The generation of abstracts by selective analysis

Horacio Saggion and Guy Lapalme
RALI
Département d'Informatique et Recherche Opérationnelle
Université de Montréal
CP 6128, Succ Centre-Ville
Montréal, Québec, Canada, H3C 3J7
Fax: +1 514-343-5834
{saggion,lapalme}@iro.umontreal.ca

Abstract

We describe work in progress in the generation of user-oriented abstracts which we define as abstracts being indicative in the content of the overall document and informative in the specific information the actual user is interested in. We generate the indicative part of the abstract using indicative sentences of the parent document. Additional information is extracted and integrated to the abstract using text spans covering the topics of the indicative material.

Introduction

An abstract is a text of a recognizable genre with a very specific purpose: to give the reader an exact and concise knowledge of a document. Abstracts of research articles are produced by their author or by professional abstractors working for abstracting services. 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"). Most studies agree on a two stage logical account for describing the human production of abstracts: the analytical stage in which the salient facts of the text are obtained and condensed a the synthetic stage in which the text of the abstract is produced. Several factors influence the process of summarizing: input factors such as form and type of the input text; purpose factors such as function and audience; and output factors such as the characteristics of the output material.

In the process of automatic abstracting several methodologies have been proposed: inference from raw text (DeJong 1982); rhetorical analysis (Miike *et al.* 1994; Marcu 1997); semantic analysis of a conceptual text representation and mapping to discourse structure (Rino & Scott 1996); selection of phrases using stylistic information, and instantiation of templates (Paice & Jones 1993); statistical analysis (Kupiec, Pedersen, & Chen 1995; Brandow, Mitze, & Rau 1995); and selection of phrases using heuristic rules and corpus analysis (Sharp 1989; Lehman 1997). Very few consider the audience in the process of abstracting and none of them consider the issue of linguistic realization of the abstract.

The objective of our work is to produce abstracts of technical and scientific documents using natural language generation techniques. We are investigating the use of specific parts of the source text in order to obtain the propositional content that will depend on the source text, the function of the abstract and the user interest. We aim at the production of user-oriented abstracts which we define as abstracts being *indicative* in the content of the overall document and *informative* in the specific information the actual user is interested in.

Selective Analysis and Generation

We have manually analyzed a corpus which consists of pairs of abstract and parent document, the abstracts were produced by professional abstractors and the parent document refers to the document used to produce the abstract. We have already collected and analyzed 25 pairs of documents and we are currently collecting the rest of our corpus (around 100 pairs of documents). We use as source for the abstracts the following journals: Library & Information Science Abstracts (LISA), Information Science Abstract (ISA), and Computer Abstracts. We want to identify which specific parts of the parent document are used to produce the abstract and we are looking at the operations to apply to the parent document in order to obtain a concise indicative text. For each sentence in the abstract we look for a match in information in the parent document. We scan the following parts of the parent document:

the title, the author abstract, the first and last sections, the titles and subtitles and the captions of tables and figures. We then construct a table which relates the sentences of the abstract to the information in the parent document. The analysis of the corpus indicates that the information in the professional abstract is extracted from the author's abstract in 18.42%, from the first section of the parent document in 31.58%, from the titles and captions in 38.82%, from the last section in 4.60% and from other parts of the document in 6.58%. The extracted information is generally reported in the professional abstract in indicative form. Table 1 shows the result of the analysis of one item of the corpus. This example shows that in order to produce the abstract the introductory part of the document was used. Another characteristic of this example is that the sentences of the original text which contains the information appearing in the abstract are indicative sentences, they contain the following indicative expressions: *a research project ... is investigating, the aim is, currently work is focused on, this has resulted in.* Our objective is to use some indicative text spans of the parent document as a starting point for generating an indicative-informative abstract which contains more information than the indicative sentences.

Our process of automatic abstracting is composed of the following steps:

- *indicative selection*: indicative sentences are extracted from the parent document in order to produce the propositional content for an indicative abstract.

- *indicative summarization*: from the pool of indicative propositions an indicative abstract is produced and presented to the user. The abstract includes some topics that will then be informatively extended upon user demand. The phrases in bold font in column two of Table 1 form an indicative abstract which includes the following topics: "a research project currently in progress", "geographic information systems", "the temporal aspects of the spatio-temporal reasoning techniques" and "a tesseral temporal reasoning system".

- *informative selection*: using the content of the indicative material text spans are selected for analysis.

- *informative summarization*: the user will select from the indicative abstract some topics to expand. An informative abstract will be produced using the text spans associated to the selected topics. In Figure 1 an informative abstract is shown for the example in Table 1, in this case the topics expanded are "geographic information systems", "the temporal aspects of the spatio-temporal reasoning techniques" and "a tesseral temporal reasoning system". The information about those topics was found in the first and second sections of the parent document.

