Generating Patient-Specific Summaries of Online Literature

Kathleen R. McKeown* and **Desmond A. Jordan†** and **Vasileios Hatzivassiloglou***

*Department of Computer Science
450 CS Building
Columbia University
New York, N.Y. 10027
{kathy, vh}@cs.columbia.edu

†Department of Anesthesiology
Columbia Presbyterian Medical Center
630 West 168th Street
New York, N.Y. 10032
{daj3}@columbia.edu

Abstract

Medical professionals increasingly use online resources to find journal articles of interest, looking for both the latest news in their specialty, and for articles that discuss results pertaining to patients currently under their care. We present a design for generating summaries of such online articles that are tailored to the characteristics of the patient under consideration. In this model, results of a document search are filtered to highlight articles that are clinically relevant to the patient by matching characteristics of the patients under study against characteristics found in the online patient record. Our summarization algorithm processes each relevant article and retrieves only those pieces of it that are appropriate to the individual patient; these are subsequently combined to produce the summary. We describe a user study which supports our summarization design, and present conclusions from it that indicate the users' expectations from a good summary. We also outline our approach to several general natural language problems that we have identified as particularly important for the summarization task.

1 Introduction

In the healthcare domain, physicians and residents increasingly turn online to find journal articles of interest. In addition to looking for the latest news in their specialty, they may also look for articles that discuss results pertaining to patients currently under their care. With the large volume of medical research articles currently published every year[1], even a specialist cannot easily keep up with all the developments in his or her field. Particularly for the resident and intern, who are still within the formal education process, it can be difficult to find literature that is clinically relevant to an individual patient.

The increased availability of online literature suggests it should now be easier to access information, even if this information is specialized and spread over several articles. In practice, however, online searches often flood users with more information than needed, much of it irrelevant. This severely impairs the efficiency of the retrieval process, which often is critical: feedback we received from medical personnel indicates that they would like to access the literature during the pre-operative stage and sometimes even during the operation itself. Access to the literature is especially important during these stages for education and practice, particularly for residents. Being abreast and aware of the most recent literature on a patient can enhance the provision of medical care.

Our approach exploits several characteristics of the healthcare domain to achieve a significant reduction of the information collected by a literature search. First, we rely on the online patient record, available through the Clinical Information System in place at Columbia Presbyterian Medical Center (Roderer and Clayton 1992), to exclude articles that do not match the patient currently under the user's care. Only articles which study patients with similar characteristics as the current patient are passed on to the summarizer, thus dramatically reducing the set of relevant articles. These are the articles that are most likely to be of interest to the medical professional at this particular moment; the online patient record serves as a readily available user model.

Our user-specific approach is further enhanced by our strategy for summarizing the individual articles

[1]For example, there are five journals which publish papers in the narrow specialty of cardiac anesthesiology, but 35 different anesthesia journals in general; approximately 100 journals in the closely related fields of cardiology (60) and cardiothoracic surgery (40); and over 1,000 journals in the more general field of internal medicine.

that *are* relevant. Much earlier work on summarization has focused on producing summaries that represent the salient points of the source articles, whether based on word counts (Luhn 1958), on extensions of these such as total frequency–inverse document frequency pairs (in many cases combined with additional discourse modules and semantic knowledge) (Paice 1990; Brandow *et al.* 1990; NetSumm 1996; Hovy and Lin 1997), or on more elaborate statistical models (Kupiec *et al.* 1995; Teufel and Moens 1997).[2] A common element in these previous systems is their uniform treatment of the source document's content. While this may be appropriate for newspaper reports, we have found that different users are interested in different parts of a medical article. While a medical researcher may want to focus on how a particular comparison between patient groups was set up and validated, a physician preparing for an operation or considering alternative drugs is only interested in the study's results, conclusions, and recommendations.[3] Thus, our summarization algorithm processes each relevant article and retrieves only pieces of it that are appropriate to the particular caregiver and patient; these are subsequently combined to produce the summary.

The selection of appropriate articles, parts of articles, and even journals and authors, is further influenced by a third component of our user model. Different specialists are interested in the same article, but frequently from a different perspective; an article about congestive heart failure may be of interest to both surgeons and anesthesiologists. While both groups may want a general understanding of the article's main results, they are likely to focus most of their attention on different parts of it. Our work investigates how we can take these user considerations into account at all levels of article, section, and text segment extraction.

