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

by

Ronilda Covar Lacson

Master of Science in Medical Informatics, Massachusetts Institute of Technology (2000)

Doctor of Medicine, University of the Philippines (1992)

B.S. Basic Medical Sciences, University of the Philippines (1989)

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September, 2005

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAR 2 8 2006
LIBRARIES

Copyright © 2005 Massachusetts Institute of Technology. All rights reserved.

Author _____

Department of Electrical Engineering and Computer Science
August 26, 2005

Certified by _____

William J. Long, PhD
Thesis Supervisor

Accepted by _____

Prof. Arthur C. Smith
Chairman, Committee on Graduate Students

ARCHIVES

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

by

Ronilda Covar Lacson

Submitted to the Department of Electrical Engineering and Computer Science on

August 26, 2005 in Partial Fulfillment of the

Requirements for the Degree of Doctor of Philosophy

Massachusetts Institute of Technology

Abstract

Spoken medical dialogue is a valuable source of information, and it forms a foundation for diagnosis, prevention and therapeutic management. However, understanding even a perfect transcript of spoken dialogue is challenging for humans because of the lack of structure and the verbosity of dialogues. 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. Our model has a classification accuracy of 73%, compared to 33% achieved by a majority baseline ($p<0.01$). 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 and significantly outperform random selections ($p<0.0001$) in precision and recall. 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 is a significant result because it spares the physician from the need to wade through irrelevant material ample in dialogue transcripts. This work demonstrates the feasibility of automatically structuring and summarizing spoken medical dialogue.

Thesis Supervisor: William J. Long, PhD
Title: Principal Research Scientist

Contents

5. Structure Induction

5.1 Semantic Taxonomy

5.2 Data Annotation

5.2.1 Agreement

5.3 Semantic-Type Classification

5.3.1 Basic Model

5.3.2 Data Augmentation with Background Knowledge

5.3.2.1 UMLS Semantic Types

5.3.2.2 Automatically Constructed Word Clusters

5.3.3 Results of Semantic Type Classification

5.3.4 Using a Sequential Model for Semantic Type Classification

5.3.4.1 Semantic Type Transition

5.3.4.2 Utilizing the Previous Turn's Label

5.3.4.3 Conditional Random Fields

6. Summarization

6.1 Summarization Method

6.2 Predicted Semantic Type vs. True Semantic Type

7. Evaluation

7.1 The "Gold Standard" – Manual Dialogue Turn Extraction

7.2 Baseline Summary

7.3 Intrinsic vs. Extrinsic Evaluation Techniques

7.4 Evaluation Results

8. Conclusions and Future Work

8.1 Future Work

9. Appendices

Appendix A: Partial Results of the Agglomerative Clustering Algorithm

Appendix B: Request for Annotation

B1. Instructions

B2. Sample of Annotated Dialogue

Appendix C: Samples of word clusters for various cluster sizes

C1. Results using 1000 clusters

C2. Results using 1500 clusters

C3. Results using 2000 clusters

C4. Results using 5000 clusters

10. References

List of Tables

List of Figures

Acknowledgements

I thank all members of my thesis committee. Fortunately, I was able to work closely with each of them at various stages of my research at MIT. My supervisor, Bill Long, encouraged me whenever I hit "dead ends" in my research while tackling a new problem or trying a new method. He challenged me to focus on my goal, to keep the basic foundations of my research stable, and to emphasize the practical and challenging aspects of my research. I am glad Regina Barzilay came to MIT as a faculty member last year. She mentored me on natural language processing, she challenged me to keep trying new scientific methods for my research, and she brought a new perspective to the evolution and development of my scientific writing style. Pete Szolovits has been a wonderful mentor to me from the first moment I entered MIT at the Division of Health Sciences and Technology (HST), unwavering until today. He supported and encouraged my interests in various areas of research in computer science and medical informatics. He constantly presents new and exciting opportunities and directions for my research.

I am also grateful to my friends and colleagues at the Medical Decision-Making Group (MEDG) at MIT. It was a pleasure coming to work everyday in a collegial work environment. I thank Fern DeOliveira for all her assistance. I enjoyed my discussions with Stanley Trepetin, my old officemate at the old NE-43 building. My current officemates, Mark Finlayson, Keith Bonawitz and Paul Keel, at the new Stata Center have likewise provided stimulating discussions. I also thank Percy Liang from the Spoken Language Systems Group who graciously shared his word clustering software when the need for it arose.

