information and to categorise large-scale structure according to which of these three types of information they emphasise. Source representations can capture linguistic information, or domain world information, or communicative information. Taking the short example text of Figure 1 as a compact working illustration, Figure 2a shows how three successive sentences from it embody, or convey, linguistic, world, and communicative information. These three types of information are of course intimately linked: for instance, the word "valuable" is associated with a property of books and with the writer's communicative interests. The types can also be assumed to be grounded in common sentence meanings. However the different types of information offer their own different perspectives on the source text.

The same applies when a source text as a whole is considered. Thus it is possible to see that a text has a large-scale structure depending on one type of information rather than another, as illustrated for the Figure 1 text, again informally, in Figure 2b. Again these structures, to be captured in the source representation, would be grounded in individual sentence meanings. But they give a very different characterisation of the text i.e. discourse, as a whole and correspondingly imply different bases for summary formation, with possible results as suggested in Figure 2b.

We are not necessarily claiming that source representations can only capture one type of information rather than another, though it is not yet clear how a single representation could capture all of them when they are as distinct as Figure 2b suggests they may be. But in our research we have deliberately chosen to be exclusive, in order to investigate the resources for, and constraints on, summmarising. Thus one important issue is whether any key source content will be identified regardless of the particular access route to it via one sort of representation rather than another.

5.2 Representation forms

It is further possible, in considering large-scale structure, to see a contrast of form between what we have called bottom-up and top-down *forms* of representation. This distinction is not a processing one, or one about the formal properties of representations, but is intended to reflect the difference between structures for a given discourse that are highly individual since they are created by the application of independent rules, and ones that are essentially instantiations of prior, generic schemas. This distinction is again a heuristic one that we are using as a lever to stake out the complex terrain of discourse structure characterisation. The structures resulting in eaither case may have the same formal properties, e.g. be hierarchical; and there are clearly intermediate possibilities of the 'grammar' kind involving interdependent rules.

6 Working procedure

Thus our experimental space for source text representation and summary formation is that defined by a grid combining choice of information type (linguistic, world, communicative) with choice of represention form (bottom-up, top-down). Each grid slot refers to a *sort* of representation, clearly with many specific possible variants. To control our experiments we have therefore adopted a working strategy as follows. First, we have chosen to to work with a set of test texts, ten paragraph-to-page length items, to which all our approaches to representation and summarising are applied. These ten texts include three 'noddy' texts of a clear and simple kind (see Figure 1), originally written for another purpose and constituting limiting test cases, and seven real weekly newspaper texts. [1] Though longer texts are desirable, our test ones nevertheless have large-scale structure, and they have provided plenty of work. They can be described, informally, as illustrating various genres: report, argument, narrative, etc.

Second, as a model for each sort of representation, we have taken an approach already proposed or developed in the field as an exemplar, e.g. scripts as a world-based, top-down sort. Whatever the merits or demerits of any particular exemplar, we believe it is better to start with approaches that have already been worked through and tested at least to some extent, in the best case computationally, without being manifestly hopeless. Of course we have had to develop our own versions of these approaches, e.g. scripts relevant to our test text topics. But we have as far as possible attempted to construct general-purpose resources, for instance scripts clearly applicable to other texts, and to exploit them as neutrally as possible: we have not designed structures to suit single test texts. We have in particular found resource development necessary in trying to apply exemplar strategies originally designed for text generation to text interpretation.

Third, we have chosen to assume a common kind of sentence representation for the source text, namely logical forms of the now conventional kind as produced by e.g. the SRI Cambridge Core Language Engine (Alshawi, 1992). If we allow for anaphor resolution, identifying the discourse referents of pronouns and other nominal expressions, as a necessary part of source interpretation, this supplies a propositional network linked by discourse entity nodes that we take as a baseline source text representation (cf Gladwin et al, 1991). Thus alternative treatments of large-scale structure can be seen as leading to richer, less or more elaborated, full source representations.

