Automatic Analysis of Medical Dialogue in the Home Hemodialysis Domain: Structure Induction and Summarization

Ronilda C. Lacson, MD, MS, Regina Barzilay, PhD, William J. Long, PhD

Computer Science and Artificial Intelligence Laboratory (CSAIL)

Massachusetts Institute of Technology (MIT), Cambridge, MA

Correspondence and reprints:

Ronilda C. Lacson, MD, PhD

32 Vassar Street, Room 32-258, Cambridge, MA 02139

e-mail: rclacson@mit.edu

phone: (617) 253-0287

fax: (781) 652-0404

Abstract

Spoken medical dialogue is a valuable source of information for patients and caregivers. This work presents a first step towards automatic analysis and summarization of spoken medical dialogue. We first abstract a dialogue into a sequence of semantic categories using linguistic and contextual features integrated in a supervised machine-learning framework. Our model has a classification accuracy of 73%, compared to 33% achieved by a majority baseline ($p<0.01$). We then describe and implement a summarizer that utilizes this automatically induced structure. Our evaluation results indicate that automatically generated summaries exhibit high resemblance to summaries written by humans. In addition, task-based evaluation shows that physicians can reasonably answer questions related to patient care by looking at the automatically-generated summaries alone, in contrast to the physicians' performance when they were given summaries from a naïve summarizer ($p<0.05$). This work demonstrates the feasibility of automatically structuring and summarizing spoken medical dialogue.

Keywords: spoken dialogue analysis, medical discourse, summarization, natural language processing, information processing

I. Introduction

Medical dialogue occurs in almost all types of patient-caregiver interaction, and forms a foundation for diagnosis, prevention and therapeutic management. In fact, studies show that up to 80% of diagnostic assessments are based solely on the patient-caregiver interview.[1] Automatic processing of medical dialogue is desirable in multiple contexts – from clinical and educational, to financial and legal. Caregivers can use the results of this processing for informed decision-making, researchers can benefit from large volumes of patient-related data currently unavailable in medical records, and health care providers can enhance communication with patients by understanding their concerns and needs. All of these users share a common constraint: none of them wants to wade through a recording or transcript of the entire interaction.

To illustrate the difficulty of accessing medical dialogue, consider 30 seconds of an error-free transcript of an interaction between a dialysis patient and a nurse (see Figure 1). This excerpt exhibits an informal, verbose style of medical dialogue – interleaved false starts (such as "**`I'll pick up, I'll give you a box of them`**"), extraneous filler words (such as "**`ok`**") and non-lexical filled pauses (such as "**`Umm`**"). This exposition also highlights the striking lack of structure in the transcript: a request for more supplies (e.g. "**`kidney`**", which in this context refers to a dialyzer) switches to a question about a patient's symptom (e.g. shoulder pain) without any visible delineation customary in written text. Therefore, a critical problem for processing dialogue transcripts is to provide information about their internal structure.

This paper presents the first attempt to analyze, structure and summarize dialogues in the medical domain. Our method operates as part of a system that analyzes telephone consultations between nurses and dialysis patients in the home hemodialysis program at Lynchburg Nephrology, the largest such program in the United States.[2] By identifying the type of a turn – Clinical, Technical, Backchannel or Miscellaneous – we are able to render the transcript into a structured format, amenable to automatic summarization. The Clinical category represents the patient's health, the Technical category encompasses problems with operating dialysis machines, the Miscellaneous category includes mostly scheduling and social concerns, while Backchannels capture greetings and acknowledgments.

In addition, automatically processing medical dialogue has important implications for the development and evaluation of conversational systems. Current methods for developing automated dialogue systems rely on large amounts of labeled data for training;[3] human annotation of this material is an expensive and lengthy process. Our system can provide an initial annotation which can be further refined by a human, if necessary. Furthermore, for evaluation of automated dialogue systems, structure of the dialogue can be analyzed and compared to human-human dialogues. An interesting direction in analyzing the performance of automated dialogue systems is their comparison with human-human dialogues. Understanding similarities and differences in structure between human-human and machine-human dialogues can further advance the development of automated systems. Our method may also be used for mixed conversational systems, in which part of the dialogue is routed to an automated system (i.e. scheduling), as opposed to a clinical or technical query, which requires the attention of a human caregiver. Finally, our classification allows a provider to distill the portions of the dialogue that support medical reasoning and are of primary interest to clinicians. In the long run, knowing the distribution of patient requests can improve the allocation of resources, and ultimately provide better quality of health care.

Our system has two main components:

Structure Induction We present a machine learning algorithm for classifying dialogue turns with respect to their semantic type. The algorithm's input is a transcription of spoken dialogue, where boundaries between speakers are identified, but the semantic type of the dialogue turn is unknown. The algorithm's output is a label for each utterance, identifying it as Clinical, Technical, Backchannel and Miscellaneous. Our algorithm makes this prediction based on a shallow meaning representation encoded in lexical and contextual features. We further improve the classification accuracy by augmenting the input representation with background medical knowledge.

Summarization We introduce a novel way to extract essential dialogue turns within our domain of spoken medical dialogue using the discourse structure just described. Our goal is to provide a caregiver with a succinct summary that preserves the content of a medical dialogue, thereby reducing the need to leaf through a massive amount of unstructured and verbose transcript.

In order to assess the performance of the summarizer and the contribution of structure induction, we describe a framework for evaluation of medical dialogues. Our first evaluation method follows an intrinsic methodology, commonly used in the text summarization community.[4] We compare automatically-generated summaries with a “gold standard” summary created by humans, assuming that a better automatic summary exhibits high overlap with a “gold standard” summary. Our second evaluation is task-based. Doctors were asked to use our summaries to answer questions concerning various aspects of patient care, ranging from clinical assessment to scheduling issues. We compare their responses to randomly-generated summaries and to a “gold standard” summary.