In addition to this multi-level emphasis on user modeling, our work differs from other summarization approaches in that it investigates how text analysis techniques can do more than shallow processing of the article text to retrieve text fragments for inclusion in the summary, rather than whole sentences. A combination of statistical methods and textual, domain-dependent clues can be employed to identify the appropriate text segments. While full text understanding is still an elusive goal, the algorithm posits a reasoning component to resolve contradictions among the multiple articles that are presented in a single summary. Furthermore, text generation techniques, such as aggregation, will be applied to combine the selected fragments into grammatical and concise sentences.

The work we report on here is still in early stages. It is part of a multi-year effort at Columbia University to generate informed summaries that explore natural language understanding and generation techniques and tailor a presentation of retrieval results for the end user. While implementation of system components is under way, our focus at this stage of the project has been on a design that meets the needs of medical specialists. In this paper, we present our overall design for generation of patient-specific summaries, describe a feasibility user study that supports this design, and several summarization issues that this approach entails. Our design was developed through consideration of the capabilities of current natural language technology in tandem with the interests of medical specialists as outlined by the cardiac anesthesiologist on our team. We further validated our design through a feasibility study we performed for this model of summarization, where a prototype of the system was shown to several residents at Columbia Presbyterian Medical Center. Feedback from the users reveals several surprising facts about the medical professionals' expectations for a good summary, which in all likelihood are not limited to this specific domain.

[2] Space does not permit a more detailed presentation of the extensive prior work on summarization, which includes discourse-based approaches such as (Boguraev and Kennedy 1997) and (Marcu 1997).

[3] Assuming the study's results should be accepted *prima facie*, which is not always the case. We return to this issue of article validation, and especially to its manifestation through contradictions between multiple related articles, in Section 5.

2 Retrieving Patient-Specific Information From The Text

Our research is predicated on the observation that there is no such thing as as a single "best" summary of an article. Rather, the content and form of a summary depends on the task for which it will be used. For example, a well known distinction made by abstractors (Borko 1975) is whether the abstract is indicative (i.e., tells what the article is about) or informative (i.e., provides information from the article and can be read to get a gist of it). In the medical domain, where one purpose for reading an article is to find the latest medical techniques that are relevant to a specific patient case, the results of an experiment that specifically pertain to the patient characteristics are important. A quick read of these results and the way in which they relate to the particular patient can be enough to determine if the full article (e.g., experimental setup, statistical techniques, related work, etc.) should be read. Furthermore, reading the summary of results may entice the physician to read an article they may not otherwise have read and which they should have read. This is a form of informative summary since it includes specific information from the article itself.

Our task is made easier by the standard format used in journal articles in medicine. Provided that the article reports on experimental results (as opposed to a review article, for example), the standard structure of the paper includes the following conceptual sections: Introduction, Methods, Statistical Analysis, Results, Discussion, Previous Work, Limitations of the Study, and Conclusions. These conceptual sections frequently correspond to formally marked article divisions, with headings that provide strong clues about the content of each division. Even when we cannot rely on such formatting information, we expect that the presence of strong textual indicators can be used to automatically separate the sections of interest with a fair degree of success.

Furthermore, within a single section, a reader is guaranteed to find certain types of information. For example, within the Methods section, the author invariably includes a description of the patients who were included in the study, providing inclusion criteria and often, exclusion criteria as well. This can be as general as overall diagnosis (e.g., heart failure), but is usually accompanied by the ranges of values that patients had for specific lab results and diagnostic procedures (e.g., "evidence of stenosis at least 70%" or "ejection fraction measured by radioisotope method at most 0.35"). In some cases, this description is presented very briefly in one paragraph or less, while in others the description may go on for a page or more. While this variability complicates the task of identifying particular types of information, the prior knowledge of what is expected to be within each section greatly simplifies the design of the extraction subsystem.

Within this general paper structure, there are three types of information that indicate whether an article is relevant to a specific patient. The first is the goal of the paper, which often relates to the diagnosis of the patient. This information is almost always included in the introduction of the paper. For example, in our domain of cardiology and cardiac anesthesiology we have worked with a set of articles that all pertain to possible treatments of congestive heart failure. In one of these articles (Nul *et al.* 1997), we find the sentence

> "The prevalence of ventricular arrhythmias and sudden death in severe congestive heart failure (CHF) has focused attention on antiarrhythmic therapy as a means for reducing mortality."

while in another (Adams *et al.* 1996), the Introduction concludes with

> "To readdress this controversy, we studied whether the clinical expression and outcome of heart failure differed between men and women and whether gender was an independent risk factor ..."

The commonality between purposes indicates that congestive heart failure is the disease being studied. This fact should be included in the summary when the patient under consideration has congestive heart failure.

