User-Sensitive Text Summarization
Thesis Summary

Noemie Elhadad
Computer Science Department
Columbia University
New York, New York 10027
noemie@cs.columbia.edu

Introduction

Text summarization has emerged as an increasingly established field over the course of the past ten years. We may soon reach a stage where researchers will be able to design, and provide everyday users with, robust text summarization systems. Users of text summarization are many and range from the Internet surfers lacking the time to locate and digest all the latest news available on the web to scientists unable to keep pace with the burgeoning number of technical publications who must, nonetheless, be familiar with the latest findings in their fields.

Given texts to summarize, there is no *a priori* criteria for determining relevance for the summary. When humans summarize texts, they identify relevant information that they think will be of interest to the readers. Summarization is not only a function of the input documents but also of the reader's mental state: who the reader is, what his knowledge before reading the summary consists of, and why he wants to know about the input texts. This fact has been long acknowledged by both the psycho-linguistic and the computational-linguistic communities. However, both communities agree that trying to model the reader's mental state is far too complicated, if not entirely impossible. Given this dilemma, most of the computational linguistic research in summarization has assumed that the "reader variable" is a constant and has focused on defining a general notion of salience, valid for all readers.

In my thesis, I investigate strategies to take user characteristics into account in the summarization process. Acquiring a user model is by itself a wide subject of research. I do not focus on ways to acquire a user model, and I assume that there is an existing user model in my framework. Rather, my focus is on the challenges entailed in incorporating knowledge about the user into summarization strategies and providing the user with a text relevant to his needs.

In my work, two types of user tailoring are examined: *individualized*, i.e., the specific facts in which the reader is interested, and *class-based*, i.e., the degree of expertise of the reader. My research framework consists of PERSIVAL, a digital library that provides tailored access to medical technical literature for both physicians and patients. When treating a specific patient, physicians will want to keep abreast of the latest findings that pertain to that patient. Likewise, patients may want to access the latest findings that are relevant to their medical situation but may be hindered by the jargon commonly used in technical medical texts. My summarizer attempts to provide both types of users with tailored syntheses of the findings reported in clinical studies. Such tailoring is accomplished at the *individual* level by taking advantage of the existing patient record in the digital library. The summarizer will also adapt the language in which the summary is generated by using *class-based* information, i.e., whether the user is a physician or a patient.

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Summaries and the Users' Interests

There are several challenges in incorporating individual-based modeling. The first one consists of deciding which stages of the summarization process should be affected by the individual characteristics of the user; should it occur when selecting the content to include in the summary, when organizing the selected bits of information, and/or when choosing verbalizations? The second challenge is to determine the degree of abstraction and inference needed to achieve satisfactory modeling. Given input texts to summarize and a patient record composed of many reports in textual format, one must choose a representation of the information conveyed both in the input and the patient record that allows for efficient and accurate matching of information.

One contribution of my thesis work is to show how individual characteristics can determine what content should be included in the summary. While this has been investigated in information retrieval and traditional semantic-to-text generation, incorporating the user in the selection of content from textual input is a novel approach for summarization. Individual modeling takes place at the content selection stage; it is the place where the user's interests have most impact on the output summary. The summarizer selects the findings that pertain to the user's characteristics. This is achieved in a simple fashion thanks to an adequate representation of the input texts and the patient record. Using information extraction techniques and relying on an existing ontology, the summarizer operates over a data-structure called *content item* that is between full-semantic analysis and agnostic extracted text. This representation enables the use of simple matching strategies between the user model and the input texts.