I thank Dr. Robert Lockridge and the nurses at the Lynchburg Home Hemodialysis Program in Lynchburg, VA – Mary Pipken, Viola Craft and Maureen Spencer. They were responsible for recruiting all the patients who participated in my research. They performed the necessary consenting procedures, they participated in my training seminar and recorded all conversations for this study. They really wanted to make a difference in how home dialysis patients are cared for and they are very conscientious in their work.

I thank all the physicians who participated in my study, all of whom are excellent clinicians and researchers – Tom Lasko (Medical Informatics, MA), Joey Boiser (Neurology, AL), Jojo Almendral (Cardiology, NY), Andrea De Leon (Pediatrics, NY), Wendy Castrence (Internal Medicine, AL), Mike Nolledo (Pulmonary and Critical Care, NJ), Pauline Lerma (Oncology, NJ), Chat Baron (Dermatology, OH), Winston Mina (Geriatrics, MO), Martin Lansang (Pulmonary and Critical Care, PA), Janus Ong (Hepatology and Gastroenterology, DC), and Eduardo Lacson (Nephrology, MA). Amidst their hectic clinical schedules on top of their research responsibilities, they performed my requested tasks promptly and professionally.

I would like to thank my previous mentors at the Decision Systems Group (DSG) at the Brigham and Women's Hospital where I did my Medical Informatics fellowship. In particular, I wisk to thank Bob Greenes, Lucila Ohno-Machado and Aziz Boxwala. Through them I was able to develop an early perspective of Medical Informatics.

I deeply thank my family for their love and support. My brother and sisters – Rainier Covar, Ronina Covar and Ronora Covar-Pena – two physicians and a computer scientist, all of whom were very supportive while I attempted to bridge Medicine and Computer Science. I thank my parents, Roger and Nina Covar, for coming all the way from the Philippines to lend support and assist our family so I can devote more energy during the final stages of this work. Both retired professors and scientists themselves, they provided me with the discipline to finish whatever educational goal I aim for. I would never have finished my medical training without them and they continue to be supportive even now.

Most importantly, I thank my husband, Eduardo Lacson Jr., for supporting me while I contemplated leaving clinical medicine, for encouraging me to embark on this new career, for trusting that I can reach what I aspire for if I try hard enough. I thank God for him and for two additional sources of inspiration, Roger and Edward, my children who were born while I was in the midst of this research endeavor. I dedicate this thesis to God and to my family, a testament of my work, a humble beginning designed to inspire future work in this evolving field, a piece of work with a dedication to last beyond my lifetime.

For my three boys

Jay-r, Roger and Edward

With Love

1. Introduction

1.1 Motivation

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 two minutes of 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 "`The same, he's on the same one`"), extraneous filler words (such as "`ok`" or "`oh`") and non-lexical filled pauses (such as "`Mmm`"). This exposition also highlights the striking lack of structure in the transcript: from reporting a patient's blood pressure readings, the topic switches to how long the patient underwent dialysis that day, and then switches again to a discussion about the patient's medications without any visible delineation customary in written text. Therefore, a critical problem for processing dialogue transcripts is to provide information about their internal structure.

P: His blood pressure's still up.

N: How high is it?

P: When we went on, it was 182/88, that was sitting. And 168/83 standing

N: Ok

P: After I hook him up and everything, and I put him on, it was 167/85

N: Ok

P: When he came off tonight, I took the same reading. The first one I took, I took minutes before he came off. And sitting, it was 216/96

N: Ok

P: When I took it again, it was 224/100.

N: Ok

P: So when he finally came off, I took the reading, the first one was 214/106

N: Mmm

P: And standing was 203/92. Last night, when he came off after the first one was 213/92 and the one sitting was 205/98, standing was 191/91.

N: Ok

P: I can give you last night's. It was high too. When he went on, it was pretty good, when he came off, it was 211/96, the first one. And the second one was 222/100, and 197/85.

N: Now what time did he come off today?

P: He went 6 hours, 5 o'clock to 11.

N: How does he feel?

P: He feels fine.

N: What medicine is he on now? What blood pressure medicine?

P: The same, he's on the same one, Corgard

N: ok, I have to talk to Dr. Lindsay tomorrow and see if he wants to put him on something else.

P His weight, the other day was 63. But he came off at 61.5

N: What's his target weight?