The second type of information is the patient characteristics that represent the patients in the study and which match the patient under consideration. Needless to say, not all characteristics mentioned in a paper will match a given patient, and not all of the descriptions provided will exactly match the form of patient information stored in the patient record. The summary should indicate which of the characteristics from the patient study are relevant.

In addition to forming part of the summary, matches against patient characteristics can also be used to determine relevance of an article. If an article describes a study on patients that match the patient record, it is more likely to be of interest than an article that does not match. Thus, we can also use this matching process to determine which of the articles returned by a search are more likely to be of interest to the user. Matching articles are ranked above non-matching articles and only matching articles are summarized.

The third type of information is the results of the study. Generally, results are stratified by patient group. Often there is a control group and a study group. Furthermore, each of these groups is divided into two classes of patients for whom the effects of therapy are measured. For example, one article (Nul *et al.* 1997) reports on the efficacy of a particular drug, amiodarone, for congestive heart failure patients. Results are reported separately for patients with a high heart rate (where the test hypothesis expects beneficial effects) and for patients whose heart rate is below a certain level, both within the study group (those who took amiodarone) and within the control group (those who did not).

In other cases, a study specifically examines the differences in, for example, prognosis across different groups. A second of our test articles that falls into this category examines gender differences (Adams *et al.* 1996). In yet other cases, different categories may be described in the section of the article that reports on results. For example, in this same article, the discussion of results presents a stratification by etiology (i.e., cause) of the disease.

In the summary, we include only results from the category that is pertinent to the characteristics of the patient. If we have an ischemic, female patient with a heart rate of 120 beats per minute, then we would only include results pertaining to the patients with high heart rate from Nul *et al.*'s article and only results pertaining to ischemic females from Adams *et al.*'s article.

The type of results that are reported depends on the type of study. We looked at three classes of articles: prognosis articles report on how well a patient with certain characteristics can be expected to do (e.g., mortality rate), diagnosis articles report on how well a certain test works to identify a disease or its complications, and treatment articles report on how a certain drug or procedure affects outcome. The class of article determines the type of result to look for. It is also useful to the medical professional in determining whether the article is of interest and thus information about the article's classification should be directly included in the summary.

3 Summarization Algorithm

Given the above summary design for retrieving matching information from the articles, our approach comprises the following steps:

1. *Match patient characteristics.* This involves three substeps:

(a) *Extract patient characteristics from the article.* We are using machine learning to build pattern matching rules to extract this information from the Methods section of the article. Thus, the Methods section must first be identified. As in the information extraction systems developed under the DARPA Message Understanding Program (MUC 1992), this requires looking for restricted types of information which typically is expressed using very similar syntactic forms. We are experimenting with two different methods. We are using the CIRCUS system developed at the University of Massachusetts (Lehnert *et al.* 1993) to learn the pattern matching rules. This approach requires hand tagging a set of training articles with semantic tags (e.g., "ejection fraction") and then running a learning component to build rules that capture all forms associated with those tags. This gives us the advantage of building on a fully developed, robust system for information extraction, but requires substantial hand tagging. The second method involves less manual effort, but runs the risk of being less accurate. Each semantic tag is associated with several key phrases by hand, and we identify patterns corresponding to each tag by measuring their recurrence rate within a corpus of articles that has been tagged by part of speech, noting when a pattern contains one of the key phrases associated with that particular semantic tag. Given the relatively stylized language used to refer to patient characteristics, this may be adequate.

(b) *Extract patient characteristics from the patient record.* In the cardiology domain, pertinent patient characteristics can be found in four different reports, the procedural report from the invasive cardiology unit (or cath lab, where catheterization procedures are done), the echocardiogram report, the radiology report, and the pre-operative history and physical. The first three of these follow a structured text format. Part of a sample cath lab report is shown in Figure 1. Given the table format, it is relatively easy to extract the needed values. The pre-operative history and physical is done by the anesthesiologist who uses an online menu-based system to check off conditions that the patient has (e.g., diabetes). Thus, this information is available in database form.

Hemodynamic Data

Heart Rate: 125

	Pressure (mmHg)			
	Sys	Dias	Mean	Sat
PA	45	26	34	
PCW	31	37	31	
AO	97	75	83	100%

For PA and PCW the three pressure values represent: A-Wave, V-Wave, Mean

Angiographic Data

Overall Status of Coronaries: Abnormal
Dominance: Right Dominance

	% Narrowing	Morphology	Collaterals
RCAM	70.00		
LADP	100.00	Thrombus	
DIA1	90.00	Ostial	
DIA2	90.00	Ostial	
LCXP	95.00		

Figure 1: Part of a sample cath lab report, indicating the rigid format used in such reports.

