Selective Analysis for the Automatic Generation of Summaries

Abstract: Selective Analysis is a new method for text summarization of technical articles whose design is based on the study of a corpus of professional abstracts and technical documents. The method emphasizes the selection of particular types of information and its elaboration exploring the issue of dynamical summarization. A computer prototype was developed to demonstrate the viability of the approach and the automatic abstracts were evaluated using human informants. The results so far obtained indicate that the summaries are acceptable in content and text quality.

1. Introduction

Selective Analysis is a new method of automatic abstracting, we have designed for technical documents and that we implemented using state of the art techniques in Natural Language Processing. The method produces short indicative-informative abstracts in two steps: first, the reader is presented with an indicative abstract which identifies the topics of the document (what the authors present, discuss, address, the problem, the solution, etc.). Then, if the reader is interested in some of the topics, information from the source document elaborating the topics will be presented. The informative abstracts are produced dynamically by this particular interaction between reader and document allowing the reader to access conceptual information about the topics. This method has been designed for tasks such as deciding if the source text is worth reading and informing the reader about specific topics.

Figure 1 shows an abstract produced with the current version of Selective Analysis (the system uses texts found on the Web). The source document “Climbing, walking and intervention robots” from the electronic version of Industrial Robot 24(2) was found on the Emerald Electronic Library. The topics identified by the system, which are terms appearing in the source document and are introduced in the automatic abstract, are presented in a separate list.

If the reader wants more information about the topic RIMHO he/she will get text spans from the source text such as those shown below which describe or define the topic:

The RIMHO walking robot is a prototype developed with the aim of studying the potential possibilities of walking locomotion in hazardous environments, especially in nuclear power plants.

The RIMHO is a quadruped walking machine of the insect type.

In this paper, we briefly describe the design of Selective Analysis, its implementation and the evaluation of the automatic abstracts.

The complex problem of inspection and maintenance of the steam generator in nuclear power plants was approached using previously gained expertise and, as a result, an innovative solution was achieved with the development of two co-operative robots, remotely controlled from a tele-operation station incorporating tele-presence. The department was leading a project for the introduction of a robotics system whose mission was to avoid human operators having to enter the steam generator's water chamber. Proposes an innovative solution to the serious problem of inspection and maintenance of the steam generator tubes, which must be checked regularly to detect leakage of radioactive water from the primary to the secondary circuit. Another interesting activity was the...
realization, within the framework of a EUREKA project, of a tele-manipulator for servicing a new concept of urban infrastructures. Shows the RIMHO walking robot and ROBUR arm exchanging gas filter in IUI (Industrializable urban infrastructures) demonstration.


Figure 1: Abstract produced using Selective Analysis

2. Conceptual Information

One of the main research questions in automatic abstracting is what information might be included in an abstract (Spark Jones, 1998). In order to address that question, we have studied abstracts written by professional abstractors together with their source documents. 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).

We manually aligned each sentence of the abstract 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.

In Table 1, we present the distribution of the sentences in the parent documents which were aligned with the professional abstracts in our corpus. The first three columns contain, 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 distribution of information in a "typical" source document. We found that 72% of the information included in the abstracts of our corpus 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.

<table>
<thead>
<tr>
<th>Documents</th>
<th>with A.Abs.</th>
<th>w/o A.Abs.</th>
<th>Avr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Title</td>
<td>10</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Author abstract</td>
<td>83</td>
<td>15</td>
<td>83</td>
</tr>
<tr>
<td>First section</td>
<td>195</td>
<td>34</td>
<td>61</td>
</tr>
<tr>
<td>Last section</td>
<td>18</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Subtitles &amp; capt.</td>
<td>191</td>
<td>33</td>
<td>76</td>
</tr>
<tr>
<td>Other sections</td>
<td>71</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 1: Distribution of Information

The analysis of the corpus allowed us to identify more than 40 concepts (author, researchers, institutions, references, problem, solution, method, etc.), more than 50 relations (make known, describe, define, identify problem, study, experiment, etc.), and more than 50 types of information which seem important for automatic abstracting and which are typical of
scientific and technical texts. We also identified some common transformation that human abstractors apply in order to produce a coherent abstract including verb transformation, concept deletion, concept reformulation, structural deletion, parenthetical deletion, clause deletion, acronym expansion, abbreviation, merge and split (see (Saggion and Lapalme, 1998) for the results). In our corpus 89% of the sentences from the professional abstracts included at least one transformation.

