Automatic Abstracting; towards a Text Based Generation.

Horacio Saggion*
Groupe Scriptum, Laboratoire Incognito
Département d'Informatique et Recherche Opérationnelle
Université de Montréal
CP 6128, Succ Centre-Ville
Montréal, Québec, Canada, H3C 3J7
saggion@iro.umontreal.ca

Abstract

We review the basic concepts involved in automatic abstracting and we discuss some aspects we will address in our work which will focus on natural language generation of abstracts.

1 Introduction

Abstracts are pragmatic texts that have been studied in Computational Linguistics since the 1960s. Our own interest in abstracts started in 1994, when we studied automatic analysis of abstracts in Portuguese [Saggion, 1995]. We are now interested in the process of producing abstracts from scientific and technical articles. Our objective is to produce them automatically using Natural Language Generation theories and techniques which have been until now practically ignored in automatic abstracting.

Nowadays, the overwhelming quantity of information and the need to access the essential content of documents accurately to satisfy users' demands have made automatic abstracting a major research area. The 1993 Seminar on Summarizing Text for Intelligent Communication, the 1995 Information Processing & Management Special Issue in Automatic Summarization and the 1997 Workshop on Intelligent Scalable Text Summarization mark a renewed interest in text summarization.

As a human activity, the production of summaries is an interesting and, unfortunately, not well studied process: a source text is read and understood to recognize its content which is then compiled in a concise text. Abstracts of research articles are produced by their author or by professional abstractors working for abstracting services. Many readers depend on the abstract of a research paper to decide whether or not to read the full paper. The idea of producing abstracts automatically is not new. Several studies and practical solutions to the problem have been proposed, but they are still not satisfactory: an abstract so produced does not meet the quality of a human produced text.

---

*Member of Departamento de Computación de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, Argentina.

In the rest of the paper we introduce the concept of an abstract and its structure. Then the processes of human and automatic abstracting are presented, and finally we discuss some issues in the field. In each section we introduce research questions for discussion in the workshop.

2 Abstracts

According to several studies in Information Science [Borko and Bernier, 1975, Rowley, 1982] an abstract is a concise and accurate representation of the contents of a document. It is the reflection of the macrostructure of the text [Hutchins, 1987]. Linguistics studies [Swales, 1990, Bhatia, 1993] and Information Science studies [Borko and Bernier, 1975, Cremmins, 1982] have shown that an abstract is a text of a recognizable genre with a very specific purpose: to give the reader an exact and concise knowledge of the full document.

Several norms have been proposed suggesting how an abstract must be written: [ABNT, 1987] for Portuguese, [AFNOR, 1984] for French, and [ANSI, 1979] for English. Some journals adopt these norms for the writing of abstracts but usually the norms act only as recommendations for the authors and they are not enforced as such.

Each type of abstract is associated with a specific communicative purpose. Two main types of abstracts can be identified: *indicative* and *informative* abstracts. The purpose of an informative abstract is to provide information from the original document (e.g. "the author concluded that..."); an indicative abstract describes information that can be found in the original without actually giving it (e.g. "conclusions are presented"). Abstracts also depend on the particular type of document to be abstracted. Articles centered on a particular subject promote informative abstracts while longer documents covering many subjects lend to indicative ones.

An automatic generation system must decide the kind of abstract to generate. The choice is in part influenced by the communicative purpose and the intended audience.

3 Structure of an Abstract

Being a textual genre, an abstract does not only have a typical organization but also the information included should be recognizable. Abstracts of research articles are theme oriented and would generally include the following categories of information: *Purpose, Methodology, Results* and *Conclusions* associated with typical aspects of the scientific research; what the author did, how the author did it, what the author found and what the author concludes. Rowley [Rowley, 1982] suggests that the abstract usually follows the style and arrangement of information of the parent document.

Liddy [Liddy, 1991] studied research abstracts to identify their information structure and to base the recognition of their structure on cue words. Liddy's model includes three typical levels of information. The most representative level, called the *Prototypical Structure*, includes the information categories: *Subjects, Purpose, Conclusions, Methods, References* and *Hypotheses*. The other two levels are the *Typical Structure* and the *Elaborated Structure* which include less frequent information. We think that these components and categories would be useful for the extraction of important information from the original document in order to automatically produce an abstract.