P: 61.5

N: ok

N: When is he scheduled to see Dr. Lindsay

P: He's scheduled to see Dr. Lindsay this week, he comes to see Dr. Lindsay on the 26th

N: Oh, not until the 26th

P: He comes in on Monday for bloodwork

N: Ok, I'll talk to Dr. Lindsay tomorrow about his pressure. And he might need to adjust his medication

P: That's about it

Figure 1: Transcribed segment of a phone dialogue

1.2 Thesis Statement

In this thesis, we show that we can automatically acquire the structure of spoken medical dialogue using techniques that have been developed for natural language processing for written text. We also demonstrate that dialogue summarization is feasible using the techniques we describe. Lastly, we design a framework to evaluate a medical dialogue summarizer that can assess the usefulness of summaries in the medical setting and effectively discriminate against baseline naïve summaries.

1.3Thesis Contributions

This thesis 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.

Our emphasis on spoken discourse sets us apart from the efforts to interpret written medical text.[3,4,5] This thesis has three main contributions:

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 derived from two sources: (1) Unified Medical Language System or UMLS,[6] a manually crafted, large-scale domain ontology, and (2) clusters of semantically related words automatically computed from a

large text corpus. Our best model has a classification accuracy of 73%, compared to 33% achieved by the majority baseline.

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. We construct such a summary by extracting dialogue turns that are representative of key topics discussed by a caregiver and a patient. The extraction algorithm relies on a variety of features to select summary sentences, including utterance length, position within the dialogue and its semantic type.

Evaluation We describe a framework for evaluating a summarizer of medical dialogues. Our first evaluation method follows an intrinsic methodology, commonly used in the text summarization community.[7] 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. Based on the evaluation results, we conclude that automatically generated summaries capture essential pieces of information about patient-caregiver interaction that can be utilized for improving quality of care.

1.4 Thesis Outline

The thesis is structured as follows. In Chapter 2, we present some related work on text summarization and medical discourse. In Chapter 3, we present our data collection technique. In Chapter 4, we describe how we performed structure induction using unsupervised clustering. In Chapter 5, we describe a semantic taxonomy and show that it can be reliably annotated by doctors. We present the basic design of our dialogue act classifier:[8,9] it predicts the semantic type of an utterance based on a shallow meaning representation encoded as simple lexical and contextual features. We then show how to enhance our machine learning algorithm with background knowledge. The experimental evaluation, described in Section 5.3, confirms that adding semantic knowledge brings some improvement to dialogue turn classification. In Chapter 6, we present the methods that we used in creating summaries from a transcribed medical dialogue based on the actual semantic category of each dialogue turn. In Chapter 7, we discuss the procedures for obtaining "gold standard" human-created summaries and baseline summaries and explain our evaluation methodology. We report the results of the intrinsic and the task-based evaluation. Finally, we present our conclusion and future work in Chapter 8.

2. Related Work

In this Chapter, we describe related work in structure induction for discourse and we focus on medical discourse in particular. We will then discuss approaches to written text summarization followed by a few related works on spoken dialogue summarization. Finally, we will discuss current approaches to summarization evaluation. We will discuss the knowledge sources that are pertinent to our thesis in Chapter 5, where we then describe augmentation of our model with background knowledge.

2.1 Discourse Structure and Dialogue Modeling

Discourse structure builds on first identifying appropriate discourse units for a given domain and task. Based on a discourse unit such as a dialogue act, various models are developed to identify the specific dialogue's structure. Dialogue act modeling is a growing area of research in natural language processing. [10,11,12] Speech act theory [13,14] has been used as a basis for much of the current dialogue act identification schemes. We will discuss the methods that have been used for dialogue modeling in the succeeding section.

2.1.1 Discourse units

Discourse units have been described in various ways and referred to by various names. A comprehensive discussion of discourse units and segments is abundant in the Computational Linguistics literature.[15,16,17,18,19] The most commonly used discourse unit is the dialogue act. A dialogue act represents the meaning of an utterance at the level of illocutionary force.[11] Based on speech act theory that dates back to 1962,[10] Austin offered an analysis of the concept of speech acts and distinguishes between three aspects: Locutionary act, Illocutionary act and Perlocutionary act. The locutionary act deals with saying something that makes sense in a certain language, including the act of producing noises and conforming to a vocabulary and grammar. Illocutionary act includes the illocutionary force that specifies the type of action performed while saying something (e.g. question, answer, etc.), and the propositional content that specifies the action in more detail. This aspect mirrors the speaker's intention behind an utterance. Perlocutionary act reflects the effects provoked by an utterance in context (e.g. surprise, persuasion). It also mirrors the listener's perception of the intentions of an utterance. Dialogue acts traditionally focus on illocutionary acts, and specifically on the illocutionary force. Speech acts, which by definition reflect underlying intentions or plans by the speaker, have long been incorporated into plan-based approaches in dialogue.[20] In these approaches, a dialogue is defined as a plan, while the speech acts are the components of a structured account of the dialogue.[21]

The intentions of the dialogue acts become the basis for dialogue act tag sets or the classification schemes that have been developed in Natural Language approaches to dialogues.[12,13,22] There are close similarities between dialogue acts in dialogue act coding schemes, such as the one described by Carletta,[19] and intention-based discourse segments, as described in Discourse Structure Theory by Grosz and Sidner.[16] Both approaches rely on dialogue units that are based on intentions underlying linguistic behavior.