(c) *Intersect article patient characteristics with database patient characteristics.* When a substantial number of characteristics match (the exact threshold and weighting method will be experimentally determined), then the article is highly ranked and will be included in the summary. The matching patient characteristics are stored and will be included in list format in the summary. An estimate of the system's confidence in the matching decision should also be included in the summary.

2. *Categorize the article as either a prognosis, diagnosis, or treatment article.* We are exploring a combination of text categorization techniques with the use of linguistic clues. It is unlikely we can use text categorization techniques alone because our corpus of medical articles is not large enough. While there are

many online abstracts (e.g., in MEDLINE (1997)), the number of online full text articles in a narrow domain is much more limited. However, articles typically include specific phrases that indicate the category; for example, a treatment article usually refers to "therapy", "drugs", and "benefits".

It is worth noting that the classes of the articles are fluid. Building an ontology of article types is a non-trivial task, but we are experimenting with hierarchical subdivisions of the broad three-way categorization listed above. For example, treatment articles may be subdivided according to whether they involve drugs or surgery; it is quite possible that a resident may be interested primarily in one of these treatment types. Further elaboration of our article classes will take place according to user feedback.

3. *Identify patient stratification and extract results.* Stratification of the study groups is sometimes present in the Methods section and is explicitly described. For example, the second of the articles mentioned in Section 2 (Nul *et al.* 1997) is a treatment article, and contains the sentences

   > "The effect of BHR *[Base Heart Rate]* was evaluated by stratifying each group (amiodarone and control) according to a mean BHR $\geq$ 90 and $<$90 beats/min, respectively. Thus, four groups were established ..."

   In other cases, the stratification is implicitly described in the Results section. In the article on gender differences mentioned earlier (Adams *et al.* 1996), multivariate analysis is performed in an attempt to associate etiology (causes) of heart failure and gender with outcome. The results are presented in several pages which describe many different possible groupings. The challenge is to automatically identify the small set of result statements which appear to be of interest to physicians treating a particular patient. For example, the system must be able to extract these two non-adjacent sentences from the Results section of this article,

   > "The risk of death was similar for the subset of men and women with ischemic heart disease as the primary cause of heart failure."

   and

   > "In contrast, men had a significantly greater risk of death than women in the subset of patients judged to have a non-ischemic etiology of heart failure."

   and form the four possible groups of patients (depending on presence of ischemia and gender). Only one of the two sentences will be used in the summary depending on which group the patient falls into. We will use the same kind of learning of extraction rules as we described for extracting patient characteristics.

4. *Merge extracted sentence fragments.* The above steps result in a set of extracted phrases (e.g., "congestive heart failure") and a list of patient characteristics plus a set of sentences or sentence fragments extracted from the Results sections of the relevant articles. To produce the summary, post-processing the sentences using symbolic techniques to group them together is required. The summary starts with a sentence describing the general phenomenon (such as congestive heart failure) discussed in all the summarized articles; this sentence is built around the predominant phrase used in the objectives statement of the articles. The patient characteristics that match one or more articles are listed next, in list form.

   The extracted result sentences are grouped by article type (e.g., all result sentences from treatment articles will be placed together). Rephrasing may be required during this process, as there may be contradictions among the articles, or there may be repetition of the same or similar conclusions across several articles. The specific types of rephrasing we are addressing are discussed in Section 5.

4 User Study

To validate our approach to patient-tailored summarization, we have organized a feasibility study, presenting a prototype of the system to four physicians in cardiac anesthesiology at Columbia Presbyterian Medical Center (CPMC).[4] The residents provided us with reactions to the summaries that are both beneficial to our ongoing development of the system and illuminative of general characteristics that end-users (at least in the medical domain) expect from the summary.

For this task, we used data from the records of four former patients at the Cardiac Thoracic Intensive Care Unit at Columbia Presbyterian Medical Center. A selection of six relevant articles was extracted from the online literature by the second author, thus bypassing the generic information retrieval step which initially procures the documents that are to be summarized. The articles were selected by hand so that the physicians' evaluations of the summaries was not obscured by incorrectly retrieved documents; our task is not document retrieval. In the final system, we plan to use an off-the-shelf document retrieval system, to which the summarizer will serve as a front end.

The patients were selected to ensure presentation of interesting, non-trivial, and significantly different cases

[4] We thank cardiac anesthesia fellows Shabina Ahmad, Terry Koch, and Jonathan Oster, and Carmen Domingez, an anesthesia attending physician, for participating in this study.

to the physicians. The variability requirement was set so that the number of features of our system that would be tested would be maximized.