The domain of discourse also influences the structure of the abstracts. Natural Sciences and Social Sciences abstracts have different structure. The first are highly structured and follow the pattern *Introduction, Method, Results* and *Discussion*, while the latter have a more idiosyncratic structure [Milas-Bracović and Zajec, 1989].

Structured abstracts [Hartley et al., 1996] have a paragraph structure each being characterized by the use of sub-headings signaling their information. They are used in medicine and seem to be more useful than standard abstracts in the search for information. Problem structured abstracts are produced for papers reporting the solution of a scientific problem and they are characterized by the following information: *Document Problem, Problem Solution, Tests, Related Problems* and *Content Elements* [Trawiński, 1989].

Anticipating the discourse structure for the abstract is an important issue to be investigated. A successful automatic system would select the structure which aid the reader in understanding the message.

4 Abstracting Procedures

Methods, rules and recommendations have been proposed for the process of human abstracting.

Questions:

> How these methods can be applied for the automatic generation of abstracts? Is it necessary to concentrate on the final product (the abstract) or on the process that generates it?

The process of abstracting has been investigated as a human and professional activity. Most studies agree on a two stage logical account for describing the process of human abstracting: the analytical and the synthetic stage [Borko and Bernier, 1975, Cremmins, 1982, Pinto Molina, 1995]. In the analytical stage the salient facts of the text are obtained and condensed. In the synthetic stage, the text of the abstract is produced.

Several factors influence the process of summarizing and the information to be included in the abstract [Cremmins, 1982, Spark Jones and Endres-Niggemeyer, 1995]. Sparck Jones and Endres-Niggemeyer categorize these factors as:

Input factors : characteristics of the input material such as form and type of the input text (e.g. research papers promote informative abstracts and proceedings of conferences and documents covering several subjects promote indicative abstracts).

Purpose factors : such as function (e.g. to inform the reader or to alert the reader) and audience (e.g. specialist, non-specialist).

Output factors : characteristics of the output material (e.g. running vs. itemized text).

As far as content selection is concerned, the type of abstract provides the basis for the inclusion of material. An indicative abstract should include information on purpose and methodology and an informative abstract should include information on purpose, methodology, results and conclusions [Russell, 1988]. An input factor to consider is the information brought by the author. Writers use linguistics features to stress their main points; expressions like "it is necessary", "it is important" are cues that guide the abstractors in the selection of information. Another factor in the selection of information is the abstractor's knowledge of the field and the policy of the abstracting service. The knowledge of the abstractor determines that certain material should be omitted in the abstracts, i.e. well known techniques and procedures. The objective of the abstracting organization also influences the selection of information. For example, chemical abstracting services would probably report only the chemical information that appear in a biochemical work [Borko and Bernier, 1975].

As far as the overall procedure is concerned, Cremmins [Cremmins, 1982] proposes a three stage analytical reading procedure: *the retrieval reading stage* responsible for the selection of relevant information *the creative reading stage* in which the information is organized and a draft written and finally *the critical reading stage* in which the draft is edited for unity and conciseness. Rules are given for the selection of material and the structuring and writing of the abstract.

Pinto Molina [Pinto Molina, 1995] suggests that the analytical stage is composed of two sequential processes which operate after reading and understanding: selection (deletion of repeated, little relevance and irrelevant units) and interpretation (assigning meaning).

It was argued that human abstractors rely on cue words, titles and sub-titles and indicative phrases in order to identify relevant information [Cremmins, 1982]. Several studies were carried with professional abstractors in order to determine how they produce an abstract. The plans and strategies that professional abstractors follow in their natural environment were investigated by introspection and case-based studies [Endres-Niggemeyer et al., 1991, Endres-Niggemeyer et al., 1995]. Typical steps for an abstractor to use were defined, the process of abstract production was modeled and a computational simulation of the working environment was produced. It was observed that for writing the "topical sentence" of an abstract the professional would use just the introductory part of the material and the information enhanced by the author.