II. Related Work

In recent years, a variety of summarization algorithms have been developed for text,[5,6] and are primarily applied for summarizing newspaper articles.[7,8] Our work builds on these approaches in the design of a summarization algorithm for medical dialogues. For instance, some of the features used in our algorithm, such as position information and sentence length, have been shown useful in summarization of written materials.[9,10,11]

Our emphasis on spoken discourse sets us apart from the efforts to interpret written medical text.[12,13,14] In particular, our work differs in two significant directions:

1. The essential component of our method is structural representation of dialogue content, tailored to the medical domain. We show that this scheme can be reliably annotated by physicians, effectively computed and integrated within a summarization system.
2. We propose a novel task-based evaluation method that assesses usefulness of our summaries in the medical setting. Research in text summarization has revealed that designing a task-based evaluation is challenging; frequently a task does not effectively discriminate between systems. In contrast, we show that our task-based evaluation does not suffer from this drawback, and can be used to evaluate other summarization systems for medical dialogues.

III. Method

A. Data Collection

We collected our data from the Lynchburg Nephrology program, the oldest and largest home hemodialysis program in the United States.[2] All phone conversations between nurses and 25 adult patients treated in the program from July to September of 2002 were recorded using a telephone handset audio tap ("QuickTap", made by Harris, Sandwich, IL) and a recorder. The home hemodialysis nurses recorded the conversations whenever a call was made and stopped the recorder when the conversation ended.

All patients and nurses whose questions and answers were recorded read and signed an informed consent form approved by the MIT Committee on the Use of Humans as Experimental Subjects. At the end of the study period, we received a total of six cassette tapes, consisting of 118 phone calls, containing 1,574 dialogue turns with 17, 384 words. The conversations were transcribed, maintaining delineations between calls and speaker turns. The data were then divided chronologically into training and testing sets. The distribution of semantic types for each set is shown in Table 1.

B. Structure Induction

1. Semantic Taxonomy

Our annotation scheme was motivated by the nature of our application – analysis of phone consultations between a nurse and a dialysis patient. It is defined by four semantic types –

Clinical, Technical, Backchannel and Miscellaneous. Examples of utterances in each semantic type are shown in Table 2.

Dialogue turns are labeled Clinical if they pertain to the patient's health, medications, laboratory tests (results) or any concerns or issues the patient or nurse has regarding the patient's health. These discussions become the basis from which a patient's diagnostic and therapeutic plans are built. Dialogue turns are labeled Technical if they relate to machine problems, troubleshooting, electrical, plumbing, or any other issues that require technical support. This category also includes problems with performing a procedure or laboratory test because of the lack of materials, as well as a request for necessary supplies. Utterances in the Technical category typically do not play a substantial role in clinical decision-making, but are important for providing quality health care. We label as Miscellaneous any other concerns primarily related to scheduling issues and family concerns. Finally, the Backchannel category covers greetings and confirmatory responses, and they carry little information value for health-care providers.

Kappa agreement

Two domain experts, specializing in Internal Medicine and Nephrology, independently labeled each dialogue turn with its semantic type. Each annotator was provided with written instructions that define each category and was given multiple examples (see Appendix A). To validate the reliability of the annotation scheme, we computed agreement using the kappa coefficient.[15] Complete agreement would correspond to a kappa of 1.0. We computed the kappa to be 0.80, which is “substantial” agreement.[15] This kappa suggests that our dialogue can be reliably annotated using the scheme we developed.

2. Basic Model

Our goal is to identify features of a dialogue turn that are indicative of its semantic type and effectively combine them. Our discussion of the selected features is followed by a presentation of the supervised framework for learning their relative weights.

Feature selection: Our basic model relies on three features that can be easily extracted from the transcript: words of a dialogue turn, its length and words of the previous turn.

Lexical Features Clearly, words of an utterance are highly predictive of its semantic type. We expect that utterances in the Clinical category would contain words like "**`pressure`**", "**`pulse`**" and "**`pain`**", while utterances in the Technical category would consist of words related to dialysis machinery, such as "**`catheter`**" and "**`port`**". To capture colloquial expressions common in everyday speech, our model includes bigrams (e.g. "I am") in addition to unigrams (e.g. "I").

Durational Features We hypothesize that the length of a dialogue turn helps to discriminate certain semantic categories. For instance, utterances in the Backchannel category are typically shorter than Technical and Clinical utterances. The length is computed by the number of words in a dialogue turn.

Contextual Features Adding the previous dialogue turn is also likely to help in classification, since it adds important contextual information about the utterance. If a dialogue is focused on a Clinical topic, succeeding turns frequently remain Clinical. For example, the question "**`How are you doing?`**" might be a Backchannel if it occurs in the beginning of a dialog whereas it would be considered Clinical if the previous statement is "**`My blood pressure is really low.`**"

Feature weighting and combination: We learn the weights of the rules in the supervised framework using Boostexter,[16] a state-of-the-art boosting classifier. Each object in the training set is represented as a vector of features and its corresponding class. Boosting works by initially learning simple weighted rules, each one using a feature to predict one of the labels with some weight. It then searches greedily for the subset of features that predict a label with high accuracy. On the test data set, the label with the highest weighted vote is the output of the algorithm.

3. Data Augmentation with Background Knowledge

Our basic model relies on the shallow representation of dialogue turns, and thus lacks the ability to generalize at the level of semantic concepts. Consider the following scenario: the test set consists of an utterance "`I have a headache`" but the training set does not contain the word "`headache`." At the same time, the word "`pain`" is present in the training set, and is found predictive of the Clinical category. If the system knows that "`headache`" is a type of "`pain`", it will be able to classify the test utterance into a correct category. In our previous work, we described methods that bridge this gap by leveraging semantic knowledge from readily available data sources.[17] These methods identify the semantic category for each word, and use this information to predict the semantic type of a dialogue turn.