Various specific textual units have been used to represent a dialogue act. These representations are primarily based on the ease of extracting text from various data sources. Some examples of commonly used representations are as follows.

1. Sentences or clauses – Sentences are obvious units of text, especially when one looks at written textual data (e.g. "My arm hurts."). Traditionally, each sentence is regarded as having a subject, an object and a verb.

2. Dialogue Segments – Dialogue segments are sequences of words, utterances, clauses or sentences that all pertain to one "topic". The notion of a topic was defined by Brown and Yule as a way of describing the unifying principle which makes one stretch of discourse about "something" and the next stretch about "something else."[23] We use dialogue segments in our preliminary analysis of dialogue structure (see section 4.2).

3. Dialogue Turns – A dialogue turn has been used extensively in the field of Conversation Analysis and is defined as "a time during which a single participant speaks, within a typical, orderly arrangement in which participants speak with minimal overlap and gap between them."[24] A dialogue turn may thus contain one to many sentences. The dialogue turns, however, more distinctly reflects the interaction between participants in a conversation. Obviously, dialogue turns are only defined for spoken dialogue.

   For various spoken language systems, dialogue turns have been utilized for question answering tasks, as well as summarization tasks dealing with human-human spoken dialogues.[25,26] Turn-taking marks an explicit boundary in a conversation when a speaker recognizes that the other speaker is done and he then initiates his own turn. We therefore use dialogue turns as our basic dialogue units.

2.1.2 Dialogue Modeling

Dialogue act modeling is a central goal in dialogue analysis. This entails developing an annotation scheme (corresponding to the dialogue act) and labeling each utterance in a dialogue with a dialogue act. It forms the basis for computing a dialogue unit's meaning, which is important for understanding a dialogue. Several researches have focused on structuring dialogue using techniques ranging from Transformation-Based Learning to Hidden Markov modeling. We will discuss several approaches and the relevant features that they identified in this section. The first three methods we describe focus on local features of a dialogue unit or the preceding unit (e.g. lexical features, duration). The next two methods incorporate the sequential nature of dialogues and models sequences of dialogue acts using machine learning techniques.

Maximum Entropy Classification

Dialogue act segmentation and classification was performed using audio recordings from the ICSI meeting corpus.[22] Automatic segmentation techniques were used based on acoustic features from the data. Subsequently, the dialogue acts were classified into five broad domain-independent categories: statements, questions, backchannels, fillers and disruptions. A maximum entropy classifier was used using the following external features:

1. length of the dialogue unit (measured by the number of words in the unit)
2. first two words of the dialogue unit

3. last two words of the dialogue unit
4. initial word of the following dialogue unit

Their model achieved a classification accuracy of 79.5%. They also used prosodic features to augment the model. Clearly, the words of the dialogue and the length of the dialogue unit are important features in identifying dialogue acts.

Transformation- Based Learning

Similar to the maximum entropy classifier described above, transformation-based learning (TBL) labels units in the data using local features trained on a separate training data set. TBL learns a sequence of rules from a tagged training corpus by finding rules that will correctly label many of the units in the data. In order to prevent an infinite number of rules that can be applied to the data, the range of patterns that the system can consider is restricted to a set of rule templates, which are manually developed by humans. This approach is similar to the method used for the Brill tagger, a part-of-speech tagger.[27] TBL has been used for dialogue act tagging in the same domain as Verbmobil,[28] using the following features in the dialogue unit:

1. cue phrases – These are manually selected sets of cue phrases that have been identified in other domains (e.g. Finally, In summary). However, cue words are less common in spoken dialogues and because these are usually determined manually by inspecting text for cue words, they are not readily available for spoken text.

2. dialogue act cues – This is defined as word substrings that appear frequently in dialogue for a particular training corpus.
3. entire utterance for a one, two or three word utterance
4. speaker information – This refers to the speaker identity.
5. punctuation marks
6. length of the utterance – This refers to the number of words in the utterance.
7. dialogue acts of the preceding utterances
8. dialogue acts of the following utterances

This approach was able to classify dialogue in the travel planning domain into 18 categories with an accuracy of 71.2%.