For each patient, a summary was constructed from the relevant subset of the articles. This subset was identified, and sentences and sentence fragments were extracted from it, according to the methods outlined in Section 3. Since neither our matching nor information extracting components were operational at the time, we constructed the summaries by following the algorithm but instantly resolving the difficult recognition and synthesis natural language problems that arose in the way a human would do. Thus, the user study can show us qualitatively whether our approach is a valid one, decoupling the evaluation of the summarization algorithm from the lower level natural language problems. We can then focus separately on these problems, to bring the actual system's performance close to human standards.

We discuss two of the four patients and the corresponding summaries here. We have implemented a user interface for the summarization system, and asked the four residents to experiment with it. The user enters the patient name or internal identification number, and an optional set of keywords; the system then retrieves matching articles along with author and publication information, ranks them according to their relevancy to this particular patient, and produces the summary of the relevant articles. The user can click on each article's title to read its full text. Hypertext links to the full text of the article(s) which support each statement in the presented results are also provided. For the purposes of this user study, the search results and the summaries were preconstructed as described above. But this was not revealed to the users, as we wanted to get an accurate feeling of what they would expect from a fully automated summarization system.[5] Figure 2 displays the interface and the summary for the first patient as it was presented to the residents. Figure 3 shows the summary for the second patient.

The records of both patients have values for the left ventricular ejection fraction (which indicates the level of heart function) and chest radiograph cardiothoracic ratio (which indicates heart size) that suggest they have heart failure. (Other measures, such as NYHA (New York Heart Association) class assignment, which indicates severity of heart failure, and end-diastolic echocardiographic diameter, which indicates size of the heart right before it pumps, also support this diagnosis). The first patient has results from an electrocardiogram (EKG) which indicate that he could possibly benefit from a pacemaker (these are the mean PR interval and first degree AV block, which are included in the produced summary). He has a history of ischemia and has a base heart rate of 85. EKG readings for the second patient do not indicate the need for a pacemaker, but this patient has a high base heart rate (> 90), is not ischemic, and has no evidence of coronary artery disease.

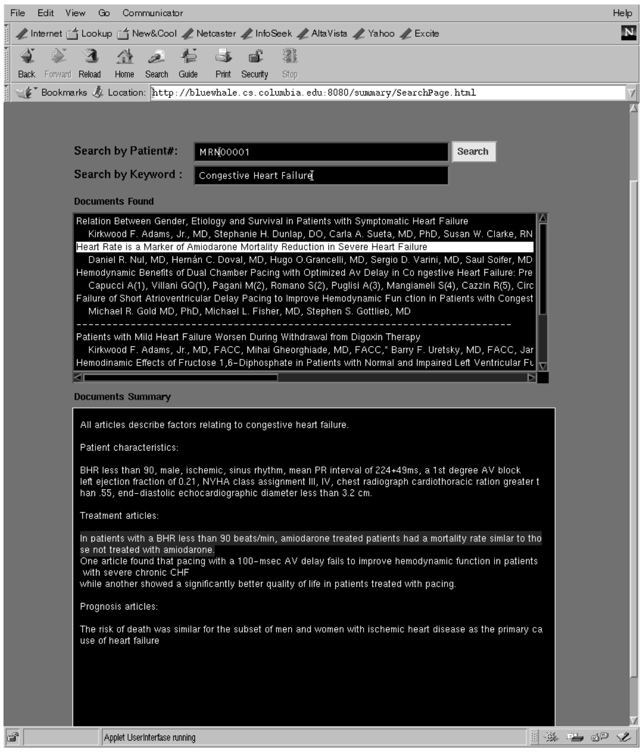

Figure 2: The user interface, displaying the summary for the first patient.

The physicians viewed our approach to summarization as a useful one that provides necessary information they don't otherwise easily get. They typically spend a significant portion of their time accessing journal articles, and selecting which ones to read is a major concern. They expressed frustration with current information retrieval technology, as they usually have to wade through a large number of unrelated documents to get to the few ones they really want. As a result, they were very supportive of our focus on patient- and user-centered summaries. The fact that sentences were used out of context and in list form, or that only a portion of the article's results were presented, was positively received and not viewed as a problem. They can quickly scan the author-provided abstracts and decide whether they should look at the full text, but they found our short summaries that aggregate results from multiple documents a more convenient means for obtaining data and filtering articles. People found summaries that pointed out contradictions between articles were particularly effective; contradictions were an en-

[5] None of the users found the supposedly automatic construction of perfectly natural text summaries suspicious.