Our best algorithm derives background knowledge from clusters of semantically-related words automatically computed from a large text corpus. Clustering provides an easy and robust solution to the problem of coverage as we can always select a large and stylistically appropriate corpus for cluster induction. This is especially important for our application, since patients often use colloquial language and jargon. In addition, similarity based clustering has been successfully used in statistical natural language processing for such tasks as name entity recognition and language modeling.[18,19]

To construct word classes, we employ a clustering algorithm that groups together words with similar distributional properties.[18] In our experiments, we applied clustering to a corpus in the domain of medical discourse that covers topics related to dialysis. We downloaded the data from a discussion group for dialysis patients available in the following url: http://health.groups.yahoo.com/group/dialysis_support. Our corpus contains more than one million words corresponding to discussions within a ten-month period. We empirically determined that the best classification results are achieved for 2000 clusters. We added cluster-based substitutions to the feature space of the basic model by substituting each word of text with their corresponding cluster identifier. An example of a cluster is shown in Figure 2. When feature space is augmented with clustering information (computed outside of Boostexter), the number of features is increased by the number of clusters.

We also used semantic types from a large-scale human-crafted resource, UMLS, for data augmentation. Unfortunately, the results we obtained are less successful than those using word clusters.[a]

4. Results of Semantic Type Classification

Table 3 displays the results of various configurations of our model on the 293 dialogue turns of the test set, held out during the development time. The basic model and the knowledge-augmented model are shown in bold. All the presented models significantly outperform the 33.4% accuracy ($p<0.01$) of a baseline model in which every turn is assigned to the most frequent class (Clinical). The best model achieves an accuracy of 73%, and it combines lexical,

[a] Lacson R, Barzilay R. Automatic processing of spoken dialogue in the home hemodialysis domain. AMIA Annual Fall Conference, 2005.

durational and contextual features augmented with background information obtained through statistical clustering.

The first four rows of Table 3 show the contribution of different features of the basic model. Words of the dialogue turn alone combined with both the length of the turn and the words of the previous utterance achieve an accuracy of 70%. Table 4 shows the most predictive features for each category. The last row in Table 3 demonstrates that adding background knowledge improves the performance of the model modestly, achieving a 3% gain over the basic model. Even at the current level of performance (73%), we are able to use this model's predicted semantic types to generate summaries that are comparable to manual summaries created by physicians (see Section V).

In an effort to determine whether predictive features are easily identifiable by humans, we also compare the accuracy of our model to that of a classification algorithm that uses manually identified words chosen from our data.[20] In this approach, a domain expert manually identifies the most predictive features for each category, instead of automatically learning it from the data. For every category, an expert assigned a set of representative words. The weight of the words is determined by the count of instances it occurred in the training data for that particular category. When a new data segment is presented, we computed the score for each class by summing the word scores derived from the training data for every word in the segment that appears in the expert's list. The class with the highest vote wins. This model achieves an accuracy of 61%. This unexpectedly low result demonstrates the complexity of semantic annotation for medical dialogues, and justifies the use of machine learning methods.

C. Summarization

In the next section, we describe our method for automatically extracting key dialogue turns using the semantic types we defined. The extracted dialogue turns will comprise the summary for each dialogue.

Telephone dialogues between caregivers and patients may provide additional information to the health team for individual management of patients as well as for identifying the bulk of patient requests. Availability of this data is important for continuity of individual patient care as well as for proper allocation of health resources. However, an entire transcript of dialogue is not helpful for caregivers who are often pressed for time. Summarized versions of the transcript, which preserves the main contents, will provide the information in a more concise form.

Our extraction method consists of three consecutive steps:

Step 1: Remove Backchannels – By definition, backchannels contain greetings and acknowledgements that carry very little information value for health care providers. Removing backchannels should not affect the quality of information that is essential in summarization. Examples of backchannels are "**`Hello.`**", "**`Hi, is Martha there?`**", "**`That's ok.`**" and "**`Thank you.`**" We remove all backchannels from the dialogues at the beginning of the process. After this, each dialogue only contains dialogue turns from the following three categories: Clinical, Technical and Miscellaneous.

Step 2: Dialogue Segmentation – Our manual corpus analysis revealed that a typical dialogue in our domain contains more than one topic.[21] Therefore, a summary has to include dialogue turns representative of each topic. We computed topics by segmenting a dialogue into blocks of consecutive turns of the same semantic type. In other words, consecutive dialogue turns with the same semantic type are considered to belong to a segment with a single topic. For instance, a dialogue with six turns of the following type “clinical, clinical, miscellaneous, technical, technical, miscellaneous” is abstracted into a sequence of four topics “clinical, miscellaneous, technical, miscellaneous. An example of such segmentation is shown in Figure 3.

Step 3: Dialogue Turn Extraction – Next, we extract key utterances from each segment. Following a commonly used strategy in text summarization, we select the leading utterance of each segment.[22] We hypothesize that the initial utterance in a segment introduces a new topic and is highly informative of the segment’s content.

This extraction strategy may be deficient for long segments since such segments may discuss several topics of the same semantic type. For instance, a patient may discuss his vital signs while doing dialysis and then proceed to talk about back pain. Thus, for segments with more than two dialogue turns, we select the longest dialogue turn in addition to the initial one. We hypothesize that introducing a new topic will contain a lot of new information and will therefore contain more words. Figure 3 shows one run of the algorithm. The summarizer compresses a conversation of 14 into five key dialogue turns.

Predicted Semantic Type vs. True Semantic Type

Our summarization takes as input a dialogue in which every turn is annotated with its semantic type. An obvious way to obtain this information is to use an automatic classification method described in Section B for generating semantic types for each dialogue turn. We refer to these automatically generated labels as “predicted semantic types.” In our experiments, we also consider summaries that use “true semantic types,” that is, types manually assigned by human experts to each dialogue turn. Analyzing the performance of the model based on the “true semantic types” would allow us to measure whether structural information helps. Comparing summaries based on “true semantic types” with summaries based on “predicted semantic types” would reveal the impact of classification accuracy on the quality of the produced summaries.

Note that there is one caveat in this comparison: summaries of the two types may have different lengths for the same dialogue. This happens because our summarization method captures changes in conversation topics by identifying switches in semantic types of the dialogue turns. We found that summaries based on “true semantic types” contain 38% of the original dialogues, compared to the summaries based on “predicted semantic types” which contained 53% of the original dialogues. The discussion of our evaluation results in the next section takes this discrepancy into account.

