Using Linguistic Knowledge in Automatic Abstracting

Horacio Saggion
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@iro.umontreal.ca

Abstract

We present work on the automatic generation of short indicative-informative abstracts of scientific and technical articles. The indicative part of the abstract identifies the topics of the document while the informative part of the abstract elaborate some topics according to the reader's interest by motivating the topics, describing entities and defining concepts. We have defined our method of automatic abstracting by studying a corpus professional abstracts. The method also considers the reader's interest as essential in the process of abstracting.

1 Introduction

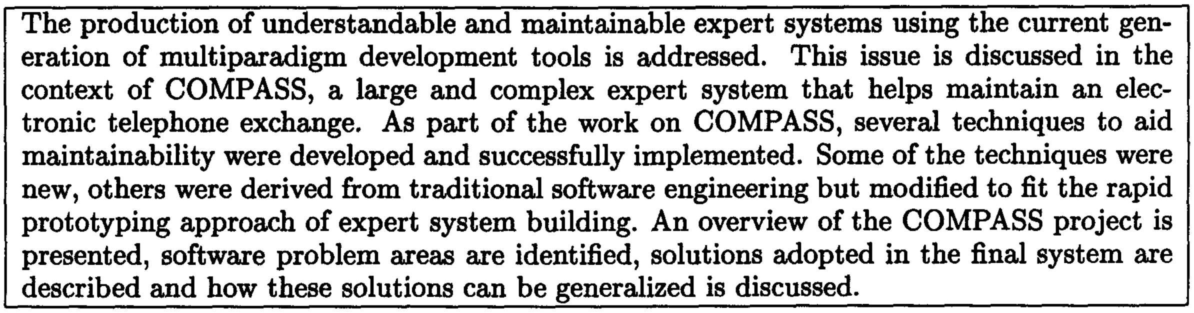
The idea of producing abstracts or summaries by automatic means is not new, several methodologies have been proposed and tested for automatic abstracting including among others: word distribution (Luhn, 1958); rhetorical analysis (Marcu, 1997); and probabilistic models (Kupiec et al., 1995). Even though some approaches produce acceptable abstracts for specific tasks, it is generally agreed that the problem of coherent selection and expression of information in automatic abstracting remains (Johnson, 1995). One of the main problems is how to ensure the preservation of the message of the original text if sentences picked up from distant parts of the source text are juxtaposed and presented to the reader. Rino and Scott (1996) address the problem of coherent selection for gist preservation, however they depend on the availability of a complex meaning representation which in practice is difficult to obtain from the raw text.

In our work, we are concerned with the automatic generation of short indicative-informative abstract for technical and scientific papers. We base our methodology on a study of a corpus of professional abstracts and source or parent documents. Our method also considers the reader's interest as essential in the process of abstracting.

2 The Corpus

The production of professional abstracts has long been object of study (Cremmins, 1982). In particular, it has been argued that structural parts of parent documents such as introductions and conclusions are important in order to obtain the information for the topical sentence (Endres-Niggemeyer et al., 1995). We have been investigating which kind of information is reported in professional abstracts as well as where the information lies in parent documents and how it is conveyed. In Figure 1, we show a professional abstract from the "Computer and Control Abstracts" journal, this kind of abstract aims to alert readers about the existence of a new article in a particular field. The example contains information about the author's interest, the author's development and the overview of the parent document. All the information reported in this abstract was found in the introduction of its parent document.

In order to study the aforementioned aspects, we have manually aligned sentences of 100 professional abstracts with sentences of parent documents containing the information reported in the abstract. In a previous study (Saggion and Lapalme, 1998), we have shown that 72% of the information in professional abstracts lies in titles, captions, first sections and last sections of parent documents while the rest of the information was found in author abstracts and other sections. These results suggest that some structural sections are particularly important in order to select information for an abstract but also



The production of understandable and maintainable expert systems using the current generation of multiparadigm development tools is addressed. This issue is discussed in the context of COMPASS, a large and complex expert system that helps maintain an electronic telephone exchange. As part of the work on COMPASS, several techniques to aid maintainability were developed and successfully implemented. Some of the techniques were new, others were derived from traditional software engineering but modified to fit the rapid prototyping approach of expert system building. An overview of the COMPASS project is presented, software problem areas are identified, solutions adopted in the final system are described and how these solutions can be generalized is discussed.

