Where does Information come from?
Corpus Analysis for Automatic Abstracting

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 report on our study of a corpus of abstracts and parent documents to determinate which structural parts of the parent document are used to extract useful information for an abstract. The results give us a sound basis for automatic abstracting of research articles. Our method for automatic abstracting, called selective analysis, is intended to produce user-oriented abstracts which are indicative in the essential content of the document and informative in the user's interest.

1 Introduction

An abstract aims at giving the reader an exact and concise knowledge of the parent document. Abstracts of research articles are produced by their author or by professional abstractors working for abstracting services. In documentary abstracting, two main types of abstracts are identified: *indicative* abstracts which point to information and *informative* abstracts which give detailed information about the findings of the work [Bernier, 1985].

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 and the *synthetic* stage in which the text of the abstract is produced [Pinto Molina, 1995]. Abstracting manuals [Cremmins, 1982, Borko and Bernier, 1975] give indications about grasping the "essential" content of a document and writing the abstract such as "scan the document to get some idea of the subject matter", "mark the material containing information on purpose, method, findings, conclusion and recommendation", "write a concise unified abstract". These instructions are very conceptual and require good abstracting skills in order to be operationalized and thus are very difficult to implement in an automatic procedure.

In this article we report on our study of a corpus of abstracts and parent documents to determine which structural parts of the parent document are used in the analytical stage. We use professional abstracts instead of author's abstracts because they are better structured in content and form and because they are produced from the reading of the document following specific strategies such as the identification of useful information using lexical clues. We undertake this study in order to use the results of the analysis as a sound basis for automatic abstracting of research articles.

In the rest of the paper we describe the corpus under analysis and the results that were obtained in this stage of the research. We also present an overview of our approach to the production of abstracts by selective analysis and a discussion of the main problems we are facing.

2 The Corpus

Our corpus is composed of 100 items each composed of a professional abstract and its parent document. We used as source for the abstracts the journals Library & Information Science Abstracts, Information Science Abstracts and Computer Abstracts while the parent documents were found in journals of Computer Science (CS) and Information Science (IS) such as AI Communications; AI Magazine; American Libraries; Annals of Library Science & Documentation; Artificial Intelligence among others (a total of 44 publications where examined). The professional abstracts contain 3 sentences on the average, with a maximum of 7 and a minimum of

1. Most of the articles are structured in sections (97 documents). We examined 62 documents in CS and 38 in IS. Some of the articles containing author provided abstract. The documents are 7 pages on the average, with a minimum of 2 and a maximum of 45.

We manually aligned each sentence of the corpus with one or more elements of the parent document. We looked for a match between the information in the professional abstract and the information in the parent document. The structural parts of the parent document we examined are: the title of the parent document, the author abstract, the first section, the last section, the subtitles and captions of tables and figures. When the information is not found, we look in other parts of the parent document. In Table 1, we present one such alignment[1] (P/T indicates the position of the information and its type).

Ex.	**Professional Abstract**	**Parent Document**	**P/T**
(1)	Presents the results of an empirical study that investigates the movement characteristics of a multi-modal mouse - a mouse that includes tactile and relevance feedback.	In this paper, we present the results of an empirical study that investigates the movement characteristics of a multi-modal mouse - a mouse that includes tactile and force feedback.	1st/Intr.
(2)	Uses a simple target selection task while varying the target distance, target size, and the sensory modality.	Our experiment used a simple target selection task while varying the target distance, target size, and the sensory modality.	1st/Intr.
(3)	Significant reduction in the overall movement times and in the time taken to stop the cursor after entering the target were discovered, indicating that modifying a mouse to include tactile feedback, and to a lesser extent, force feedback, offers performance advantages in target selecting tasks.	We found significant reductions in the overall movement time and in the time to stop the cursor after entering the target.	-/Abs.
		The results indicate that modifying a mouse to include tactile feedback, and to a lesses extend, force feedback, offers performance advantages in target selection tasks.	-/Abs.

Table 1: *Item of Corpus*

The 3 sentences of the professional abstract were aligned with 4 elements of the parent document, 2 in the introduction and 2 in the author provided abstract. In this example the information of the abstract was "litteraly" found in the parent document. The differences between the sentences of the professional abstract and those of the parent document are the

[1] Professional Abstract: Library & Information Science Abstract 3024 and, Parent Document: "Movement characteristics using a mouse with tactile and force feedback" International Journal of Human-Computer Studies, 45(5), Oct'96 p483-93.

tenses of the verbs ("Presents" vs. "We present" in alignment (1)), the verbs themself ("were discovered" vs. "We found" in alignment (3)), the impersonal vs. personal styles ("Uses" vs. "Our experiment used" in alignment (2)) and the use of markers in the parent document ("In this paper," in alignment (1)).