IV. Evaluation Method

We first describe two alternative summarization strategies that we use for comparison with our system. We then introduce two evaluation frameworks for testing our summarizer.

A. The “Gold Standard” – Manual Dialogue Turn Extraction

We created a “gold standard” summary for evaluating our automatically extracted dialogue turns. Two physicians were given instructions to select dialogue turns that cover the most essential topics within each dialogue. For each dialogue, we limited the number of dialogue turns the human subjects could select ranging from a single turn up to 1/3 of the total turns in a dialogue (see instructions in Appendix B). We obtained summaries for 80 dialogues. Twenty summaries were summarized by two physicians while the remaining 60 were summarized by a single physician.

Measure of Agreement

We assess the degree of agreement between two humans by comparing selected dialogue turns for 20 dialogues that both physicians summarized. First, we calculated their percentage of agreement in manually selecting dialogue turns that best represent each dialogue. Second, we calculated an odds ratio to further illustrate agreement. Percentage of agreement is defined as the number of dialogue turns that both physicians included in the summary, divided by the total number of dialogue turns in the summary. The actual observed agreement is 81.8% between the two physicians. In addition, we computed the kappa to be 0.5, which is "substantial" agreement.[15] We also computed the odds ratio, which shows the relative increase in the odds of one subject making a given decision, given that the other subject made the same decision, is 10.8. It indicates that the odds of Subject 2 making a positive decision increases 10.8 times for cases where Subject 1 makes a positive decision, which is statistically significant ($p<0.0003$, log odds ratio).[23] These two measurements indicate that dialogue turn extraction can be reliably performed by humans in our domain.

B. Baseline Summary

The baseline summaries were produced by randomly selecting a third of the dialogue turns within each dialogue, independently of their semantic types. Random baselines are routinely used for comparison in the natural language domain.[24,25] In a task-based evaluation, random extraction methods commonly rival automatic methods since humans can compensate for poor summary quality by their background knowledge.

We therefore have the complete dialogue and four types of summaries for each dialogue: the “gold standard”, a randomly generated baseline, summaries based on “true semantic types”, and summaries based on “predicted semantic types”. Appendix C shows a sample of all four summaries with the original complete dialogue.

C. Intrinsic vs. Extrinsic Evaluation

Our evaluation is composed of two parts – intrinsic and extrinsic.[26,27] In the intrinsic part, we compare the automatically generated summaries to the “gold standard.” The key assumption is that automatically generated summaries that have higher overlap with the “gold standard” are better summaries. In the extrinsic part, we do a task-based evaluation and measure how useful the summaries are in preserving information important in the medical setting.

Intrinsic Evaluation

To measure the degree of overlap between an automatically computed summary and the “gold standard,” we use precision and recall. Precision penalizes **false positives** chosen by the system in question. It is similar to **positive predictive value** in the biomedical literature and is expressed as:

$$precision \equiv \frac{\text{\# Dialogue Turns } \textit{Correctly Chosen}}{\text{\# Dialogue Turns } \textit{Chosen}}$$

Recall penalizes **false negatives** chosen by the system. It is similar to **sensitivity** in the biomedical literature and is expressed as:

$$recall \equiv \frac{\text{\# Dialogue Turns } \textit{Correctly}\ \mathrm{Re}\,\textit{cognized}}{\text{\# Dialogue Turns } \textit{Should Have Been}\ \mathrm{Re}\,\textit{cognized}}$$

To have a single measure of a system’s performance, we also use the F-measure, defined as a weighted combination of precision and recall. It is expressed as:

$$F\text{–}measure \equiv \frac{2 * precision * recall}{precision + recall}$$

Using these measures, we compare automatically generated summaries using "predicted semantic types" and "true semantic types" with the "gold standard" and the random baseline. We use 2-tailed Fisher's Exact test to determine statistical significance.

Extrinsic (Task-Based) Evaluation

Our goal in this section is to determine whether the summaries are sufficient to provide caregivers with information that is important for patient care. We consulted with dialysis physicians and nurses to create a list of key questions based on topics that commonly arise between hemodialysis patients and caregivers.[2,28] (see Table 5) The questions address relevant issues in clinical assessment, technical support and overall delivery of quality patient care.

We distributed 200 dialogues, comprised of the complete version of 40 dialogues and four "summaries" of these same dialogues: (1) the manually created summaries; (2) the summaries based on randomly-extracted dialogue turns; (3) summaries based on the "true semantic types" of the dialogue turns; and (4) summaries based on the "predicted semantic types" of the dialogue turns. We had five licensed physicians (who did not participate in the selection of questions or in the manual summarization process) answer each of the six "yes/no" questions using each of 40 dialogues. It has been noted previously that when humans are asked to answer otherwise (e.g. NA or unknown), other factors come into play, such as a person's degree of decisiveness in committing to a response. Humans who are indecisive may tend to answer "NA" to a lot of questions and further bias the results. A similar experimental design has been adopted in other summarization systems.[29] The physicians received written instructions prior to performing their task (see Appendix D). Each physician only saw one version of every dialogue. Based on self-reporting, they completed the task of answering six questions for 40 dialogues in approximately one hour. Based on the complete dialogue, 30% of the answers to these questions are "yes" and 70% are "no." The characteristics of the complete data set are provided in Table 6 below. We compare the number of questions that physicians answered correctly using our summaries with answers based on the "gold standard" and the random baseline. Sign test was used to measure statistical significance.

V. Summarization Results

We report the results of the intrinsic and extrinsic evaluation.

A. Intrinsic Evaluation

The precision, recall and F-measure for the random baseline and the computer-generated summaries are shown in Table 7. The results indicate that machine-generated summaries outperform random summaries by a wide margin. The results of 2-tailed Fisher's Exact test comparing various summaries is shown in Table 8. As expected, recall was better for the summary that was generated using the predicted semantic types compared to true semantic types because it contained more dialogue turns. It is more important to note the effect on precision, which is less influenced by the length of the summaries. Precision was significantly better for both summaries compared to the random baseline and there was no difference between the precision of the two summaries. These results demonstrate the contribution of structural information to text summarization.