All articles describe factors relating to congestive heart failure.

Patient characteristics: BHR $>$ 90, male, no coronary artery disease, left ejection fraction of 0.21, NYHA class assignment III, IV, chest radiograph cardiothoracic ration $>$ .55, end-diastolic echocardiographic diameter $>$ 3.2 cm.

Treatment articles:

The 2 year mortality rate was higher in patients with a BHR $\geq$ 90 beats/min compared with patients with a BHR $<$ 90 beats/min. A significant reduction in sudden death was observed in patients with a BHR $\geq$ 90 beats/min. treated with amiodarone. {*article #4*}

Prognosis articles:

Men without coronary artery disease had a greater risk of death than women ($N = 369$; men vs. women: RR 2.38, 95% CI 1.58 to 3.58, $p <$ 0.001). {*article #1*}

Figure 3: Summary produced for the second patient.

ticement to issue further searches or read the articles. They also felt that the summaries would be even more time efficient when a larger number of articles are summarized. They suggested further reductions in the set of retrieved articles and the length of the produced summary, with decisions based on the quality of the journal where the article appears. Our classification of articles as prognosis, diagnosis, or treatment was regarded as a positive feature that helps them look at the specific results they are interested in. Extending and refining this classification of article types, tailoring the output even more to the specialty of the user (e.g., anesthesiologist versus surgeon) and the task at hand (e.g., displaying only prognosis articles), and aggregating similar statements from multiple documents were viewed as highly important.

At the same time, the physicians could not easily separate the functionality of the summarization component from its sister information retrieval and database matching (for patient records) components. On receiving a summary of a set of articles, the physicians often wanted to explore literature on related topics. They expressed disappointment when the articles retrieved were not exactly what they were looking for, when the system could not produce additional articles as they varied the keywords, or when it would not list the information from the patient record that wasn't relevant to any of the articles in our collection. While this indicates the benefits of the proposed system in encouraging browsing of the literature for educational purposes, it also indicates problems in integrating summarization with information retrieval. These concerns are not the responsibility of the summarizer, but they are indicative that summarization is strongly tied to information retrieval in the users' eyes. The saleability of natural language summarization technology in the real world, even as it succeeds in overcoming technical problems, may thus be less than what we think, given the current quality of retrieval systems.

To further test our hypothesis that users want narrowly focused summaries, which tend to be much shorter than typical informative summaries, we prepared one longer summary for the two articles that relate to the fourth patient's characteristics. This summary was comparable to a normal abstract, and was met with dismay by our evaluators. They stated that if they were going to get a summary about half as long as the two papers' abstracts combined, they would prefer to quickly scan the abstracts instead. It appears that a summarization system needs to offer a significantly higher rate of reducing information overload before the users will concede their powers of article selection and relevance verification to it.

5 Some Summarization-Related Natural Language Issues

Throughout the development of our summarization method and from the analysis of the users' reactions to its prototype version, we have identified several natural language processing issues that bear particularly on the summarization task. These problems are not related to the main architecture and algorithm discussed in Sections 2 and 3; rather, they become a concern when the automated system has to perform tasks at the sentence level which seem trivial to a human speaker of the language.

Overall, our approach results in a promising new strategy for intelligent summarization that goes beyond the usual across-the-board sentence extraction. Summarization is made do-able, without having to go all the way to full semantic analysis, as seems to be the case with the much harder and more general domain independent case. We still work at the sentence and phrase rather than the single word level, but by taking advantage of appropriate textual clues and prior domain knowledge, we reduce the problem mainly to the three tasks of stratifying, classifying, and matching the appropriate sentences. Yet, several natural language related problems remain.

Contradictions across multiple related documents Whenever a summarizer processes multiple documents, there is the chance that conflicting results would be extracted from two or more of these. This may be due to mistakes during the text analysis phase; for example, the summarizer may misclassify a particular sentence in one of the articles as belonging to the wrong stratum and thus attempt to apply its recommendation to the wrong group of people. However, there are also genuine cases where the two documents disagree; for factual statements, at least one of the documents must then be wrong.

Humans can reason about the supporting evidence each article offers, and resolve such cases in favor of one of them. This is very hard to do automatically. Instead, operators that combine the disagreeing statements and provide them to the user for further analysis can be employed, as has been suggested in the past (McKeown and Radev 1995).