For our first studies we sought at least one exemplar for each source representation sort, though we have not yet completely filled the grid in this way. At the same time, for some slots we have tried more than one exemplar, reflecting the pragmatic course of the research and also an interest in comparing only slightly elaborated full representations with much richer ones of the same sort. We have also been motivated in these cases by the results of summarising following natural strategies for a particular choice of representation.

[1] Noddy texts are simplistic ones in children's book style.

7 Strategies studied

The representation choices we have investigated are as follows.

7.1 Linguistic approaches

For large-scale linguistic structure we have sought to apply Mann and Thompson (1987)'s Rhetorical Structure Theory (RST) for the bottom-up case and, as this is a quite rich structure, have also simply recorded the current focus history as defined by Sidner's focusing algorithm (Sidner, 1983). The focus history leads to a quite natural approach to summarising, choosing the discourse entity most often in focus as identifying the key text topic. For the example text of Figure 1 this identifies the discourse entity 'biographies' as most focused, as shown in detail in Gladwin et al (1991). Of course this does not also choose what to say about the key entity, and it would still, even if thus supplemented in some suitable way, be a very weak, selective strategy.

We take the line that RST can be viewed as dealing with linguistic information since rhetorical relations, especially as flagged by lexical and other surface cues, may be seen as embodying standard, known ways of organising text that are conventionally associated with seeking to achieve certain communicative effects. RST is thus not directly a communicative theory; equally, it is a way of organising text explicitly and not simply of transmitting domain relationships. RST has hitherto been computationally applied for discourse generation. We found it difficult to apply in interpretation in a reliable way, especially since identifying relations presupposes full text understanding and also, for instance, because we could not always obtain a strict hierarchical structure. However RST does have a natural and quite intuitively appealling summarising corollary, namely picking up the nucleus in each relation working recursively down the structure from the top to the ground nucleus clause. This strategy, again a selective one, is illustrated in Figure 3a, with a quite straightforward state-of-the-art style of text realisation from the information supplied by the operations on the source representation.

We had problems in finding a top-down linguistic strategy. Past work on rhetorical schemas (e.g. McKeown, 1985) would appear to suggest that text may instantiate such schemas. But so far these have only been used automatically to generate text, and we could not easily analyse our test texts as instances of existing schemas, even of Maybury (1991)'s relatively flexible kind, or of generally comparable ones we might have constructed. Developing adequate new schemas would have been a major exterprise in its own right. We were therefore obliged in this case to adopt an intermediate sort of representation motivated by Rumelhart (1977)'s story grammars, and constructed text grammars for our (and hopefully other) texts. These are naturally somehat genre specific, and even (as with RST) appear to deal in domain-oriented rather than purely linguistic notions. But they can nevertheless properly be regarded as discourse grammars since notions like 'event' or 'episode' are far more general than domain notions like 'eating event'. For summarising, we pursued the grammar analogy and associated summarising rules, essentially of a selective sort, with each 'syntactic' grammar rule, with results as shown in Figure 3b.

7.2 Domain approaches

For the structures motivated by discourse or domain world information, we have so far only experimented with a very rudimentary version of the bottom-up approach, somewhat related to e.g. Lehnert (1992)'s work. In this case we look at the baseline predication network and note the way discourse entities participate, in any argument role, in predications, so we are effectively amplifying the initial representation with explicit participation frequency information, as shown in Figure 3c. It is natural then to choose the most frequent entity as the summarising vehicle, though as with the earlier linguistic case it is not clear what predication information should be supplied for the chosen entity either by selection or generalisation. Though the source predications stem from linguistic processing, this approach to representation is really divorced from the text structure itself and is rather concentrating on the exhibited properties of the world the discourse is about. However it would be desirable to investigate a more substantial bottom-up domain structure, for instance by filling out the predication data extracted from the source text with additional inferences (as might naturally occur during interpretation).