B. Extrinsic (Task-Based) Evaluation

We report the results of physicians' answers to each of our six questions when given various summaries for 40 dialogues. We assume that answers based on the complete dialogues are the correct ones. The numbers of correct responses are shown in Table 9 for each summary type. The summaries based on true semantic types outperformed all other summaries. Computer generated summaries based on predicted semantic types performed comparably, allowing physicians to correctly answer 81% of questions.

Statistical significance was measured using Sign test comparing summaries generated using our method to random summaries as shown in Table 10. Sign test has been used in the speech recognition domain to show systematic evidence of differences in a consistent direction, even if the magnitudes of the differences are small.[30] The automatically generated summaries outperform random summaries on five questions, with a tie for the sixth (see Table 9). Using one-tailed Sign test, this difference was significant. This test is applicable for our evaluation: we want to measure the degree of improvement our method has over the random baseline. Using two-tailed Sign test, there is no significant difference between computer-generated summaries and manually-generated or random summaries.

VI. Conclusion and Future Work

This work presents a first step towards automatic analysis of spoken medical dialogue. The backbone of our approach is an abstraction of a dialogue into a sequence of semantic categories. This abstraction uncovers structure in informal, verbose conversation between a caregiver and a patient, thereby facilitating automatic processing of dialogue content. Our method induces this structure based on a range of linguistic and contextual features that are integrated in a supervised machine-learning framework. We demonstrate the utility of this structural abstraction by incorporating it into an automatic dialogue summarizer. Our evaluation results indicate that automatically generated summaries exhibit high resemblance to summaries written by humans. Our task-based evaluation shows that physicians can reasonably answer questions related to patient care by looking at the summaries alone, without reading a full transcript of a dialogue. We believe that further refinement of the presented summarizer would ultimately spare the physician from the need to wade through irrelevant material ample in dialogue transcripts. Automatic segmentation of dialogues by topics and the use of more expressive statistical models to capture the sequential structure of dialogue will likely enhance the summarizer's performance.

In the future, we plan to extend this work in three main directions. First, we will apply our method to automatically recognized conversations. Clearly, automatic speech recognition will introduce mistakes in a transcript. At the same time, we will have access to a wealth of acoustic features that provide additional cues about dialogue content. For instance, a pause may be a strong indicator of topic switch. Therefore, we will explore the use of acoustic features to compensate for recognition errors in the transcript. Second, we will refine our annotation scheme to include more semantic categories. This would support a deeper analysis of medical dialogue. To achieve this goal, we will experiment with more expressive statistical models able to capture the sequential structure of medical dialogue. Possible modeling methods include hidden Markov models and conditional random fields.[31,32] Finally, we will explore query-based summarization as opposed to generic summarization[33]. In our current implementation, the summaries are not tailored to specific information needs of a care provider. By knowing what information is of interest to different categories of care providers, we can personalize the summaries towards their needs.

Acknowledgements

The authors acknowledge the support of the National Science Foundation (Barzilay; Career grant IIS-0448168 and grant IIS-0415865). We thank Dr. Robert Lockridge and the nurses at the Lynchburg Nephrology nightly home hemodialysis program, Mary Pipken, Viola Craft and Maureen Spencer for recording the conversations. We thank Percy Liang for letting us use his word clustering software, Peter Szolovits and Eduardo Lacson Jr. for reviewing the manuscript and all the physicians who participated in the annotation and evaluation of the system. Any opinions, findings and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A: Request for Annotation

We provide here the instructions and examples for annotating dialogue turns within our dialogues.

A.1. Instructions

Dear Doctor,

I would like to request your participation in annotating a transcription of a telephone dialogue between dialysis nurses and patients. This annotation will be used to help identify the most frequent reasons for calls to a dialysis unit by actual patients. It will be used in conjunction with other methods in helping identify the topics that are pertinent to patients who undergo home hemodialysis.

The dialog will be segmented by utterances or each person's turn in the actual dialogue. Each turn will be labeled as belonging to one of several categories:

1. Clinical
2. Technical
3. Greetings and acknowledgements
4. Miscellaneous

As implied by the category names, a **clinical** utterance is anything that pertains to a clinical topic, such as the patient's health, medications, laboratory tests (results) or any concerns or issues the patient or nurse has regarding the patient's health. Examples include:

1. You see, his pressure's dropping during his treatments.

2. Do you want me to do blood test?

A **technical** utterance relates to machine problems, troubleshooting, electrical, plumbing, or any other issues that require technical support. This also includes problems with performing a procedure or laboratory test because of lack of or defective materials, as well as a request for necessary supplies. Procedures for doing a laboratory test will also be classified as technical. Examples include:

1. The machine is stuck

2. That's where you spike it, the second port is the one where you draw from.

Greetings include "**hellos**" and "**goodbyes**" that are typically located at the beginning and end of a call.

Acknowledgements and confirmatory responses to questions include "**aha**", "**ok**", "**alright**", "**yes**", etc. Examples of this category include:

1. Hello, is S__ there?

2. Thanks for calling.

Any other utterances can be classified as **miscellaneous**. These include (but are not exclusive to) scheduling (a clinical or technical meeting or appointment), personal conversations, etc. Examples include:

1. I'll call you back

2. I'm just helping out till they get back from vacation

An utterance should be taken within the context of the conversation. (e.g. "**I'm taking two**" should be categorized as clinical if the conversation is regarding how many tablets a patient is taking.) However, "**ok**", "**yes**" and other acknowledgements should be categorized as confirmations.

Please indicate the categorizations by marking the clinical utterances with "C", the technical utterances with "T", acknowledgements/greetings with "A" and miscellaneous utterances with "M". A sample annotation is given below.