Figure 1: Professional Abstract: CCA 58293 (1990 vol.25 no.293). Parent Document: "Maintainability Techniques in Developing Large Expert Systems." D.S. Prerau *et al.* IEEE Expert, vol.5, no.3, p.71-80, June 1990.

that it is not enough to produce a good informative abstract (i.e. we hardly find the results of an investigation in the introduction of a research paper).

3 Conceptual and Linguistic Information

The complex process of scientific discovery that starts with the identification of a research problem and eventually ends with an answer to the problem (Bunge, 1967), would generally be disseminated in a technical or scientific paper: a complex record of knowledge containing, among others, references to the following concepts *the author*, *the author's affiliation*, *others authors*, *the authors' development*, *the authors' interest*, *the research article and its components* (sections, figures, tables, etc.), *the problem under consideration*, *the authors' solution*, *others' solution*, *the topics of the research article*, *the motivation for the study*, *the importance of the study*, *what the author found*, *what the author think*, *what others have done*, and so forth. Those concepts are systematically selected for inclusion in professional abstracts. We have noted that some of them are lexically marked while others appear as arguments of predicates conveying specific relations in the domain of discourse. For example, in an expression such as "We found significant reductions in ..." the verb "find" takes as an argument a *result* and in the expression "The lack of a library severely limits the impact of..." the verb "limit" entails a *problem.*

We have used our corpus and a set of more than 50 complete technical articles in order to deduce a conceptual model and to gather lexical information conveying concepts and relations. Although our conceptual model does not deal with all the intricacies of the domain, we believe it covers most of the important information relevant for an abstract. In order to obtain linguistic expressions marking concepts and relation, we have tagged our corpus with a POS tagger (Foster, 1991) and we have used a thesaurus (Vianna, 1980) to semantically classify the lexical items (most of them are polysemous). Figure 2, gives an overview of some concepts, relations and lexical items so far identified.

The information we collected allow the definition of patterns of two kinds: (i) linguistic patterns for the identification of noun groups and verb groups; and (ii) domain specific patterns for the identification of entities and relations in the conceptual model. This allows for the identification of complex noun groups such as "The TIGER condition monitoring system" in the sentence "The TIGER gas turbine condition monitoring system addresses the performance monitoring aspects" and the interpretation of strings such as "University of Montréal" as a reference to an institution and verb forms such as "have presented" as a reference to a predicate possibly introducing the topic of the document. The patterns have been specified according to the linguistic constructions found in the corpus and then expanded to cope with other valid linguistic patterns, though not observed in our data.

Concepts/Relations	Explanation	Lexical Items
make know	The author mark the topic of the document	describe, expose, present, ...
study	The author is engaged in study	analyze, examine, explore, ...
express interest	The author is interested in	address, concern, interest,...
experiment	The author is engaged in experimentation	experiment, test, try out, ...
identify goal	The author identify the research goal	necessary, focus on, ...
explain	The author gives explanations	explain, interpret, justify,...
define	a concept is being defined	define, be, ...
describe	entity is being described	compose, form, ...
authors	The authors of the article	We, I, author,...
paper	The technical article	article, here, paper, study, ...
institutions	authors' affiliation	University, Universitė, ...
other researchers	Other researchers	Proper Noun (Year), ...
problem	The problem under consideration	difficulty, issue, problem, ...
method	The method used in the study	equipment, methodology, ...
results	The results obtained	result, find, reveal, ...
hypotheses	The assumptions of the author	assumption, hypothesis, ...

Figure 2: Some Conceptual and Linguistic Information

4 Generating Abstracts

It is generally accepted that there is no such thing as an ideal abstract, but different kinds of abstracts for different purposes and tasks (McKeown et al., 1998). We aim at the generation of a type of abstract well recognized in the literature: short indicative-informative abstracts. The indicative part identifies the topics of the document (what the authors present, discuss, address, etc.) while the informative part elaborates some topics according to the reader's interest by motivating the topics, describing entities, defining concepts and so on. This kind of abstract could be used in tasks such as accessing the content of the document and deciding if the parent document is worth reading. Our method of automatic abstracting relies on:

- the identification of sentences containing domain specific linguistic patterns;
- the instantiation of templates using the selected sentences;
- the identification of the topics of the document and;
- the presentation of the information using re-generation techniques.