As a top-down strategy for world information we have experimented, following the obvious lead supplied by DeJong (1979) and Tait (1983), with scripts and frames. (Though these may be deemed to refer to events and things respectively, we call them all scripts.) As we do not have access to any existing stock of scripts we have had to construct our own, hopefully ones as generally applicable as possible. For a given text we allow multiple current, and also nested, scripts. Our initial strategy for summarising with scripts was along the lines of previous work, instantiating 'overview' or 'covering' templates already provided for a script as a whole, as shown in Figure 3d. But we feel this is too crude, and we have therefore been experimenting with a more discriminating approach based on density functions, aimed at exploiting those parts of scripts which are most fully instantiated and thus presumably of most significance in the source. The summarising strategies using scripts are the only ones we have tried so far which generalise rather than just select, but this is still not in a wholly constructive way from the actual text data.

7.3 Communicative approaches

We have so far done least on the communicative case, primarily for lack of prior models. Our major effort has been in some joint work with Professor Barbara Grosz, who provided analyses under the Grosz and Sidner (1986) model giving the intentional structure for our test texts. In representation this is the intentional structure recognised by the reader, which may thus refer only to the encapsulated beliefs etc that the writer intends to communicate; and it is a bottom-up strategy in our terms. As with the domain information case, the intentions associated with discourse segments may be distant from their linguistic expression in the text. There is, however, a natural summarising procedure: since the intentional structure for a text is hierarchical, it is possible to take the topmost segment intentions for the summary representation, or the segment material, as shown in Figure 3e.

8 Discussion

The figures in this paper illustrate our investigations only for a single 'noddy' text. We have applied all the representation and summarising strategies to our complete set of ten test texts, including the newspaper ones, which were rather more challenging for our sorts of representation.

Our view of our results is necessarily impressionistic. All the strategies, for both representation and summarising, have *some* plausibility. Thus considering accuracy and coverage respectively, the different sorts of representation have fair accuracy and coverage is reasonable though there are, not surprisingly, holes where sorts are not geared to certain kinds of concept. For instance, the domain representations do not capture the force of the concluding recommendations in the 'Bibliographies' text. The summarising results are a similar mix of success and failure, but as the summaries are all very brief, and there is not much work to be done with such short sources as our test ones, little more can be said than that the strategies have merits. An analogous picture emerges when the different strategies are compared, rather than considered individually: they to some extent pick up the same source content, but also complement one another in capturing or failing to capture particular source items.

With respect to the role of large-scale structure, the summaries we have obtained suggest that this can be a valuable support for summarising, even if we cannot conclude that it is essential. The different sorts of structure vary in their distance from the surface text, but this distance is not tied to the type of information involved, the representation form, or the general complexity of the structure, as sort or instance. We thus cannot yet draw any conclusions about the characteristics of large scale structures that may be important and also about whether simple and shallow strategies can capture, indirectly, enough information about text structure to be useful approaches to summarising.

The work described clearly has many limitations. There is no guarantee that any of the resources used for the representations are in any sense correct; and both the construction and use of the representations has been an individual and fundamentally intuitive operation. Beyond this, we have not considered whether any of the models we have adopted has any real standing, as opposed to mere plausibility. We nevertheless believe that what we have done is a reasonable as well as practicable way to begin our investigation, and that even our informal results are of some suggestive value. These follow from our explicit use of large scale structure, and our comparative approach.

However even on this basis, we have hardly started. We have to try to fill all the slots in our grid, and to explore more alternatives per slot. This applies both to exploiting any one choice of representation in different ways for summarising, and to examining different representation choices per slot. There is furthermore the problem of evaluating the results properly (even when the way they are obtained is taken as sound throughout). As mentioned earlier, there is no such thing as a correct summary for a text, even in a given relatively well-defined context. There can in any case be no serious way of operationally evaluating summaries for such minimal test data as ours. However what can be done is to judge whether the derived summaries seem at least reasonable, or alternatively manifestly inadequate, on a careful point by point analysis comparing source with summary and one summary with another, and to proceed from there in strategy selection and development.