An utterance can be categorized into more than one topic. If any utterance appears to belong to more than one topic, please indicate both categories. For example,

1. "**You know the meter on the machine, and I couldn't get it to come out so I called technical support. He said someone will call him but nobody called me.**" This can be technical because it concerns the machine or miscellaneous because it refers to someone who needs to call. You can indicate "T" or "M" in this case.
2. "**Ok, how many hours did you run M_**". This can be clinical because knowing how long the patient dialyzed impacts their health. It can also be technical if taken in context with the machine not working anymore after this run. You can indicate "C" or "T" in this case.

This participation is voluntary and any specific data you provide will not be published or made available without your consent. Thank you.

A.2. Sample of Annotated Dialogue

C Just changed it this morning, he said it’s not sore. It’s still got the dressing on it, didn’t take it out last night in case it drains again.

C Have you looked at it this morning?

C It hasn’t drained overnight. Just a little bit. It’s not clear, it’s pussy looking

A Ok

C It’s not red like it was last night.

C ok, let me Dr. M_ is on call for the weekend, let me give her a call. See if he wants to put him on any antibiotics. You, know, preventatively

A ok

M and I’ll call you back

T, M You know the meter on the machine, and I couldn’t get it to come out so I called technical support. He said someone will call him but nobody called me

C, T ok, how many hours did you run M_

C, T 3 and a half

C, T You ran 3 and a half?

A aha.

M ok, well nobody will be coming out here today anyway to do anything about your machine

A aha

M At least, till tomorrow morning. And I will go ahead and call them to see if we can get somebody to come out there tomorrow to do something

M It’s the same thing.

M Oh you’re kidding

Appendix B: Instructions given to physicians for manually selecting dialogue turns

B.1. Instructions

Dear Doctor,

1. Please select dialogue turns from each phone call, which are most representative of the entire dialogue and would give the reader an idea about the topics within the conversation. In particular, please pick dialogue turns that are important to the patient's health and dialysis management. Information about their relatives, their homes, etc. is not relevant unless these impact the delivery of their care.
2. A dialogue turn starts with N: (for a nurse's turn) or P: (for a patient's turn).
3. You are allowed to pick at least one dialogue turn, up to a specified number of turns that will best summarize the conversation, at your discretion.
4. Please highlight your choices with the highlighter provided.
5. See example below.

Thank you.

B.2. Example

Select up to 3 turns

N: ok

P: I was making cabbage rolls and a little bit of rice. And I have to cook the rice and put it in there. And it's the regular long grain rice. And I thought it would cook, you know, in the rolls.

N: Right

P: But it appears not to get done so the first half of the cabbage rolls I ate was crunchy rice.

N: Oh, ok.

P: I just wanted to ask if there's anything I should watch out for because I know raw rice is not a good thing for you. (laughs)

N: I'll ask Dr. LAWSON ok, coz I'm not sure to be honest with you, but I'll ask Dr. LAWSON. I'll call you back and let you know, ok?

P: Ok. I'm just concerned because people stop throwing rice at weddings because birds would eat it. And they get stuck in their stomachs. Now they probably don't have enough enzymes, but we can probably break down rice and stuff but I just called to make sure.

N: Ok, well I'll ask her and I'll call you back and let you know, ok?

P: ok, Thanks.

N: Bye-bye.

Appendix C: Complete and Summarized Dialogues

C.1 Complete Dialogue With each Turn Labeled with Corresponding Summaries that contain this Turn (G: gold standard, R: random, T: true semantic type, P: predicted semantic type)

R T P	P: It's the machine, I couldn't turn it on
P	N: What's the matter?
G P	P: The pressure, arterial pressure, I mean the venous pressure, I couldn't even turn the pump on
G R P	N: Did you have the transducer hooked up? Your monitor is on?
G R T P	P: Yes ma'am, my blood won't flush, every time I try to turn the pump on, its either I got a negative pressure, arterial has a pressure now, and both of my catheters, I have an arterial pressure of 220 and a venous pressure of 180. I don't even have my pump open.
R	N: You don't have any pumps open where? On your catheter?
	P: I have pressures a little bit there.
R	N: I can hear the warning. Does it flush ok?
	P: Yeah
	N: I will try switching the ports. Start the pump and clamp off your lines and try switching the ports. And then turn it on and see what happens
G	N: Can you come off and put your blood in recirculation? I'll go ahead and call technical support and see if they have any suggestions. I can't think of anything else that can be causing it.
T P	N: How are you feeling?

R	P: I feel fine.
G R	N: You feel better? Your target weight's ok?
G T P	P: My blood pressure was 147/79, when I sit it drops to 139/73. My pulse is good, 80 and 84.
R	N: And how's your weight now
	P: 129.2
G	N: Your blood pressure medicine, I'll have you finish that.
	P: I finished taking that on Friday
G	N: Oh, so you finished taking that Friday, and the diarrhea and nausea, all that stopped.
	P: Yuh
T P	N: Ok, that's good. Go ahead and call technical support and then just call me back and let me know what they say, ok?

C.2 The four summaries (P: patient, N: nurse)

Gold Standard	P: The pressure, arterial pressure, I mean the venous pressure, I couldn't even turn the pump on
	N: Did you have the transducer hooked up? Your monitor is on?
	P: Yes ma'am, my blood won't flush, every time I try to turn the pump on, its either I got a negative pressure, arterial has a pressure now, and both of my catheters, I have an arterial pressure of 220 and a venous pressure of 180. I don't even have my pump open.
	N: Can you come off and put your blood in recirculation? I'll go ahead and call technical support and see if they

	have any suggestions. I can't think of anything else that can be causing it. N: You feel better? Your target weight's ok? P: My blood pressure was 147/79, when I sit it drops to 139/73. My pulse is good, 80 and 84. N: Your blood pressure medicine, I'll have you finish that. N: Oh, so you finished taking that Friday, and the diarrhea and nausea, all that stopped.
Random	P: It's the machine, I couldn't turn it on N: Did you have the transducer hooked up? Your monitor is on? P: Yes ma'am, my blood won't flush, every time I try to turn the pump on, its either I got a negative pressure, arterial has a pressure now, and both of my catheters, I have an arterial pressure of 220 and a venous pressure of 180. I don't even have my pump open. N: You don't have any pumps open where? On your catheter? N: I can hear the warning. Does it flush ok? P: I feel fine. N: You feel better? Your target weight's ok? N: And how's your weight now
"True semantic type"- based	P: It's the machine I couldn't turn it on P: Yes ma'am, my blood won't flush, every time I try to turn the pump on, its either I got a negative pressure, arterial has a pressure now, and both of my catheters, I have an arterial pressure of 220 and a venous pressure of 180. I don't even have my pump open.