While most approaches to automatic text summarization present the extracted information in both the order and the form of the original, this is not the case in human produced abstracts. Nevertheless, some transformations in the source document can be implemented by computers with state of the art techniques in natural language processing in order to improve the quality of the automatic summaries. Our concerns regarding the presentation of the information are now being addressed by other researchers as well (Jing and McKeown, 1999).

3. Producing Abstracts

The implementation of our method relies on: the selection of particular types of information from the source text; the instantiation of different types of templates; the selection of some of the templates in order to produce an indicative abstract; the (re)generation of a short but novel text which indicates the topics of the document and; the expansion of the indicative text with topic elaborations. The overall process of automatic abstracting is composed of the following steps (for a complete description refer to (Saggion, 1999)):

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. Each sentence is syntactically and semantically interpreted using partial analysis. The process also extracts terms, computes terms distribution and based on the titles generates the topical structure of the paper (i.e. list of terms) which 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. 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 among others the topic candidate slot which is filled in with some terms of the sentence used for instantiation. In order to select the content for the indicative abstract the system looks for a "match" between the topical structure and the topic candidate slots in the templates in the indicative data base. The selected templates constitute the indicative content and the terms appearing in the indicative content form 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. The process considers sentences containing the potential topics and matching informative patterns in order to instantiate informative templates. 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, the content is sorted using a conceptual order and then it generates sentences from instantiated templates according to the style observed in the corpus (i.e. verbs in the impersonal and reformulation of some domain concepts), when possible the system integrates information from different templates to produce a single sentence. Second, the informative generation, where 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.

While the indicative abstract depends on the structure, content, and to some extent, on
specific types of information generally reported in this kind of summary, the informative abstract depends on the interest of the reader to know more about the topics.

(Paice and Jones, 1993) have already addressed the issue of content identification and expression in technical summarization using templates, but while they produced indicative abstracts for a specific domain, we are producing domain independent indicative-informative abstracts. Being designed for one specific domain, their abstracts are fixed in structure while our abstracts are dynamically constructed.

4. Evaluation

The quality and success of human produced abstracts have already been addressed in the literature (Gibson, 1993) using linguistic criteria such as cohesion and coherence, thematic structure, sentence structure and lexical density. But in automatic text summarization, this is an emergent research topic. The most extensive evaluation to date was done in the context of the TIPSTER SUMMAC evaluation (Mani et al., 1998). That evaluation was extrinsic in the sense that it addressed the usefulness of automatic abstracts in the completion of a given task (i.e. categorization, question-answering, etc.). Even if this effort created a corpus of full-texts for evaluation purposes there is a clear lack of resources for the evaluation of technical articles.

Our experiments addressed the evaluation of content and text quality relying on human judges using technical articles found on the Web. We compared abstracts produced by our method with abstracts produced by Microsoft’97 Summarizer and with others published with source documents (usually author abstracts). In order to evaluate content, we presented judges with randomly selected abstracts and five lists of keywords (content indicators). The judges had to decide to which list of keywords the abstract belongs given the fact that different lists share some keywords and they belong to the same technical domain. Those lists were obtained from the journals where the source documents were published. The idea behind this evaluation is to see if the abstract convey the essential content of the source document.