The templates represent different kinds of information we have identified as important for inclusion in an abstract. They are classified in: **indicative templates** used to represent concepts and relations usually present in indicative abstracts such as "the topic of the document", "the structure of the document", "the identification of main entities", "the problem", "the need for research", "the identification of the solution", "the development of the author" and so on; and **informative templates** representing concepts that appear in informative abstracts such as "entity/concept definition", "entity/concept description", "entity/concept relevance", "entity/concept function", "the motivation for the work", "the description of the experiments", "the description of the methodology", "the results", "the main conclusions" and so on. Associated with each template is a set of rules used to identify potential sentences which could be used to instantiate the template. For example, the rules for the topic of the document template, specify to search the category *make know* in the introduction and conclusion of the paper while the rules for the entity description specify the search for the *describe* category in all the text. Only sentences matching specific patterns are retained in order to instantiate the templates and this reduces in part the problem of polysemy of the lexical items.

The overall process of automatic abstracting shown in Figure 3 is composed of the following steps:

Pre-processing and Interpretation: The raw text is tagged and transformed in a structured representation allowing the following processes to access the structure of the text (words, groups of words, titles, sentences, paragraphs, sections, and so on). Domain specific transducers are applied in order to identify possible concepts in the discourse domain (such as the authors, the paper, references to other authors, institutions and so on) and linguistic transducers are applied in order to identify noun groups and verb groups. Afterwards, semantic tags marking discourse domain relations and concepts are added to the different elements of the structure.

Additionally, the process extracts noun groups, computes noun group distribution (assigning a weight to each noun group) and generates the topical structure of the paper: a structure with $n + 1$ components where n is the number of sections in the document. Component i ($0 \leq i \leq n$) contains the noun groups extracted from the title of section i (0 indicates the title of the document). The structure is used in the selection of the content for the indicative abstract.

Indicative Selection: Its function is to identify potential topics of the document and to construct a pool of "propositions" introducing the topics. The indicative templates are used to this end: sentences are selected, filtered and used to instantiate the templates using patterns identified during the analysis of the corpus. The instantiated templates obtained in this step constitute the indicative data base. Each template contains, in addition to their specific slots, the following: the *topic candidate* slot which is filled in with the noun groups of the sentence used for instantiation, the *weight* slot filled in with the sum of the weights of the noun groups in the *topic candidate* slot and, the *position* slot filled in with the position of the sentence (section number and sentence number) which instantiated the template. In Figure 4, the "topic of the document" template appears instantiated using the sentence "this paper describes the Active Telepresence System with an integrated AR system to enhance the operator's sense of presence in hazardous environments."

In order to select the content for the indicative abstract the system looks for a "match" between the topical structure and the templates in the indicative data base: the system tries all the matches between noun groups in the topical structure and noun groups in the topic candidate slots. One template is selected for each component of the topical structure: the template with more matches. The selected templates constitute the content of the indicative abstract and the noun groups in the topic candidate slots constitute the potential topics.

Informative Selection: this process aims to confirm which of the potential topics computed by the indicative selection are actual topics (i.e. topics the system could informatively expand according to the reader interest) and produces a pool of "propositions" elaborating the topics. All informative templates are used in this step, the process considers sentences containing the potential topics and matching informative patterns. The instantiated informative templates constitute the informative data base and the potential topics appearing in the informative templates form the topics of the document.

Generation: This is a two step process. First, in the indicative generation, the templates selected by the indicative selection are presented to the reader in a short text which contains the topics identified by the informative selection and the kind of information the user could ask for. Second, in the informative generation, the reader selects some of the topics asking for specific types of information. The informative templates associated with the selected topics are used to present the required information to the reader using expansion operators such as the "description" operator whose effect is to present the description of the selected topic. For example, if the "topic of the document" template (Figure 4) is selected by the informative selection the following indicative text will be presented:

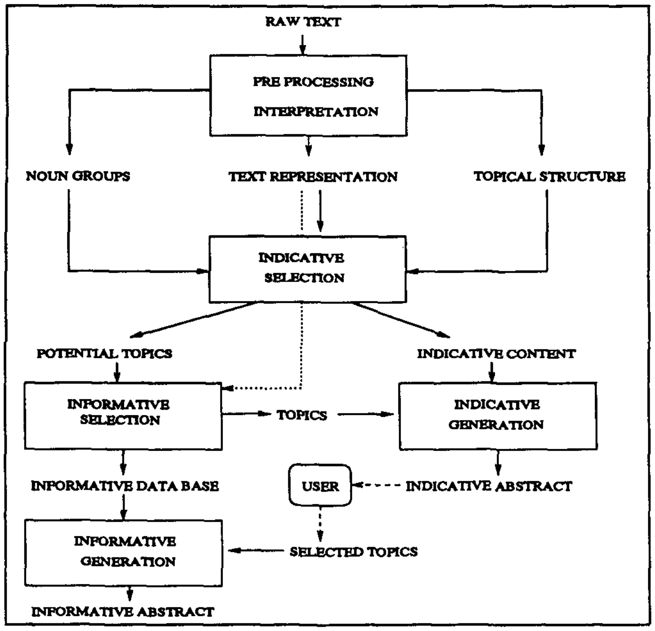


Figure 3: System Architecture

Templates and Instantiated Slots	
Topic of the document template	*Entity description template*
Main predicate: "describes" : DESCRIBE	**Main predicate:** "consist of" : CONSIST OF
Where: nil	**Topical entity:** "The Active Telepresence System"
Who: "This paper" : PAPER	**Related entities:** "three distinct elements", "the stereo head", "its controller", "the display device"
What: "the Active Telepresence System with an integrated AR system to enhance the operator's sense of presence in hazardous environments"	**Position:** Number 4 from "The Active Telepresence System" Section
Position: Number 1 from "Conclusion" Section	
Topic candidates: "the Active Telepresence System", "an integrated AR system", "the operator's sense", "presence", "hazardous environments"	**Weight:...**
Weight:...	

Figure 4: Some Instantiated Templates for the article "Augmenting reality for telerobotics: unifying real and virtual worlds" J. Pretlove, Industrial Robot, vol.25, issue 6, 1998.

Describes the Active Telepresence System with an integrated AR system to enhance the operator's sense of presence in hazardous environments.

Topics: *Active Telepresence System (description); AR system (description); AR (definition)*

If the reader choses to expand the description of the topic "Active Telepresence System", the following text will be presented:

The Active Telepresence System consists of three distinct elements: the stereo head, its controller and the display device.

The pre-processing and interpretation step are currently implemented. We are testing the processes of indicative and informative selection and we are developping the generation step.

5 Discussion

In this paper, we have presented a new method of automatic abstracting based on the results obtained from the study of a corpus of professional abstracts and parent documents. In order to implement the model, we rely on techniques in finite state processing, instantiation of templates and re-generation techniques. Paice and Jones (1993) have already used templates representing specific information in a restricted domain in order to generate indicative abstracts. Instead, we aim at the generation of indicative-informative abstracts for domain independent texts. Radev and McKeown (1998) also used instantiated templates, but in order to produce summaries of multiple documents. They focus on the generation of the text while we are addressing the overall process of automatic abstracting.

We are testing our method using long technical articles found on the "Web." Some outstanding issues are: the problem of co-reference, the problem of polysemy of the lexical items, the re-generation techniques and the evaluation of the methodology which will be based on the judgment of readers.

Acknowledgments

I would like to thank my adviser, Prof. Guy Lapalme for encouraging me to present this work. 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.

References

M. Bunge. 1967. *Scientifc Research I. The Search for System.* Springer-Verlag New York Inc.

E.T. Cremmins. 1982. *The Art of Abstracting.* ISI PRESS.

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

G. Foster. 1991. Statistical lexical disambiguation. Master's thesis, McGill University, School of Computer Science.

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

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

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

D. Marcu. 1997. From discourse structures to text summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11.

K. McKeown, D. Jordan, and V. Hatzivassiloglou. 1998. Generating patient-specific summaries of on-line literature. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 34–43, Standford (CA), USA, March 23-25. The AAAI Press.

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

D.R. Radev and K.R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

L.H.M. Rino and D. Scott. 1996. A discourse model for gist preservation. In D.L. Borges and C.A.A. Kaestner, editors, *Proceedings of the 13th Brazilian Symposium on Artificial Intelligence, SBIA'96*, Advances in Artificial Intelligence, pages 131–140. Springer, October 23-25, Curitiba, Brazil.

H. Saggion and G. Lapalme. 1998. Where does information come from? corpus analysis for automatic abstracting. In *RIFRA'98. Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique*, pages 72–83.

F. de M. Vianna, editor. 1980. *Roget's II. The New Thesaurus.* Houghton Mifflin Company, Boston.