	N: How are you feeling **P: My blood pressure was 147/79, when I sit it drops to 139/73. My pulse is good, 80 and 84.** **N: Ok, that’s good. Go ahead and call technical support and then just call me back and let me know what they say, ok?**
“Predicted semantic type”- based	**P: It’s the machine, I couldn’t turn it on** **N: What’s the matter?** **P: The pressure, arterial pressure, I mean the venous pressure, I couldn’t even turn the pump on** **N: Did you have the transducer hooked up? Your monitor is on?** **P: Yes ma’am, my blood won’t flush, every time I try to turn the pump on, its either I got a negative pressure, arterial has a pressure now, and both of my catheters, I have an arterial pressure of 220 and a venous pressure of 180. I don’t even have my pump open.** **N: How are you feeling?** **P: My blood pressure was 147/79, when I sit it drops to 139/73. My pulse is good, 80 and 84.** **N: ok, that’s good. Go ahead and call technical support and then just call me back and let me know what they say, ok?**

Appendix D: Instructions given to evaluators

D.1. Instructions

Dear Doctor,

Below are some dialogues between dialysis nurses and patients. After reading each dialogue, please answer the 6 (yes/no) questions that follow. Some dialogues are incomplete, so just answer the best you can. Thanks a lot for doing this amidst your busy schedule.

Questions:

1. Did a clinical problem require urgent intervention?
2. Did the patient mention either his vital signs (blood pressure, pulse rate, temperature), his weight, any symptoms, or his medications?
3. Was there a problem with the machine that required technical support?
4. Did the call require a follow-up (i.e. need to consult with another nurse, a physician, a technician or a supplier and/or require further laboratory investigation outside of the current call)?
5. Did the patient need to make, verify, cancel or reschedule an appointment?
6. Did the patient need to be dialyzed in-center?

VIII. References

[1] Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. BMJ. 1975; 2(5969): 486-9.

[2] Lockridge RS Jr. Daily dialysis and long-term outcomes-the Lynchburg Nephrology NHHD experience. Nephrol News Issues. 1999; 13(12): 16, 19, 23-6.

[3] Walker MA, Langkilde-Geary I, Hastie HW, Wright J, Gorin A. Automatically training a problematic dialogue predictor for a spoken dialogue system. J of Artificial Intelligence Research. 2002; 16:293-319.

[4] Document Understanding Workshop. HLT/NAACL Annual Meeting. Boston, MA. May, 2004. In: http://duc.nist.gov/. Accessed on: June 16, 2005.

[5] Kupiec J, Pedersen J, Chen F. A trainable document summarizer. Research and Development in Information Retrieval. 1995: 68-73. In: http://citeseer.csail.mit.edu/kupiec95trainable.html.

[6] Edmundson HP. New methods in automatic extracting. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 23-42.

[7] McKeown K, Radeev DR. Generating summaries of multiple news articles. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 381-390.

[8] Merlino A, Maybury M. An empirical study of the optimal presentation of multimedia summaries of broadcast news. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 391-402.

[9] Marcu D. Discourse trees are good indicators of importance in text. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 123-136.

[10] Teufel S, Moens M. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 155-176.

[11] McKeown K, Hirschberg J, Galley M, Maskey S. From text to speech summarization. ICASSP. 2005. Philadelphia, PA. In: http://www1.cs.columbia.edu/~galley/papers/from_txt_to_speech.pdf. Last accessed: June 20, 2005.

[12] McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc. 1993; 81(2): 184-94.

[13] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo. 2004: 565-72.

[14] Hsieh Y, Hardardottir GA, Brennan PF. Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses. Medinfo. 2004: 511-5.

[15] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-174.

[16] Schapire R, Singer Y. Boostexter: A boosting-based system for text categorization. Machine Learning. 2000; 39(2/3):135-168.

[17] Lacson R, Barzilay R. Automatic Processing of Spoken Dialogue in the Home Hemodialysis Domain. AMIA 2005 Fall Conference proceedings. (accepted for publication)

[18] Brown PF, DeSouza PV, Mercer R, Della Pietra VJ, Lai JC. Class-based n-gram models of natural language. Computational Linguistics. 1992; 18: 467-479.

[19] Miller S, Guinness J, Zamanian A. Name tagging with word clusters and discriminative training. HLT-NAACL. 2004: 337-342.

[20] Lacson R, Long W. How Accurate are Experts in Spotting the Right Words in a Dialogue? In: http://people.csail.mit.edu/rlacson/dialogue_poster.doc. Last accessed: June 20, 2005.

[21] Lacson R, Lacson E, Szolovits P. Discourse structure of medical dialogue. Proceedings of MEDINFO. 2004: 1703.

[22] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In Proceedings of the 22nd ACM SIGIR. 1999: 121-128.

[23] Bland JM, Altman DG. The Odds Ratio. BMJ. 2000; 320:1468.

[24] Zechner K, Waibel A. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. The 18th International Conference on Computational Linguistics. 2000; 1: C00-2140.

[25] Allan J, Gupta R, Khandelwal V. Temporal summaries of news topics. Proceedings of SIGIR. 2001: 10-18.

[26] Jones KS, GAlliers JR. Evaluating natural language processing systems: an analysis and review. New York, Springer (eds). 1996.

[27] Jing H, Barzilay R, McKeown K, Elhadad M. Summarization evaluation methods: experiments and analysis. AAAI Intelligent Text Summarization Workshop (Stanford, CA); Mar. 1998: 60-68.