Other examples of aligned sentences which come from several items of the corpus are shown in Table 2. They give an insight about the alignments of sentences in the abstract with each type of structural element in the parent document.

3 Results

The 309 sentences of the professional abstracts were manually aligned with 568 elements in the parent document. We were not able to align 6 sentences of the professional abstract. Other studies have already investigated the alignment between sentences in the abstract and sentences in the parent document. Kupiec et *al.* [Kupiec et al., 1995] report on the semi-automatic alignment of 79% of sentences of professional abstracts in a corpus of 188 documents with professional abstracts. Using automatic means it is difficult to deal with conceptual alignments that appeared in our corpus such as the relation we show in example (6). Teufel and Moens [Teufel and Moens, 1998] report on a similar work but this time on the alignment of sentences from author provided abstracts. They use a corpus of 201 articles obtaining only 31% of alignable sentences by automatic means. No information is given about the distribution of the sentences in structural parts in the parent document.

In Table 3, we present the distribution of the sentences in the parent documents which were aligned with the professional abstracts in our corpus. We consider all the structured documents of our corpus (97 documents). The first three columns contains respectively the information for all the documents, for documents with author abstract and for documents without author abstract (the information is given in total of elements and in percent). The last column indicates the average of the information. We found that 72% of the information for the analytical stage comes from the following structural parts of the parent document: the title of the document, the first section, the last section and the subtitles and captions. It is important to note that

Ex.	Professional Abstract	Parent Document	P/T
(4)	Presents a more efficient Distributed Breadth-First Search algorithm for an asynchronous communication network.	Efficient distributed breadth-first search algorithm	-/Title
		In this paper we have presented a more efficient distributed algorithm which construct a breadth-first search tree in an asynchronous communication network	Lst/ Concl.
(5)	Summarizes and make suggestions for future research.	Summary and future research	-/Subt.
(6)	Gives a formal description of the problem of optimal pruning of decision trees...	Pruning a decision tree is the process of replacing some of the subtrees of DT by leaves.	1st/Intr.
(7)	Compares France and US perspectives.	Comparison between the French and US perspectives	-/Capt.
(8)	It began in Feb 95, and comprises 4 phases: management, research, technical development and evaluation.	The DECIMAL project commenced in February 1995 and will run for two years.	4th/-
		The project involves three practical phases (in addition to a management 'phase'): research, technical development and evaluation.	4th/-
(9)	Shows how relation algebras can be used to handle interval reasoning.	The idea in this paper is to see how relation algebras can be used to handle interval reasoning.	1st/Intr.
(10)	Analyzes the complexity of the algorithm, and gives some examples of performance on typical networks.	We analise the complexity of our algorithm, and give some examples of performance on typical networks.	1st/Intr.
(11)	Investigates the phase transition in binary constraint satisfaction problems.	The phase transition in binary constraint satisfaction problems, i.e. the transition from a region in which almost all problem have many solutions to a region in which almost all problems have no solutions, as the constraints become tighter, is investigated.	-/Abs.
(12)	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/Intr.
	It offers the advantage that it is compatible with existing GIS technology.		

Table 2: *Example of Alignments*

	Documents		with A.Abs.		w/o A.Abs.		Average
	#	%	#	%	#	%	%
Title	10	2	6	2	4	1	2
Author Abstract	83	15	83	34			20
First section	195	34	61	26	134	42	40
Last section	18	3	6	2	12	4	4
Subtitles & Capt.	191	33	76	31	115	36	23
Other sections	71	13	13	5	58	17	11

Table 3: *Distribution of Information*

even in the case of documents with author provided abstract, information from other parts of the parent document would be used in the professional abstract as table shows. In addition, a "typical" element of the corpus (average information) will contain 40% from the introduction. Our findings are in agreement with Endres-Niggemeyer et *al.* study in cognitive science [Endres-Niggemeyer et al., 1995]. They found that in order to produce the "topical" sentence of an abstract the professional abstractor will use the introduction and conclusion of the parent document.

4 Analysis of the Results

The results so far indicate a correlation between the information in the abstract and structural parts of the parent document. But it is necessary to understand why some information is considered important, how to identify the information and how to use it in order to produce an abstract.

Sharp [Sharp, 1989] reports on experiments carried out with abstractors were it is shown that introductions and conclusions provide a basis for producing a coherent and informative abstract. In fact abstractors use a "short-cut" strategy (looking at introduction and conclusion) prior regarding the whole paper. But our results indicate that using just those parts is not enough to produce a good informative abstract. Important information is found in sections other than introduction and conclusion. The use of those elements in automatic abstracting has already been tested [Jang and Myaeng, 1997] but the success of the resulting abstract depends on factors such as the type of parent document and the type of information [Spark Jones, 1993]. Our observations regarding the use of structural parts of the parent document are as follows:

Use of titles. Titles of articles usually contain complete descriptions of the themes of the document so they could be used by the abstractor in order to convey the information in a more precise form. An example of such a situation can be seen in the alignment (4) where the complete descriptions of the title is used in a sentence.