9 Conclusion

This is work in progress. Our intention is therefore to continue along the lines described, especially with more and longer texts, since some of the strategies currently delivering acceptable results may fall apart in more challenging situations. We also need enough data to sort out whether different approaches tend to identify the same source content by different means, or whether they in fact differ, and if so which give better data for summarising; and whether different genres require different treatments. Finally, we want to compare the kinds of approach described, which in principle involve deep text interpretation, with much more shallow ones using only statistical, locational, and other surface cues, but in conjunction with proper sentence processing. From a general discourse analysis point of view it seems any one of our sorts of representation must be inadequate, and that a more comprehensive model of discourse structure is required. Work is in progress on this, but specifically for summarising we believe that more comparative studies of different representations, like those reported here, are the best immediate way forward.

In conclusion, returning to the relation between summarising and indexing, the work described here suggests two specific investigations. One is to see whether the methods of identifying key source content for summarising are of particular merit for indexing; and the other is to see how far the kinds of surface data used in indexing, notably statistical data, can enhance any deep source interpretation required for summarising. Thus one potential study concerns the relation between summarising and indexing, the other the relation between two kinds of information that may be jointly used for both.

10 Acknowledgements

This work was carried out under Grant GR/F 24678, "General techniques for automatic summarising", funded by the UK Science and Engineering Research Council. The detailed project studies were done by Philip Gladwin and John Beaven, and the figures are taken from their analyses. I am also grateful to John Beaven and Professor Barbara Grosz for valuable discussions.

References

[1] Alshawi, H. (Ed) *The core language engine,* Cambridge MA: MIT Press, 1992.

[2] Cremmins, E.T. *The art of abstracting*, Philadelphia: ISI Press, 1982.

[3] DeJong, G.F. *Skimming stories in real time: an experiment in integrated understanding,* Thesis, Yale University, 1979.

[4] Gladwin, P., Pulman, S. and Sparck Jones, K. *Shallow processing and automatic summarising: a first study,* Technical Report 223, Computer Laboratory, University of Cambridge, 1991.

[5] Grosz, B.J. and Sidner, C.L. 'Attention, intention, and the structure of discourse', *Computational Linguistics* 12, 1986, 175-204.

[6] Hahn, U. 'Topic parsing: accounting for text macro structures in full-text analysis', *Information Processing and Management,* 26, 1990, 135-170.

[7] Halliday, M.A.K. and Hasan, R. *Cohesion in English,* London: Longman, 1976.

[8] Hobbs, J.R. *On the coherence and structure of discourse,* Report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.

[9] Kintsch, W. and van Dijk, T.A. 'Toward a model of text comprehension and production', *Psychological Review* 85, 1978, 363-394.

[10] Lehnert, W.G. 'Plot units: a narrative summarisation strategy' in *Strategies for natural language processing* (Ed Lehnert and Ringle), Hillsdale NJ: Lawrence Erlbaum, 1982.

[11] Liddy, E.D. et al 'Development, implementation and testing of a discourse model for newspaper texts', *Proceedings of the ARPA Workshop on Human Language Technology*, Plainsboro NJ, 1993.

[12] Luhn, H.P. 'The automatic creation of literature abstracts', *IBM Journal of Research and Development* 2, 1958, 159-165.

[13] Mann, W.C. and Thompson, S.A. *Rhetorical structure theory: a theory of text organisation,* Report ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.

[14] Maybury, M.T. *Planning multisentential English text using communicative acts,* Thesis, University of Cambridge, 1991; Technical Report 239, 1991.

[15] McKeown, K.R. *Text generation,* Cambridge: Cambridge University Press, 1985.

