Introduction to "Text Summarization" workshop

Schloss Dagstuhl, December 1993

John Hutchins (University of East Anglia, Norwich, UK)

1. Why summaries?

Why are summaries made at all? The simple answer is that in order to gain access to and control the flood of information, everyone needs to know in brief what is worth reading and what is useful for a particular purpose. Nobody wants to waste time reading what is useless. By giving an overview or outline of content, summaries save readers' time. This is not, we must stress, a new phenomena – it began even before the invention of printing. Scholars in the early Middle Ages found it necessary to compile summaries of what was known. These were the first encyclopedias. With this long history, it is somewhat surprising that the process of summarization has itself been so neglected. Only relatively recently, with access to computers capable of dealing with large textual databases, has there been serious or substantial research in this field. It is to be hoped that this workshop will stimulate further activity in this important field of research.

As an introduction I have attempted to 'summarize' the basic features of this general field.

2. Typology.

A typology of summaries can be made on four sets of parameters:

(a) coverage

(b) informativeness

(c) selectivity

(d) recipients

2.1. *Text coverage*. Summaries can be made of individual texts (documents, speeches, presentations, books) or of collections of texts (proceedings of a debate, arguments of the prosecution and defence in a law court, documents in an archive, collections of papers or books, etc.)

2.2. *Informativeness*. On this parameter we may identify three types: indicative, informative, and evaluative. An informative summary reports on the factual data conveyed or the details of the opinions expressed. An indicative summary states only that certain topics were covered without conveying the precise content of the facts or opinions described in the original text. An evaluative summary locates the content or opinions of the text within the context of other texts treating similar topics.

2.3. *Selectivity*. Summaries can attempt to cover all 'important' aspects of texts (global) or they can report only part of the contents (selective). Typically, selective summaries are made for specific purposes or clientele.

2.4. *Recipients*. Summaries can be made for specific groups of recipients or readers (directed summaries), i.e. they can be targeted to their particular needs – by being evaluative or selective summaries. They are possible only when the specific needs of users are known or predictable. However, summaries can also be for general consumption (undirected), i.e. for use in an information system where the background knowledge of users cannot be predicted. In such cases, summaries are normally indicative. The production of informative summaries intended for 'unpredictable' needs is probably the most difficult of all.

2.5. In theory, all combinations are possible. In practice, some types are found more commonly together than others, e.g. directed summaries are usually evaluative or selective; general summaries are normally indicative or informative.

3. Methodology of summarization.

In the broadest terms, summarization can be regarded as a conjunction (with varying emphases) of the following processes:

(a) selection of what is 'important'

(b) omission of what is 'unimportant'

(c) generalization from the particular and specific

(d) identification of general (global) structures

Methodologically, summaries can build up from the details (of a micro-level) by generalization, selection, and omission, or they can work by extraction from within global frameworks (macro-level).

Summaries differ according to the emphases placed on each of these processes. In evaluative and selective summaries the first two processes (a) and (b) dominate. The dominant process in informative summaries is (c) 'generalization'. By contrast, indicative summaries may not necessarily involve generalization at all.

Automatic summarization was initially based entirely on the first two methods: extraction of what it is hoped is 'important' (on the basis of textual cues or triggers) and thus the omission of everything not extracted. This method is still widely used alone or in conjunction with other methods.

More sophisticated and more recent research on automatic summarization involve 'generalization', i.e. attempts to derive general statements of content from particular textual elements using (typically) semantic/lexical hierarchies, networks, thesauri and macrostructural frameworks, and (in many cases) databases of domain knowledge. Usually, the subject ranges of such experiments are highly restricted, and apparently not easily transferable to other domains.

Attempts to automate indicative summarization are rare. Whereas informative summaries are based on methods which consider the whole text as source data for deriving summaries (whatever the methods used), indicative summaries would probably have to be based on those parts of texts which state the 'topics' (i.e. what the text/sentence, etc. is 'about'). They would ignore those parts which convey what is 'new' ('rheme'). In some early efforts at automatic summarization, the 'topics' of texts were selected from the initial sentences of paragraphs. Although the aim was probably the production of 'informative' summaries, in effect the method was more appropriate to indicative summarization.

A more sophisticated approach to indicative summarization might well involve the identification of thematic progression in texts (i.e. the ways in which sentences relate to each other in terms of thematic and rhematic links) and the identification of 'theme' sentences in paragraphs and in texts as wholes.