Use of introductions. Usually, statements related to the purpose, the method and the problem studied in a technical article are found in introductions, those conceptual categories have to be reported in abstracts and so are easily extracted from that section because they are lexically marked, in alignment (9) the objective of the paper is stated in the introduction and in alignment (10) the plan of the document it is so.

Use of conclusions. In the conclusion, the author usually restates the objective of the document, this case is exemplified in alignment (4).

Use of subtitles. Subtitles of sections usually indicate the sub-themes of the document but also complete descriptions of entities could be found there, in alignment (5) the subtitle indicates the content of the section so it is used in the abstract to indicate that information.

Use of captioning. As tables (and figures) usually convey information about the results of an investigation, the captioning could be used to indicate the content reported in tables as alignment (7) shows.

Use of other sections. A very important part of the abstracts came from other sections of the document one example is given in (8) where the details about the project being described in the article are stated. In this case the information is relevant because it refers to the main entity being described in the document and not because of any lexical marker.

Use of author's abstract. Sometimes author provided abstracts are of a lower quality (or are less structured) than professional abstracts [Teufel and Moens, 1998] but abstractors would use the information found in the author abstract because it is clearly stated. As alignment (3) shows the information about the results of the investigation is stated in the author provided abstract and used in the professional abstract. Sometimes results of scientific work are difficult to extract and abstract from the main sections of parent documents [Paice and Jones, 1993].

Regarding the identification of the information, we found that the information extracted by professionals contains lexical markers of relevance ("*The principal distinguishing features* of EQLIPSE are..."), theme ("*The subject of this paper is* the concept of descriptor equivalence and...”), purpose (“*The purpose of this paper is* to assess retrospectively ...”), conclusion (“*Our conclusion was that* simple and local transformations can be automatized...”), results (“*We found that* significant...”) and plan of the document (“*we will first* put the Word Manager project in perspective...*we will then* describe the progress made...”). In fact 205 aligned sentences from the sections of the documents contain “indicative expressions” which represents 35% of the total of the aligned sentences. If, in addition, we take into account the fact that 35% of the aligned elements come from titles and subtitles, we obtain that 70% of the information is somehow “indicative” of the content of the documents. In summary, looking for text spans containing “indicative expressions” and using titles and subtitles when they clearly mark the themes being described is a good strategy for grasping the content of a text. This will be elaborated in the following section.

According to Cremmins [Cremmins, 1982] and Bernier [Bernier, 1985], the information obtained in the analytical stage will be “edited” in order to produce a concise text, but no operational method is given to do so. Having no access to the abstractors’ working environment, we can just make the following observations about the use of the information. We have already identified the following “operations” in the extracted material: *deletion of structural markers* to obtain a self contained text, as in alignment (1); *tense verb transformations* to make the style impersonal, as exemplified in (1) and (4); *deletion of clauses* which contain too much information, such as in (11); *join of information* for succinctness as in (3), (4) and (8); *split of complex clauses* as in (12) and; *theme indication* to point to information in the paper, as in (6) and (7).

The results of our analysis indicate that it is useful to look for information in specific parts of the parent document using lexical markers to obtain part of the information for the abstract but, the material obtained in this manner is sometimes too indicative. In order to produce a good informative abstract the information from introductions and conclusions have to be expanded somehow. Which information to expand depends in part on the reader’s interests. In fact if the abstract is to be used as a decision tool an indicative abstract could serve but if the reader wants more information about the entities being described in the document, additional information has to be obtained and integrated.

5 Abstracting by Selective Analysis

Our objective is to produce user-oriented abstracts which we define as abstract being indicative in the content of the document (i.e. what the authors present, discuss or show is stated in a short text) and informative in the particular details the reader of the document is interested in (i.e. identification of entities being described, advantages and drawbacks of a solution, description of the problem being investigated, etc.). The abstract produced will be a concise text without length relation with the parent document. The input to the system is a structured scientific or technical text in English. Our process of automatic abstracting is composed of four main steps:

Indicative Selection: sentences containing indicative expressions are extracted from specific parts of the parent document in order to produce the propositional content for an indicative abstract. This decision is based on the results of the corpus analysis. From these sentences, a pool of propositions is produced and the list of potential topics is obtained.

Informative Selection: texts spans which refer to the elements of the list of topics are selected from the parent document. These text spans represent additional information to convey in the abstract. We look for specific types of information (i.e. definitions, identification of entities being described, description of the problem and solution, etc.). The text spans are analyzed in order to produce the informative propositions for the abstract.