At the current stage of my thesis work, I have designed and implemented a multi-document summarizer which generates tailored summaries of clinical studies (Elhadad *et al.* 2004). It is integrated into the overall PERSIVAL digital library. It tailors its output based on the record of the patient under care. Using information extraction techniques, *content items* are extracted at the sentence level from the input articles. Each content item represents a finding reported in a clinical study. It contains extracted pieces of information (namely, a set of parameters and a relation which links them) and meta-information (such as article and sentence numbers in which the finding was reported). The content items are then filtered against the patient record using a simple strategy. The strategy was reviewed and approved by the medical experts working on the PERSIVAL project. Next, the content items are organized: they are first merged into blocks of semantically related items using clustering. The similarity measure used for clustering is based on features of the content items: the more parameters two content items have in common, the more similar they are. Repetitions are identified as identical content items and contradictions are identified as content items with identical parameters but with contradictory relations. Once the blocks are produced, each block is then assigned a priority weight based on salient features of its content items. This determines the order in which the blocks will be displayed in the summary. The summary itself is obtained by using phrasal generation, relying on the verbalization used by the input articles of the different content items. The resulting summary is a coherent and fluent re-generated (not extracted) text. The filtering against the patient record allows for the individual tailoring. To date the language the summarizer generates is targeted at physicians. I now plan to focus on generating summaries that are appropriate to users with a different level of expertise, such as patients.

Summaries and the Users' Expertise Level

I model the generation as a text simplification process. Given a summary originally produced for expert users, the goal is to generate a simplified version readable by lay users. Text simplification is an incipient field of computational linguistics, and many challenges presently exist. While most of the content is already selected in the original summary, not all the pieces of information should be included in the simplified version. Some details might be too technical for instance, and including them in a simplified version would only confuse the reader. Conversely, it might be necessary to introduce additional content such as background knowledge to enhance the reader's comprehension. Wording is obviously affected by the simplification process, whether one looks at the syntactic or lexical level.

I propose to approach the simplification process as applying a set of rewriting rules which affect both the technical lexical items and the technical sentences as whole. I plan to learn the rewriting rules by relying on a comparable corpus of technical texts and their simplified versions targeted at lay users. Lexical rules will include term simplification and definition insertion, while sentence rules will consist of paraphrasing the input sentence into a lay equivalent.

The first step is to collect instances of technical/lay sentence pairs from the comparable corpus automatically so that learning on many instances can be done. In preliminary work (Barzilay & Elhadad 2003), we have designed an algorithm for aligning sentences in comparable texts, such as encyclopedia articles written for children and adults. Our method relies on a simple lexical similarity measure and the automatically identified topical structure of typical articles. However, in the medical domain, new challenges arise: there is very little lexical similarity between technical and lay texts, and the topical structure is harder to induce automatically. I plan to investigate ways to adapt the algorithm to these new challenges.

The next step will be to learn the actual rewriting rules. For lexical rewriting rules, I will investigate ways to choose between alternative verbalizations of technical terms based on their frequency counts in lay and technical corpora. I plan to research when a definition is needed for a specific term and how it can be inserted into the simplified text. At the sentence level, I plan to investigate methods for unsupervised acquisition of the rewriting rules based on the automatically aligned comparable corpus.

Finally, I plan to evaluate my summarizer overall both with physicians and lay users. For physicians, I plan to assess whether the summaries help them access relevant information efficiently. Working with cognitive scientists members of the PERSIVAL team, we have set up a set of medical scenarios in which physicians have to find some information in the literature in order to treat a specific patient. We plan to compare how easy it is for the physicians to find the relevant information when presented with (a) the summarizer's output, (b) a modified version of the summarizer which does not tailors for individual content, or (c) results of a search. For lay users, I plan to evaluate whether their comprehension is eased by the simplification. In the planned user study, lay users will be presented with a technical summary or its lay version and a set of comprehension questions. To date, I have evaluated individual components of my summarizer, such as the content extraction and the content filtering module. The evaluation of the sentence alignment and the validity of the rewriting rules is planned work.

Acknowledgments

I would like to thank Kathleen McKeown, Regina Barzilay and Smaranda Muresan for the useful discussions. This work is supported by the National Science Foundation Digital Library Initiative Phase II Grant No. IIS-98-17434.

References

Barzilay, R., and Elhadad, N. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 25–32.

Elhadad, N.; Kan, M.-Y.; Klavans, J.; and McKeown, K. 2004. Customization in a unified framework for summarizing medical literature. *Journal of Artificial Intelligence in Medicine*. to appear.