[16] MUC-4, *Proceedings, Fourth Message Understanding Conference (MUC-4)*, San Mateo, CA: Morgan Kaufmann, 1992.

[17] Paice, C.D. 'Constructing literature abstracts by computer: techniques abd prospects', *Information Processing and Management,* 26, 1990, 171-186.

[18] Rau, L.F. 'Conceptual information extraction and information retrieval from natural language input', *RAIO-88, Conference on User-Oriented, Content-Based, Text and Image Handling*, Cambridge MA, 1988, 424-437.

[19] Rowley, J. *Abstracting and indexing,* London: Bingley, 1982.

[20] Rumelhart, D.E. 'Understanding and summarising brief stories' in *Basic processes in reading,* (Ed Laberge and Samuels), Hillsdale NJ: Lawrence Erlbaum, 1977.

[21] Sidner, C.L. 'Focusing in the comprehension of definite anaphora' in *Computational models of discourse'* (Ed Brady and Berwick), Cambridge MA: MIT Press, 1983.

[22] Sparck Jones, K. 'What sort of a thing is an AI experiment?' in *The foundations of artificial intelligence: a sourcebook* (Ed Partridge and Wilks), Cambridge: Cambridge University Press, 1990.

[23] Sparck Jones, K. *Discourse modelling for automatic summarising*, Technical Report 290, Computer Laboratory, University of Cambridge, 1993; and in press.

[24] Tait, J.I. *Automatic summarising of English texts,* Thesis, University of Cambridge, 1983; Technical Report 47, Computer Laboratory, University of Cambridge, 1983.

[25] Young, S.R. and Hayes, P.J. 'Automatic classification and summarisation of banking telexes', *Proceedings, Second Conference on Artificial Intelligence Applications,* 1985, 402-408.

```
1.  Biographies are the best books.
2.  They are about real things.
3.  They tell a true story.
4.  They are about particular people.
5.  They have lessons for their readers.
6.  Smiles wrote biographies.
7.  They are about successful workers.
8.  These biographies inspire their readers.
9.  Their readers imitate these workers.
10. History books are about real things.
11. They tell true stories.
12. They are about states.
13. They are not about particular people.
14. They do not inspire their readers.
15. Novels tell a story.
16. They are about particular characters.
17. They are not about real things.
18. They are dangerous.
19. Biographies are valuable books.
20. Histories are useful.
21. Novels are deceitful books.
22. Give children biographies.
23. Give them histories.
24. Do not give them novels.
```

FIGURE 1 : Example text, 'Biographies'.

```
   ...
Biographies are valuable books.
Histories are useful books.
Novels are deceitful books.
   ...

LINGUISTIC
   sentence type : declarative, generic ...
   lexical item :  "valuable", "book" ...
   discourse form : parallel sentences, words

WORLD
   books
     type : biography
        property : valuable
     type : history ...

COMMUNICATIVE
   speaker cause hearer believe :
       biography valuable
       history ...
```

FIGURE 2a : Types of information in discourse

```
1 = biography, 2 = history, 3 = novel, 4 = giving
1,1a; 2,2a etc refer to first and second text chunks
S ? = possible summary material ('most strong')

LINGUISTIC
  argument
   comparative description
     description 1  2  3
     description 1a 2a 3a            )
   instruction 4                     ) S ?

WORLD
  books
   type [1 2 3]
    properties 1 1a  2 2a  3 3a
     possible effects 1  2  3 [some]  )
      implied actions 4              ) S ?

COMMUNICATIVE
 S intend H act 4
  S motivate H
   S ensure H know (inform 1  2  3)
   S justify H act (inform 1a 2a 3a)  )
  S instruct H act                   ) S ?

SUMMARIES :

 Linguistic:
   Biographies are valuable, histories useful, novels deceitful. Give
   children biographies and histories, not novels.

 World:
   Biographies are true and inspire, so give them to children. Histories
   are true, so give them. Novels are false, so do not give them.

 Communicative:
   Give children biographies because they are valuable. Give histories
   because they are useful. Do not give novels as they deceive.
```