Questions:

> How can the steps that an abstractor follows be modeled and implemented to obtain an automatic system? What is the relation between the knowledge of the abstractor and the process of selection of information? How is the knowledge used?

5 Automatic Abstracting

Two approaches are generally considered in automatic abstracting: the *statistical* and the *understand and generate* approaches [Johnson, 1995]. The statistical approach is concerned with the problem of extracting the salient information of the text using *word frequency* and *term distribution*. This is based on the idea that the author of a document usually expresses the essential content of the text using words that are much more frequent in their text than in an usual text [Luhn, 1958]. This approach also includes cue methods which rely on cue words to mark positive, negative or neutral information, title methods that select sentences containing words that appeared in the title, and location methods that focus on sentences in first or last positions in a paragraph which are considered topical. The identification of text positions which carry rich semantic information continues to be investigated nowadays [Lin and Hovy, 1997].

In the statistical approach, the synthetic stage consists of the concatenation of the extracted sentences, so the result is not an abstract but an extract of the original text. This often produces an abstract of poor quality lacking cohesion and coherence. In order to improve the quality of the extract, attention was given to sentences containing anaphoric expressions [Mathis and Rush, 1985, Paice, 1990]. However, the idea of producing extracts without a deep analysis of the source text is still studied and applied [Brandow et al., 1995, Maybury, 1995, Lehman, 1997] mainly for its ease of implementation.

The understand and generate approach was studied in restricted domains such as understanding and summarization of stories and news [Alterman, 1992]. There seems to be few studies in the field of Natural Language Generation. The ones we found are linguistics approaches to the problem such as the use of Centering Theory [Hoffman, 1996] and the use of Coherence and Rhetorical Relations [Rino and Scott, 1994]. Both address the issue of selecting proposition from the text representation in order to obtain its essential content. In the first approach, segments of text (sets of propositions) are selected based on the most salient entities of the text; elaborations between propositions are then deleted and remaining propositions are considered as the basis for the summary. In the second approach heuristics are used to delete optional material from the relations that appear in the representation (i.e. the satellite part of a rhetorical relation), in this way essential information is not omitted. For linguistic realization it was noted that human abstracts are propositionally dense and contain syntactic forms which contribute to compaction [Kaplan et al., 1994]. This issue was addressed in the domain of summarization of sports events and in the domain of automated documentation of telephone planning engineer activities [McKeown et al., 1995], but not in text summarization. Little attention was paid to the problem of selecting linguistics expressions to convey the message.

Questions:

> Is the selection constrained by the source text? Are there particular syntactic structures to be used?

6 Discussion

In this paper, we showed several issues in automatic abstracting. As a human activity the process of summarization was studied in Artificial Intelligence, and some proposals emerged in restricted domains to demonstrate the understanding of an intelligent agent.

Several works aim at the production of an abstract without applying the steps and knowledge that affect the abstracting operation and its product. The resulting text does not meet the demands of the intended audience because it lacks cohesion and coherence. We saw that the generation of summaries is a difficult process affected by several factors. The exact relation between the factors and the operation and the abstracts remains unknown. In Figure 1 we present the main sources of knowledge in the process of abstracting.

Abstracts have been investigated as independent texts, but their relation with source text needs to be addressed. We think that the organization and structure of the source influences the deriva-