Our system adopts this approach. For one of the patients in our user study, we extracted conflicting statements; the produced summary reads, in part,

> "One article found that pacing with a 100-msec AV delay fails to improve hemodynamic function in patients with severe chronic CHF {*article #11*}, while another showed a significantly better quality of life in patients treated with pacing {*article #10*}."

providing hypertext links to the full text documents by which the user can examine the competing evidence. The problem for our summarization system is to identify that a contradiction between sentences exists, to place them next to each other, removing repetitive portions, and selecting an appropriate connective. Detecting contradictions from full text is quite difficult. We will limit the types of contradictions we can detect, using clues from phrasing such as "fails to improve" and "showed a significantly better quality of life."

Aggregation Aggregating multiple related statements, whether similar or competing ones, is akin to the problem of conflict resolution in that some information has to be lost for the sake of conciseness. Sometimes the statements can be simply connected with *and*; but in other cases, especially those involving many related statements, a trade-off between the accuracy of the description and its coverage of the original statements must be adopted. The evaluators strongly prefer a description like "Most of the papers support treatment A although a few adopt treatment B" to an alternative like "Eight documents recommend treatment A and three treatment B; yet, treatment C is supported by two documents and treatments D and E by one each".

We are borrowing from and extending our earlier work on aggregation (Passonneau *et al.* 1996) to the cardiology domain to select a solution that best satisfies the competing goals of coverage, specificity, homogeneity of descriptions, and conciseness. An objective function scores potential descriptions against these goals, and an optimization algorithm identifies the description that maximizes this score.

Reference resolution and rephrasing Another common natural language problem that reappears in the context of summarization is anaphora resolution, both for pronouns and definite noun phrases. Frequently, the authors of the article define a particular expression (such as "the control group" of patients) early in the document, and then refer to it much later in the results and recommendations. For example, one of the articles in our collection (Leschke *et al.* 1996) defines two groups, A and B, of patients that received different treatments with the sentences

> "Ninety-eight patients with chronic refractory and end-stage coronary artery disease were randomly assigned to two treatment groups: group A (49 patients) received 50,000 IU and group B (49 patients) 500,000 IU of urokinase ..."

near the start of the document. It then proceeds to refer to groups A and B throughout the article, including, for example,

> "At the end of the treatment period, 7% were completely asymptomatic in group A compared with 17% in group B."

in the Results section much later. Unless the system can successfully identify the referents for "group A" and "group B" in the last sentence, it cannot extract the fact that larger doses of urokinase help the patients.

The importance of anaphoric resolution is not limited to our summarization method. Other methods, based on sentence extraction, also encounter it and generally either miss the sentences that contain such references completely, or include them in the summary resulting in dangling references. Sometimes, the summaries are extended to include earlier sentences than those which are assigned the best content scores, in an attempt to include, by luck, any such antecedents (Paice 1990). We believe that unresolved anaphoric references may contribute to the well-established fact that the first sentence of any article is an excellent candidate for an informative summary. In a separate summary evaluation study performed at Columbia University (Jing *et al.*

1998), some of the human constructors of summaries reported that they used the first sentence of the article not necessarily because it was so good content-wise, but because dangling references prevented them from selecting subsequent sentences in the first paragraph without including the very first sentence too.

We are currently implementing a general tool for the resolution of anaphora, which is now operational for pronominal anaphora and will be extended to definite anaphora. Definite anaphora resolution requires a much higher amount of lexical semantic and world knowledge, so it is practical for limited domains only. We will replace anaphoric expressions in the text with their automatically identified antecedents before processing the articles through the summarizer. This will enable the matching of sentences with such expressions, and also make these sentences understandable when they are presented in the summary out of their original context.

6 Conclusions

Our work proposes a design for patient-specific summarization, supported by a user feasibility study, that is quite different from current approaches to summarization. It exemplifies the position that the content and form of a summary must be shaped according to the specific task it will be used for, as well as the genre of articles it summarizes. This is in contrast to the more typical, statistical sentence extraction system which generically considers every sentence with equal potential for extraction, depending on various measures related to its word frequencies. Instead, we use the patient record, the task of the particular specialist, and the type of article (whether prognosis, diagnosis, or treatment) to determine the information that will be included. The resulting design uses focused sentence extraction, looking for particular types of sentences in particular places, and is enhanced with the use of symbolic generation techniques to aggregate and rephrase in certain specific cases.

Acknowledgements

We are grateful to the four cardiac anesthesiology physicians, Shabina Ahmad, Carmen Domingez, Terry Koch, and Jonathan Oster, who provided us with valuable feedback during the user study. We also thank Federico Kattan, who implemented the user interface in Java, and helped with the collection of medical articles and abstracts from various online sources. The research work is supported in part by the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare, funded by the New York State Science and Technology Foundation.