FIGURE 2b : Types of large-scale structure in discourse
and notional associated summaries

```
Biographies <-  <--------   -----------
   ...       _|Evidence  |Elaboration  |
Smiles <-   -------------              |
   ...  _|Elaboration                  |Evidence
Histories <-   ------------------------+---
   ...     _|Elaboration               |   |Evidence
Novels <-   ---------------------------+---+---
   ...  _|Elaboration                  |   |   |Evidence
Biographies <-------------------------   -+---+---
                                          |   |   |Justification
Histories <--------------------------------   -+---+---
                                              |   |   |Justification
Novels <--------------------------------------   -+---+---
                                                  |   |   |Justification
Give biographies <---------------------------------    |   |
                                                      |   |
Give histories <----------------------------------------    |
                                                          |
Do not give novels <--------------------------------------
```

SUMMARY : nucleus clause of topmost nucleus

Give children biographies. Give children histories. Do not give children novels.

FIGURE 3a : Example linguistic, bottom-up representation and summary (Rhetorical Structure Theory)

```
biographies      state  )       )
   ...           state  )-      )
   ...                  )       )_state-unit      )
smiles           event     )    )                 )
   ...           state  )_ )-   )                 )
    ...                 )  )    )                 )
history books    state          )_state-unit      )
   ...                          )                 )-argument
novels           state          )_state-unit      )
   ...                          )                 )
biographies      epitome        )_epitome         )
   ...                          )                 )
give             recommendation )_recommendation  )
   ...                          )
```

SUMMARY : rule for top node e.g. {epitome} recommendation

Biographies are valuable, histories useful, novels deceitful. Give children biographies, histories, not novels.

FIGURE 3b : Example linguistic, intermediate representation and summary (Text grammar)

```
be (books biographies)
have (lessons readers biographies)
inspire (readers smiles-biographies)
tell (stories history books)
not (inspire (readers history books))
tell (stories novels)
be (valuable books biographies)
be (useful history books)
give (biographies children)
give (histories children)
    .....

SUMMARY : most active entity

        biographies      7
        history books    7
```

FIGURE 3c : Example domain, bottom-up representation and summary (Predication participation)

```
BOOK
subset : BIOGRAPHY
          subject : real things
               ...
          property : true story
               ...
          subset : (BIOGRAPHY)
               author : smiles
               ...

subset : HISTORY
         ...

COMPARISON
       compared : BIOGRAPHY, HISTORY, NOVEL
       parameter : ?
       comparison :  best BIOGRAPHY

GIVING
       recipient : children
       gift : BIOGRAPHY, HISTORY

SUMMARY : instantiated script header predication

 There are biographies, histories, novels with various properties. In
 comparison, biographies are best. Biographies, histories are gifts
 for children.
```

FIGURE 3d : Example domain, top-down representation and summary (Scripts and frames)

```
DS1 [Purpose: convince biographies best, histories alright for children.
              novels not alright
        biographies best ... real ... true ... particular

    DS2 [Purpose: provide illustrative example
            smiles wrote ... workers ... inspire ... imitate
          ] DS2

    DS3 [Purpose: convince "real people" feature important
            histories real ... true ... states ... not people ... not inspire
          ] DS3

    DS4 [Purpose: convince "real" is important
            novels ... story ... characters ... not real ... dangerous
          ] DS4

    DS5 [Purpose: recapitulate important properties
            biographies valuable ... histories useful ... novels deceitful
          ] DS5

        give children biographies ... histories ... not novels
      ] DS1

SUMMRY : topmost segment material

 Biographies are the best books. Biographies are about real things,
 true stories, particular people. Biographies have lessons for readers.
 Give children biographies, histories, not novels.
```

FIGURE 3e : Example text communicative representation
and summary (Intentional structure)