In order to evaluate the quality of the text, we asked the judges to provide an acceptability score between 0-5 for the abstract (0 for unacceptable and 5 for acceptable) based on criteria taken from (Rowley, 1982) such as “good spelling and grammar”, “indication of the topic”, “readable and understandable”, etc. (they were only suggestions to the evaluators and were not enforced). We have told the judges that we would consider the abstracts with scores above 2.5 as acceptable.

This experiment was repeated two times with different human informants and data. Each abstract was evaluated by three different judges. For each abstract, we aggregate the information in averages of indicativeness (i.e. at least two judges have chosen the correct descriptor and so we considered that the abstract helped the reader to chose the descriptor) and text quality. The results are reported in Table 2. Regarding content, both automatic methods performed similarly. The abstract provided with the journal was found more helpful than the automatic abstracts. But it is interesting to note that in the second experiment only 80% of the author provided abstracts were correctly classified. Regarding text quality, the abstract provided with the original document are always of better quality than the automatic abstracts. Taking into account the given criteria, abstract produced by Selective Analysis are considered of acceptable quality by the evaluators while abstract produced by Microsoft Summarizer are not.

In previous experiments, we have compared the accuracy of the terms identified by Selective Analysis with terms produced by other two different summarization methods. So, Selective Analysis performed better that the other methods but there is no indication of better performance in a majority of the cases. This experiment was reported in (Saggion and Lapalme, 2000).
Table 2: Results of Human Judgment about Indicativeness and Text Quality

<table>
<thead>
<tr>
<th></th>
<th>Microsoft Abstract</th>
<th>S.A Abstract</th>
<th>S.D. Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indicative</td>
<td>Quality</td>
<td>Indicative</td>
</tr>
<tr>
<td>1st Exp.</td>
<td>80%</td>
<td>1.46</td>
<td>80%</td>
</tr>
<tr>
<td>2nd Exp.</td>
<td>70%</td>
<td>1.98</td>
<td>70%</td>
</tr>
</tbody>
</table>

5. Conclusion

We have designed and implemented a new method of automatic abstracting that produces indicative-informative abstracts. This is accomplished by processes of partial syntactic and semantic analysis, sentence classification, template instantiation, content selection, text regeneration and topic elaboration. Our method was deeply influenced by the results of our corpus study, nevertheless, it has many points in common with recent theoretical and programmatic directions in automatic text summarization.

Spark Jones argues in favor of a kind of "indicative, skeletal summary" and the need to explore dynamic, context-sensitive summarization in interactive situations where the summary changes according to the user needs. Hutchins (Hutchins, 1995) advocates for indicative summaries, produced from parts of the document where the topics are likely to be stated. These abstracts are well suited for situations in which the actual user is unknown (i.e., a general reader), and so the abstract will provide the reader with good entry points for the information.

If the users were known, the abstract could be tailored towards their specific profiles (profiles could include the specification of readers interested in types of information like conclusions, definitions, methods, or user needs expressed in a "query" to an information retrieval system (Tombros et al., 1998)), but our method was designed without any particular reader in mind and with the assumption that a text does have a "main" topic. Our method also mirrors somehow professional techniques used in abstract writing (see for example (Cleveland and Cleveland, 1983)).

In the experiments described in this paper, we only addressed the evaluation of the indicative abstract in content and text quality. Our future work includes the evaluation of the content of the informative abstract relying on the comparison with "ideal" informative abstracts produced by humans. Even if several issues have been addressed as part of this research, important concerns will be subject of future improvements, specially for topic elaboration.

Acknowledgments

We would like to thank Professor Michèle Hudon from Université de Montréal for fruitful discussion and Mme Gracia Pagola from EBSI Université de Montréal and Professor John E. Leide from McGill Graduate School of Library & Information Studies for providing with human informants for the experiments. The first author is supported by Agence Canadienne de Développement International (ACDI) and Fundación Antorchas (A-13671/1-47), Argentina. He was previously supported by Ministerio de Educación de la Nación de la República Argentina (Resolución 1041/96) and Departamento de Computación, Facultad de Ciencias Exactas y Naturales, UBA, Argentina.

References