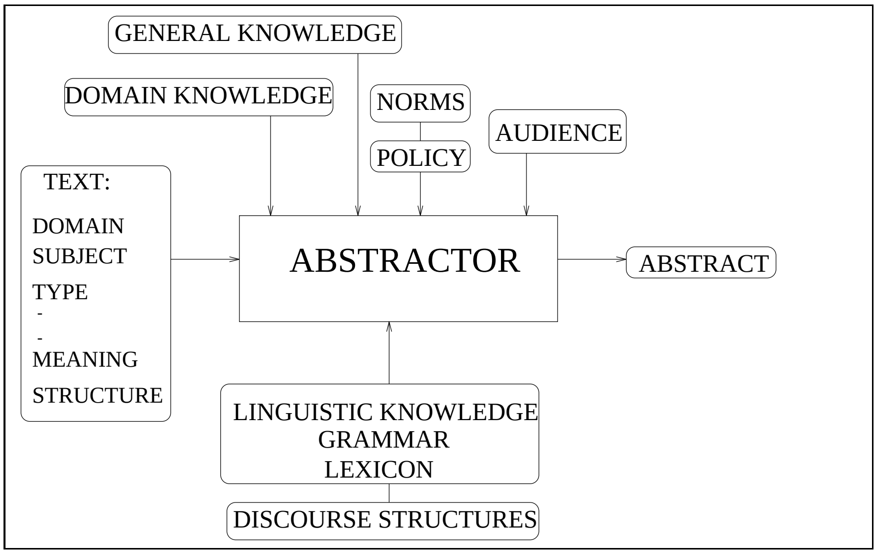

Figure 1: Sources of Knowledge in the Process of Abstracting

tive text.

Questions:

Is it necessary to read and understand the full document in order to obtain its content? If not, which parts of the documents are candidates for deeper analysis? What are appropriate measures of relevance for deciding which propositions to select (audience, knowledge of the abstractor, abstracting policy, etc)?
Once a pool of propositions has been selected what logical operations should be applied in order to compress the information (abstraction, generalization, deletion, etc)? To do so, how to recognize irrelevant information or repeated information? How to identify information on purpose, methodology, results and conclusion?
Are the statistical techniques candidates appropriate for the process of selection in a NLG environment? Which are the more appropriate?
What particular discursive structures exist for abstracts? Are there particular syntactic structures for abstract's sentences? Is the lexicon of the abstract completely determined by the lexicon of the source document? Which linguistic devices a language offers to convey the information succinctly? And finally, how can we evaluate an automatic abstract?

Acknowledgments

I would like to thank my adviser, Prof. Guy Lapalme for encouraging me to present this work in progress. This work is supported by Agence Canadienne de Développement International (ACDI) and Ministerio de Educación de la Nación de la República Argentina, Resolución 1041/96.

List of questions raised in the paper

1. Is it necessary to concentrate on the abstract or on the process which generates the abstract?

2. How can the steps an abstractor follows be modeled and implemented to obtain an automatic system?

3. What is the relation between the knowledge of the abstractor and the process of selection of information?

4. How is the knowledge used?

5. Is the selection of information constrained by the source text?

6. Is it necessary to read and understand the full document in order to obtain its content?
7. If not, which parts of the documents are candidates for deeper analysis?

8. What are appropriate measures of relevance for deciding which propositions to select?

9. Once a pool of propositions has been selected what logical operations should be applied in order to compress the information (abstraction, generalization, deletion, etc)?
10. To do so, how to recognize irrelevant information or repeated information?
11. How to identify information on purpose, methodology, results and conclusion?

12. Are the statistical techniques candidates appropriate for the process of selection in a NLG environment?
13. Which are the more appropriate?

14. What particular discursive structures exist for abstracts?
15. Are there particular syntactic structures for abstracts sentences?
16. Is the lexicon of the abstract completely determined by the lexicon of the source document?
17. Which linguistic devices a language offers to convey the information succinctly?

18. And finally, how can we evaluate an automatic abstract?

References

[ABNT, 1987] ABNT (1987). *Resumos.* Associação Brasileira de Normas Técnicas.

[AFNOR, 1984] AFNOR (1984). *Recommandations aux Auteurs des Articles Scientifiques et Techniques pour la Rédaction des Résumés.* Association Française de Normalisation.

[Alterman, 1992] Alterman, R. (1992). Text summarization. In Shapiro, S., editor, *Encyclopedia of Artificial Intelligence*, volume 2, pages 1579–1587. Jonh Wiley & Sons, Inc.

[ANSI, 1979] ANSI (1979). *Writing Abstracts.* American National Standards Institute.

[Bhatia, 1993] Bhatia, V. (1993). *Analysing Genre. Language Use in Professional Settings.* Longman.

[Borko and Bernier, 1975] Borko, H. and Bernier, C. (1975). *Abstracting Concepts and Methods.* Academic Press.