> A geographic information system (GIS) is a computer system that contains spatially referenced data that can be analysed and converted to information for specific set of purposes or applications. A research project currently in progress is investigating the application of knowledge based system techniques to GIS. Current work is focused on the temporal aspects of the spatio-temporal reasoning techniques to be applied to GIS. Temporal reasoning is concerned with the deduction of the interrelationships between events displaced through time. The result is a tesseral temporal reasoning system which offers the advantage that it is directly compatible with existing GIS technology. For the tesseral temporal reasoning system described here we reason about events which have temporal attributes associated with them (for most applications they will also have non-spatio-temporal attributes).

Figure 1: Informative abstract manually produced integrating some text spans to the indicative abstract

Following this methodology several abstract can be obtained from a single text depending on the user's interests to produce an abstract such as the one given in Figure 1. The text in Figure 1 has some problems such as the use of the verb form "we reason" in the last sentence in an impersonal text. These problems can be solved using deep analysis and natural language generation techniques.

Discussion

In this paper we have described a new methodology for the process of automatic abstracting. Two aspects are new in our approach: the consideration of the reader interests as essential in the selection of the information to be conveyed and the generation of a new text using natural language generation techniques.
We are currently investigating the following issues: the production of an indicative abstract using indicative phrases of the parent document; the selection of text spans for the topics using information extraction techniques; the syntactic and rhetoric analysis of the text spans; and the generation of the final informative abstract using natural language generation techniques.

Professional Abstract	Parent Document	Position/Type
Research is being carried out at the University of Liverpool, UK, to investigate the application of knowledge based systems techniques to geographic information systems (GIS).	**A research project currently in progress** at the University of Liverpool, the dynamic Geographic Knowledge Based Information System (dGKBIS) project, **is investigating the application of KBS techniques to GIS.**	1st/Introduction
The current focus is on developing spatio-temporal reasoning techniques which can be applied to GIS.	The aim is to develop spatio-temporal reasoning techniques which can be applied to GIS and to enhace the potential of such systems...**Currently work is focused on the temporal aspects of the spatio-temporal reasoning techniques to be applied to GIS.**	1st/Introduction
A tesseral temporal reasoning system has been designed, based on tesseral addressing and using tesseral arithmetic.	**This has resulted in a tesseral temporal reasoning system**, based on tesseral addressing and using tesseral arithmetic, **which offers the advantage that it is directly compatible with existing GIS technology.**	1st/Introduction
It offers the advantage that it is compatible with existing GIS technology.		

Table 1: Professional Abstract and Parent Document. Column one contains the sentences of the professional abstract. Column two contains the information in the parent document Column three contains the number and name of the section where the information was found in the parent document. Professional Abstract: Lisa Abstract 3090. Parent Document: Temporal reasoning using tesseral addressing : towards an intelligent environmental impact assessment system. F.Coenen *et al.*. Knowledge-Based Systems, 9(5), p287-300.

Acknowledgments

The first author is supported by Agence Canadienne de Développement International (ACDI), Departamento de Computación (UBA) and Ministerio de Educación de la Nación de la República Argentina, Resolución 1041/96.

References

Brandow, R.; Mitze, K.; and Rau, L. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management* 31(5):675–685.

DeJong, G. 1982. An overview of the frump system. In Lehnert, W., and Ringle, M., eds., *Strategies for Natural Language Processing.* Lawrence Erlbaum Associates, Publishers. 149–176.

Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A trainable document summarizer. In *Proc. of the 18th ACM-SIGIR Conference*, 68–73.

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*, 175–182.

Marcu, D. 1997. From discourse structures to text summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 82–88.

Miike, S.; Itoh, E.; Ono, K.; and Sumita, K. 1994. A full-text retrieval system with a dynamic abstract generation function. In Croft, W., and van Rijsbergen, C., eds., *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 152–161.

Paice, C., and Jones, P. 1993. The identification of important concepts in highly structured technical papers. In Korfhage, R.; Rasmussen, E.; and Willett, P., eds., *Proc. of the 16th ACM-SIGIR Conference*, 69–78.

Rino, L., and Scott, D. 1996. A discourse model for gist preservation. In Borges, D., and Kaestner, C., eds., *Proceedings of the 13th Brazilian Symposium on Artificial Intelligence, SBIA'96*, Advances in Artificial Intelligence, 131–140. Springer.

Sharp, B. 1989. *Elaboration and testing of new methodologies for automatic abstracting.* Ph.D. Dissertation, The University of Aston in Birmingham.