[28] Lehoux P. Patients' perspectives on high-tech home care: a qualitative inquiry into the user-friendliness of four technologies. BMC Health Serv Res. 2004 Oct 5; 4(1): 28.

[29]Teufel S. Task-based evaluation of summary quality: Describing relationships between scientific papers. In Proceedings of NAACL-01 Workshop on Automatic Text Summarization.2001.

[30] Pallett D, Fiscus J, Garofolo J. Resource Management Corpus: September 1992 Test Set Benchmark Test Results, Proceedings of ARPA Microelectronics Technology Office Continuous Speech Recognition Workshop (Stanford, CA); September 21-22, 1992.

[31] Conroy J, O'Leary D. Text summarization via hidden Markov models. Proc 24th annual international ACM SIGIR conference on research and development in information retrieval. 2001: 406-407.

[32] Lafferty J, Pereira F, McCallum A. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning. 2001: 282-289.

[33] Sakai T, Sparck-Jones K. Generic summaries for indexing in information retrieval. Proc 24th annual international ACM SIGIR conference on research and development in information retrieval. 2001: 190-198.

Table 1: Semantic Type Distribution of Dialogue Turns in Training and Testing Data Sets

Category	Training (n=1281)	Testing (n=293)
Clinical	33.4%	20.8%
Technical	14.6%	18.1%
Backchannel	27.2%	34.5%
Miscellaneous	24.7%	26.6%

Table 2: Examples of dialogue for each semantic type

Clinical: 1. Ok, how’s the Vioxx helping your shoulder? 2. You see, his pressure is dropping during his treatments.
Technical: 1. Umm, I’m out of kidneys. 2. That's where you spike it; the second port is the one where you draw from.
Miscellaneous: 1. Martha wants me to remind you of your appointment today at 8:30. 2. I'm just helping out 'til they get back from vacation.
Backchannel: 1. Hello. How are you doing? 2. Yeah.

Table 3: Accuracy of the models based on various feature combinations

Models	**Accuracy**
Dialogue turn	69%
Dialogue turn with length	70%
Dialogue turn with previous turn	68%
Basic Model (Dialogue turn with length and previous turn)	**70%**
Knowledge-Augmented Model	**73%**

Table 4: Examples of predictive features

Category	Current Dialogue Turn	Previous Dialogue Turn
Clinical	**weight, blood, low, feel, pulse**	**weight, take integer, you**
Technical	**filter, box, leaking**	**machine, a little**
Backchannel	**thanks, ok, and, umm**	**hi, make, sure, lab**
Miscellaneous	**appointment, hold, phone**	**can, o clock, what, time**

Table 5: Questions used in task based evaluation

1. Did a clinical problem require urgent intervention?
2. Did the patient mention either his vital signs (blood pressure, pulse rate, temperature), his weight, any symptoms, or his medications?
3. Was there a problem with the machine that required technical support?
4. Did the call require a follow-up (i.e. need to consult with another nurse, a physician, a technician or a supplier and/or require further laboratory investigation outside of the current call)?
5. Did the patient need to make, verify, cancel or reschedule an appointment?
6. Did the patient need to be dialyzed in-center?

Table 6: Answer distribution across six questions based on full transcripts of dialogues

Number of dialogues	40
Average number of dialogue turns per dialogue	13
Number of “yes” answers to question 1	12 (0.30)
Number of “yes” answers to question 2	33 (0.41)
Number of “yes” answers to question 3	20 (0.25)
Number of “yes” answers to question 4	38 (0.48)
Number of “yes” answers to question 5	26 (0.32)
Number of “yes” answers to question 6	8 (0.10)
Total number of “yes” answers	143 (0.30)

Table 7: Intrinsic Evaluation Results with Precision, Recall and F-measure for 40 Dialogues

	Random	Computer-generated using true semantic type	Computer-generated using predicted semantic type
Precision	62/183 (33.88%)	107/199 (53.77%)	139/277 (50.18%)
Recall	62/177 (35.03%)	107/177 (60.45%)	139/177 (78.53%)
F-measure	34.45	56.91	61.23
Number of dialogue turns	183/516 (35.47%)	199/516 (38.57%)	277/516 (53.68%)

Table 8: Fisher's Exact test comparing the precision and recall of pairs of summary types ($p<0.05$ is statistically significant)

	Computer-generated using true semantic type vs. Random	Computer-generated using predicted semantic type vs. Random	Computer-generated using predicted semantic type vs. true semantic type
Precision	$1.38x10^{-4}$	$7.94x10^{-4}$	0.4580
Recall	$2.53x10^{-6}$	$1.03x10^{-16}$	$3.23x10^{-4}$

Table 9: Number of correct responses for each summary type

	Q1	Q2	Q3	Q4	Q5	Q6	Total
Random	27 (67.5%)	28 (70.0%)	33 (82.5%)	26 (65.0%)	29 (72.5%)	38 (95.0%)	181 (75.4%)
Manual	31 (77.5%)	34 (85.0%)	35 (87.5%)	28 (70.0%)	24 (60.0%)	38 (95.0%)	190 (79.2%)
Computer-generated using true-label	31 (77.5%)	34 (85.0%)	37 (92.5%)	27 (67.5%)	33 (82.5%)	38 (95.0%)	200 (83.3%)
Computer-generated using predicted-label	29 (72.5%)	32 (80.0%)	38 (95.0%)	28 (70.0%)	29 (72.5%)	39 (97.5%)	195 (81.2%)

Table 10: Comparison of the accuracy of the summaries using Sign Test (p<0.05 is statistically significant, NS=not significant)

	Sign Test (One-tailed, n=5)	Sign Test (Two-tailed, n=5)
Computer-generated using true semantic type vs. Random	p=0.031	p=0.062
Computer-generated using predicted semantic type vs. Random	p=0.031	p=0.062
Computer-generated using true semantic type vs. Manual	NS	NS
Computer-generated using predicted semantic type vs. Manual	NS	NS
Computer-generated using predicted semantic type vs. true semantic type	NS	NS