Imitating Human Strategies for Information Summarization

Kokil Jaidka
Wee Kim Wee School of Communication & Information
Nanyang Technological University
Singapore
+65 - 67904460
kokil@pmail.ntu.edu.sg

ABSTRACT
In this paper, I provide my preliminary findings from my study of researchers’ preferences in selecting and summarizing information from source papers to include in a literature review. Preliminary findings indicate that researchers show a strong preference for selecting information from certain sections of a research paper, such as its Abstract, Introduction and Conclusion. This preference is affected by the style of literature review they are constructing. In future work, these findings will be applied to emulate human summarization behavior in the information selection and text generation processes for an automatic literature review generation system.

Keywords
Multi-document summarization, literature reviews, natural language processing, discourse analysis, rhetorical structure.

1. INTRODUCTION
This study is a part of a larger project to mimic the characteristics of naturally-written literature review sections, in an automatic summary of research papers. The domain selected for the study is the field of information science, and the corpus of literature reviews used for analysis and model development are the literature review sections of articles published in the top journals in the field, namely the Journal of the American Society for Information Science and Technology, Journal of Documentation and Journal of Information Science.

The first step in this research project is to study the strategies used by researchers to select and extract relevant information. The next step is to conduct a content analysis to identify the structural layout and linguistic expressions applied by researchers to fulfill different functions in a literature review. Thus, the tasks have to be carried out include:

- A source-reference analysis to identify researchers’ preferences in selecting and changing source information from research papers to summarize as referencing information in a literature review.
- A content analysis at three levels to identify the functional and linguistic structure of naturally-written literature reviews:
  - At the document level, to determine the layout and discourse elements of a literature review
  - At the sentence level, to determine the linguistic expressions used to realize different informational functions
  - At the clause level, to identify the kinds of information which is provided and how it is compared across articles

The context of this study lies in multi-document summarization. The proposed literature review generation system is shown in Figure 1. Although the automatic summarization process is comprised of four main steps: text pre-processing, information selection and integration, summary generation and post-processing, the novel approaches proposed in my method would mainly be at the information selection and text generation stage, to select information from different semantic levels and draft the literature review out of sentence templates. These processes will be based on my analysis of human summarization strategies.

In this paper, I describe my methodology for the cross-document level analysis, wherein I mapped sentences from literature reviews to the references source papers in order to identify researchers’ preferences in selecting and transforming information for a literature review. Although there has been some research on human summarizations strategies before [1], it focused on news-related and event-based summarizing at the syntactic level; however, this study will map information to source text at different semantic levels, and specifically focuses on the summarizing strategies in information science literature reviews. Applying these findings, I aim to use a combination of surface, lexical and semantic features to select common as well as unique information from the source documents to generate a comparative literature review.

3. METHODOLOGY
In the exploratory phase, I sampled twenty research articles from 8 volumes of JASIST (2001-2008), two or three articles from each year. I extracted their literature review sections for analysis and analyzed them line-by-line to identify four types of information:

- The information types in the referencing sentences. I annotated the referencing sentences according to the one or more types of information (i.e., objective, methodology, results and critical summary) which they provided. These categories were adapted from an earlier study [2] of the kinds of research information provided in a technical article.
- The location in the source papers. I annotated the selected source sentences with the sections (i.e., Abstract, Introduction, Methodology, Results and Conclusion) where they were found. If there were multiple candidate sources, the sentence which was worded as close as possible to the referencing sentence in wording was noted as the original source.
- The transformations performed on the source information. I annotated every source with the series of transformations it underwent to become the reference, namely, cut-pasted or word-for-word; paraphrased or some words substituted; and higher-level summary or critical overview. These transformation categories were identified during this analysis and are similar to the categories in [1].
The rejected alternatives which provided similar information. I analyzed these in order to surmise the reasons for choosing one source over other alternatives, despite providing similar information.

Figure 1 The proposed literature review generation system

4. PRELIMINARY RESULTS

In the content analysis of literature review sections, an XML document structure language was developed to annotate and identify the different sections of a literature review [3]. It was observed that information science literature reviews can be classified as descriptive or integrative in style, and their discourse structures are correspondingly different [3]. Integrative literature review articles represent literature on a topic as a critical review [4], whereas descriptive literature reviews briefly summarizes the research information of a research study [5]. Both types of information science literature reviews follow a hierarchical structure and have a typical composition of discourse elements. On the basis of these findings, generic templates have been designed for the literature review generation system.

For each of these literature review styles, the following preferences for information selection were observed:

- Of the sources of information, the Abstract section is referenced more often than others. In integrative literature reviews, source sentences from the Abstract tend to be paraphrased whereas in descriptive literature reviews, they tend to be cut-pasted.
- Integrative literature reviews reference significantly more information from the Conclusion, Results and Related work sections, whereas descriptive literature reviews reference more information from the Abstract and Introduction sections.
- There is more cut-paste in descriptive literature reviews than integrative literature reviews. A large proportion of the research objective and research method information are summarized at a high level, with no specific source sentences.

5. CONCLUSION

This part of the study has mapped references in the literature review to the original source papers in a source-reference analysis, and drawn a correlation between the kind of information selected, its location in the source and the type of transformation performed. Different types of information is chosen from different locations and transformed in different ways. These choices are also different for integrative versus descriptive styles of literature review writing.

In my confirmatory study, I will analyze integrative summary sentences to identify the types of information they provide and the different purposes they fulfill while comparing and critiquing information. In doing so, I will relate my source-reference analysis to my findings of the sentence-level analysis. However, due to the integrative and critical nature of these sentences, they pose a difficult challenge for analysis and classification. At the same time, they provide the opportunity for a significant contribution towards emulating human-written summaries.

The techniques developed for the purpose of multi-document summarization through this study would contribute to applications of multi-document auto-summarization to research information; if successful it would also prove to be a viable and feasible research tool for the review of large quantities of literature. Its applications can also be imagined in web-based databases and digital libraries to allow users a convenient and speedy method for browsing available content.
6. REFERENCES