[Brandow et al., 1995] Brandow, R., Mitze, K., and Rau, L. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management.*

[Cremmins, 1982] Cremmins, E. (1982). *The Art of Abstracting.* ISI PRESS.

[Endres-Niggemeyer et al., 1995] Endres-Niggemeyer, B., Maier, E., and Sigel, A. (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing & Management*, 31(5):631–674.

[Endres-Niggemeyer et al., 1991] Endres-Niggemeyer, B., Waumans, W., and Yamashita, H. (1991). Modelling summary writting by introspection: A small-scale demonstrative study. *Text*, 11(4):523–552.

[Hartley et al., 1996] Hartley, J., Sydes, M., and Blurton, A. (1996). Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science*, 22(5):349–356.

[Hoffman, 1996] Hoffman, B. (1996). Summarization: an application for NL generation. In *INLG'96 - Demonstrations and Posters*, pages 37–40.

[Hutchins, 1987] Hutchins, J. (1987). Summarization: Some problems and methods. In Jones, K., editor, *Meaning: The Frontier of Informatics*, volume 9, pages 151–173. Aslib.

[Johnson, 1995] Johnson, F. (1995). Automatic abstracting research. *Library Review*, 44(8):28–36.

[Kaplan et al., 1994] Kaplan, R., Cantor, S., Hagstrom, C., Kamhi-Stein, L., Shiotani, Y., and Zimmerman, C. (1994). On abstract writing. *Text*, 14(3):401–426.

[Lehman, 1997] Lehman, A. (1997). Une structuration de texte conduisant à la construction d'un système de résumé automatique. In *Actes des Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, pages 175–182.

[Liddy, 1991] Liddy, E. (1991). The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

[Lin and Hovy, 1997] Lin, C. and Hovy, E. (1997). Identifying topics by position. In *Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics.

[Luhn, 1958] Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

[Mathis and Rush, 1985] Mathis, B. and Rush, J. (1985). Abstracting. In Dym, E., editor, *Subject and Information Analysis*, volume 47 of *Books in Library and Information Science*, pages 445–484. Marcel Dekker, Inc.

[Maybury, 1995] Maybury, M. (1995). Generating summaries from event data. *Information Processing & Management*, 31(5):735–751.

[McKeown et al., 1995] McKeown, K., Robin, J., and Kukich, K. (1995). Generating concise natural language summaries. *Information Processing & Management*, 31(5):702–733.

[Milas-Bracović and Zajec, 1989] Milas-Bracović, M. and Zajec, J. (1989). Author abstracts of research articles published in scholarly journals in Croatia (Yugoslavia): An evaluation. *Libri*, 39(4):303–318.

[Paice, 1990] Paice, C. (1990). Constructing literature abstracts by computer: Technics and prospects. *Information Processing & Management*, 26(1):171–186.

[Pinto Molina, 1995] Pinto Molina, M. (1995). Documentary abstracting: Towards a methodological model. *Journal of the American Society for Information Science*, 46(3):225–234.

[Rino and Scott, 1994] Rino, L. and Scott, D. (1994). Automatic generation of draft summaries: Heuristics for content selection. Technical Report ITRI-94-8, Information Technology Research Institute.

[Rowley, 1982] Rowley, J. (1982). *Abstracting and Indexing*. Clive Bingley, London.

[Russell, 1988] Russell, P. (1988). *How to Write a Précis*. University of Ottawa Press.

[Saggion, 1995] Saggion, H. (1995). Análise automática de sumários em língua portuguesa: uma aproximação ao tratamento da estructura de um texto. Master's thesis, IMECC, Universidade Estadual de Campinas.

[Spark Jones and Endres-Niggemeyer, 1995] Spark Jones, K. and Endres-Niggemeyer, B. (1995). Automatic summarizing. *Information Processing & Management*, 31(5):625–630.

[Swales, 1990] Swales, J. (1990). *Genre Analysis. English in Academic and Research Settings.* Cambridge. Applied Linguistics.

[Trawiński, 1989] Trawiński, B. (1989). A methodology for writing problem structured abstract. *Information Processing & Management*, 25(6):693–702.