Indicative Generation: 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.

Informative Generation: the user will select some topics to expand from the indicative abstract and an informative abstract will be produced using the informative propositions obtained so far.

5.1 *A Short Example*

We manually apply the selective analysis to the paper "Case-based planning: selected methods and systems" from AI Communications 9 (1996) p128-137. This is a structured paper with an introduction but without a conclusion. The selective analysis proceeds as follows.

Indicative selection: in the introduction of the paper the following sentence, which indicates the main topic of the document is found (the marker "in this paper" is used to locate this sentence).

> (1) In this paper, we present three systems that integrate generative and case-based planning.

In fact, this is the only marked sentence from the introduction and it conveys the content of the overall paper (the paper just describes case-based planning and presents three systems in that domain). From it two propositions are obtained:

```
PRESENT(AUTHOR,‘‘three systems’’)
INTEGRATE(‘‘three systems’’,‘‘generative and case-based planning’’)
```

The entities "three systems" and "generative and case-based planning" are the topics of the sentence.

Indicative generation: from these propositions, the following indicative abstract is generated:

> *Presents* **three systems** *that integrate* **generative and case-based planning**.

Informative selection: in the parent document, the following sentences relating to the marked entities are found:

> (2) PRODIGY/ANALOGY was the first system that...
> (3) CAPLAN/CBC is a generic case-based reasoning system that...
> (4) PARIS (Plan Abstraction and Refinement in an Integrated System) is a domain-independent case-based planning system which...
> (5) The classical generative planning process consists mainly of a search through the space of possible sets of operators to solve a given problem.
> (6) In pure case-based planning instead, new problems are solved by reusing plans or portions of plans from previous cases.

From these sentences, the following informative propositions are obtained:

```
IDENTIFICATION(‘‘three systems’’,(PRODIGY/ANALOGY,CAPLAN/CBC,PARIS))
DEFINITION(‘‘generative planning’’,‘‘The classical generative...’’)
DEFINITION(‘‘case-based planning’’,‘‘In pure case-based planning...’’)
```

Informative generation: if the user is interested in additional information about the "three systems" and "generative and case-based planning" the following informative abstract could be produced:

> *The classical generative planning process consists mainly of a search through the space of possible sets of operators to solve a given problem. In pure case-based planning instead, new problems are solved by reusing plans or portions of plans from previous cases. The paper presents three systems that integrate generative and case-based planning: the PRODIGY/ANALOGY system, the CAPLAN/CBC system and the PARIS system.*

The abstract so produced is coherent because of the schemas used to expand the text: definitions and identifications of the main topics being described.

6 Discussion

Our methodology of automatic abstracting, called selective analysis, is intended to produce short indicative-informative texts for scientific and technical articles. The selection of the text spans for the indicative abstract is based on the searching for indicative expressions identified during the analysis of the corpus. The selection of material for the informative abstract is based on the search for expressions involving the main entities of the document.

The fact that we rely on such type of stylistic information is a limitation of the approach. In the case of articles in which the author does not mark explicitly their topics, other techniques have to be applied. We are investigating such a situation. A typical case we found in our corpus is the problem-solution structure of research papers in which some problem (or goal) and the possible solution are semantically expressed and the overall article talks about these two entities.

In the actual stage of our research we are implementing in Prolog the process of mapping the text spans to the propositional representation and in the meantime, we are defining the schemas for the integration of the informative material.

Acknowledgments

The first author is supported by ACDI and Ministerio de Educación de la Nación de la República Argentina, Resolución 1041/96. We would like to thank Sandra Szasz for her help in collecting part of the corpus. We also thank two anonymous reviewers for their comments.

References

[Bernier, 1985] Bernier, C. (1985). Abstracts and abstracting. In Dym, E., editor, *Subject and Information Analysis*, volume 47 of *Books in Library and Information Science*, pages 423–444. Marcel Dekker, Inc.

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

[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.

[Jang and Myaeng, 1997] Jang, D. and Myaeng, S. (1997). Development of a document summarization system for effective information services. In *RIAO-97. Computer-Assisted Information Searching on Internet.*, pages 101–111.

[Kupiec et al., 1995] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proc. of the 18th ACM-SIGIR Conference*, pages 68–73.

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

[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.

[Sharp, 1989] Sharp, B. (1989). *Elaboration and testing of new methodologies for automatic abstracting.* PhD thesis, The University of Aston in Birmingham.

[Spark Jones, 1993] Spark Jones, K. (1993). Discourse modelling for automatic summarising. Technical Report 290, University of Cambridge, Computer Laboratory.

[Teufel and Moens, 1998] Teufel, S. and Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In *Intelligent Text Summarization*, pages 16–25.