References

(Adams *et al.* 1996) Kirkwood F. Adams, Stephanie H. Dunlap, Carla A. Sueta, Susan W. Clarke, James Herbert Patterson, Mary Beth Blauwet, Lynda R. Jensen, Lisa Tomasko, and Gary Koch. Relation between Gender, Etiology and Survival in Patients with Symptomatic Heart Failure. *Journal of the American College of Cardiology*, **28**(7):1781–8, December 1996.

(Boguraev and Kennedy 1997) Branimir Boguraev and Christopher Kennedy. Salience-based Content Characterisation of Text Documents. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 2–9, Madrid, Spain, August 1997. Association for Computational Linguistics.

(Borko 1975) Harold Borko. *Abstracting Concepts and Methods*. Academic Press, New York, 1975.

(Brandow *et al.* 1990) Ronald Brandow, Karl Mitze, and Lisa F. Rau. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, **26**:135–170, 1990.

(Hovy and Lin 1997) Eduard Hovy and Chin Yew Lin. Automated Text Summarization in SUMMARIST. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 18–24, Madrid, Spain, August 1997. Association for Computational Linguistics.

(Jing *et al.* 1998) Hongyan Jing, Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *Proceedings of the AAAI 1998 Spring Symposium on Intelligent Text Summarization* (this volume), Stanford, California, March 1998. American Association for Artificial Intelligence.

(Kupiec *et al.* 1995) Julian M. Kupiec, Jan Pedersen, and Francine Chen. A Trainable Document Summarizer. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, July 1995.

(Lehnert *et al.* 1993) Wendy Lehnert, Joe McCarthy, Stephen Soderland, Ellen Riloff, Claire Cardie, Jonathan Peterson, and Fangfang Feng. UMass/Hughes:

Description of the CIRCUS System Used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 277–291, Baltimore, Maryland, August 1993.

(Leschke *et al.* 1996) Matthias Leschke, Frank-Chris Schoebel, Wolfgang Mecklenbeck, Dirk Stein, Thomas Walter Jax, Hans-Wilheim Müller-Gärtner, and Bodo-Eckehard Strauer. Long-Term Intermittent Urokinase Therapy in Patients With End-Stage Coronary Artery Disease and Refractory Angina Pectoris: A Randomized Dose-Response Trial. *Journal of the American College of Cardiology*, **27**(3):575–84, March 1996.

(Luhn 1958) Hans P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal*, pages 159–165, 1958.

(Marcu 1997) Daniel Marcu. From Discourse Structures to Text Summaries. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, August 1997. Association for Computational Linguistics.

(McKeown and Radev 1995) Kathleen R. McKeown and Dragomir Radev. Generating Summaries of Multiple News Articles. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, July 1995.

(MEDLINE 1997) National Library of Medicine. MEDLINE. World Wide Web site, URL: `http://www.nlm.nih.gov/databases/medline.html`. Updated weekly January through October and twice in the month of December, 1997.

(MUC 1992) DARPA Software and Intelligent Systems Technology Office. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, 1992.

(NetSumm 1996) British Telecom Laboratories. NetSumm Home Page. World Wide Web site, URL: `http://www.labs.bt.com/innovate/informat/netsumm/index.htm`, 1996.

(Nul *et al.* 1997) Daniel R. Nul, Hernan C. Doval, Hugo O. Grancelli, Sergio D. Varini, Saul Soifer, Sergio V. Perrone, Noemi Prieto, and Omar Scapin. Heart Rate is a Marker of Amiodarone Mortality Reduction in Severe Heart Failure. *Journal of the American College of Cardiology*, **29**(6):1199–205, May 1997.

(Paice 1990) Chris D. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management*, **26**:171–186, 1990.

(Passonneau *et al.* 1996) Rebecca J. Passonneau, Karen K. Kukich, Jacques Robin, Vasileios Hatzivassiloglou, Larry Lefkowitz, and Hongyan Jing. Generating Summaries of Work Flow Diagrams. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, pages 204–210, New Brunswick, Canada, June 1996. University of Moncton.

(Roderer and Clayton 1992) N. Roderer and P. Clayton. IAIMS at Columbia Presbyterian Medical Center: Accomplishments and Challenges. *Bulletin of the American Medical Libraries Association*, pages 253–262, July 1992.

(Teufel and Moens 1997) Simone Teufel and Marc Moens. Sentence Extraction as a Classification Task. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain, August 1997. Association for Computational Linguistics.