4. Purposes and roles of summarization

Some contexts in which summarization is important (and the type(s) which are appropriate) are:

4.1. *Abstracts*

Abstracts are a vital component of communication of research, saving the reading time of individual researchers and improving the control of information, e.g. in computer-based 'free text' document retrieval systems.

In addition, they provide important clues and sources of index terms (keywords) in bibliographic databases. (In practice, most automatic indexing systems are based not on the full texts of documents but on their abstracts - produced manually, of course.)

Abstracts can be informative or indicative. In general, abstracts produced by authors themselves are invariably informative, but also selective. Abstracts produced by others can be either informative or indicative. If intended for a general-purpose database (e.g. bibliography or abstract journal) they are normally non-evaluative and non-selective. If intended for a specific audience, they will normally be both selective and evaluative.

Indicative abstracts are most appropriate in circumstances where the backgrounds of users are unknown. I have argued elsewhere (Hutchins 1977a, 1977b) that indicative abstracts based on the 'thematic' approach (outlined very briefly above) can provide the best 'starting points' for users who are seeking documents in unfamiliar fields of knowledge. In general, such users will know what the ‘themes’ should be but not what the ‘rhemes’ are going to be.

By contrast, an informative abstract is appropriate for the researcher looking for documents treating subjects at the forefront of research – and these will be the subjects mentioned in the ‘rhemes’ of texts. For this purpose the well-established statistical methods of extraction provide a familiar basis.

4.2. *Review articles*

Typically an article reviewing progress in a specific research field covers a wide range of documents. In some cases, only the bare contents of texts are mentioned (indicative); in others, more substance is reported (informative). But most importantly, the review article weighs up the current status and indicates the important contributions (evaluative and selective).

In the present state of knowledge, automatic production of review articles would appear possible only for indicative reports.

4.3. *Encyclopedias*

Encyclopedia articles review the state of the art in a more global fashion. They are evaluative (and almost necessarily selective) summaries of ‘what is known’ about a particular topic (informative). They represent ‘starting points’ for readers, and hence frequently refer to other encyclopedia articles or ‘further reading’ (in this respect they are indicative).

4.4. *Legal summaries*

The trial summaries by lawyers and judges are examples of evaluative summaries of informative nature. The summaries of prosecuting and defending lawyers are necessarily and deliberately evaluative (and normally selective); the summaries of judges when speaking to juries are supposedly neutral, non-evaluative, but equally selective.

4.5. *Journalism*

Summaries are the stock in trade of most journalists. Many newspaper reports are extracts from other texts (e.g. reports of debates and official documents). Most journalist summaries are selective (often evaluative).

4.6. *Market surveys and reports*

These basic information sources for business people are intrinsically compilations of (evaluative) summaries of documentation produced by companies and of evaluations of products.

4.6. *Translation*

While the translating of a text is not as such a summary, it can be argued that what the recipients of translations often need is not a translation of the whole text but something which conveys the basic message (essence) of the text in the unknown language. In other words, they do not want just a translation but also a summary.

The conjunction of machine translation and automatic summarization would be an obvious and valued desideratum for future NLP research. (Whether MT itself offers methods applicable to automatic summarization *per se* is another question.)

In addition, the demand for multilingual access to databases is growing rapidly: MT can offer help in searching databases in unknown languages as well as (obviously) translating the results of searches.

5. Conclusion

In view of the central function of summarization in society it is surprising that it is a topic which has been so often neglected. (Indeed, it could be argued that summarization is central to all communication: every expression of a 'theme' is in essence a 'summary' of some preceding statement – in either the same 'text' or in some other earlier 'text'.)

It is a personal pleasure to me that this workshop is taking place, since it is devoted to a topic in which I have had an interest for many years (although without being actively involved in research as such), principally in relation to the interface between machine translation and information retrieval. This workshop is bringing together researchers from a wide range of disciplines, and I am sure that discussions in the coming week will bear fruit in both the immediate and more distant future.

6. References

Hutchins, W.J. (1970) Linguistic processes in the indexing and retrieval of documents. *Linguistics* 61, 1970, 29-64

Hutchins, W.J. (1977a) On the problem of 'aboutness' in document analysis. *Journal of Informatics* 1(1), April 1977, 17-35

Hutchins, W.J. (1977b) On the structure of scientific texts. *UEA Papers in Linguistics* 5, September 1977, 18-39.

Hutchins, W.J. (1985) Information retrieval and text analysis. In: T.A.van Dijk (ed) *Discourse and communication* (Berlin: de Gruyter, 1985); 106-125.