Feature Latent Semantic Analysis

Feature latent semantic analysis (FLSA) is an extension to Latent Semantic Analysis (LSA) and has been used for dialogue act classification.[29] The corpora used for this method include the Spanish CallHome, MapTask and DIAG-NLP.[30,31,32] FLSA enables the use of additional features, other than words, to be added into the typical LSA model. In LSA, a dialogue unit is represented as a vector of all words in the dialogue, where the meaning of the dialogue unit is a reflection of the meaning of all the words it contains. Singular value decomposition is then performed to find the major associative patterns in the data. After representing all dialogue units using words and some related features, classification is performed for each new dialogue unit by comparing it to each vector in

the training data. The tag of the vector with the highest similarity (using cosine measure) to the new dialogue unit becomes its label. This method is similar to the one used for a question-answering application developed by Bell Laboratories where calls are represented as vectors of words.[9] Calls are then classified into their corresponding destinations; each destination within the call center is likewise represented as an n-dimensional destination vector. Using cosine as a similarity measure between 2 vectors, the destination for a call is determined. The features used in this method, in addition to words, include:

1. previous dialogue act
2. initiative (refers to the person who is in control of the flow of conversation)
3. game (another set of labels, such as "information" or "directive" based on the MapTask notion of a dialogue game)[19]
4. duration (length of the dialogue unit)
5. speaker identity

In this method, the feature they identified to be most predictive is the notion of "Game." The problem, as they have identified, is that this feature is not readily available in real time. In this study, it is not clear whether using the previous dialogue act was any help in improving classification accuracy.

N-gram Language Modeling

Language modeling has been used for predicting dialogue acts given a sequence of dialogue units. Verbmobil was initially developed as a translation system for travel

planning.[28] In addition to finite state transducers, it uses a knowledge-rich approach for dialogue processing and summary generation. It has a rich hierarchy of dialogue acts which at the roots are domain-independent and at the leaves are very domain-specific. It performs dialogue act processing using a three-tiered approach: a planner, a finite state machine and a statistical n-gram model.

The planner understands the thematic structure of the domain and a contextual model allows backup strategies for the dialogue. The finite state machine describes the sequences of dialogue acts that are admissible in their domain and checks whether sequences of ongoing dialogue labels conform to these expectations. Finally, the statistical n-gram model predicts the appropriate speech act for a dialogue unit using unigrams, bigrams and trigrams that have been trained on dialogue annotated with the corresponding dialogue acts. Deleted interpolation is used for smoothing. In this method, we observe that the sequence of dialogue acts clearly play a role in dialogue act classification.

Hidden Markov Modeling

One research study compared various machine learning techniques for automatically computing dialogue acts from both transcribed and automatically recognized conversational telephone speech.[11] Using a standard for discourse annotation, Discourse Annotation and Markup System of Labeling (DAMSL),[33] 205000 dialogue acts from 1155 Switchboard conversations[34] were labeled into 42 categories. The dialogue model

is based on a Hidden Markov Model where individual dialogue acts correspond to observations emanating from the dialogue states. The goal is maximizing the probability of an utterance given all the observable features they selected. A dialogue grammar is described to constrain the sequence of dialogue acts using dialogue act n-grams. It determines the prior probability of utterance sequences. The features that were used in the model include the following:

1. speaker information
2. word n-grams – n-grams use n consecutive words (instead of just one word) as an additional feature in textual data representation with a reasonable bound on the length. They have been used for language modeling and for automatic spelling correction in the medical domain.[35]
3. prosodic features – These features were taken from the spoken dialogue data.
4. recognized words – These include words that were recognized by the automatic speech recognizer. A separate analysis was conducted using recognized words and manually transcribed words for dialogue act classification.

The model was highly successful and achieved a labeling accuracy of 71% on manual transcripts of their data.

2.1.3 Relevant Applications in Medicine

In the medical domain, discourse literature has been limited to theoretical study of dialogue representation. Discourse literature has mainly focused on sequential speech activities. In particular, discussions have focused on the rituals associated with conversational talk – the type of speech is predetermined, the place of occurrence in sequential talk is prescribed for a given topic, and phrasing is routinized.[36] Consider the conversation between a caregiver and a patient about a patient's complaint shown in Figure 2. Out of necessity, it is imperative that the caregiver seeks to find answers to specific questions in what appears as an "interrogative" manner to clarify an illness, which in this case is a cough. This is then followed by discussion about how the cough is treated and managed. Thus, on the surface, the conversation appears ritualized and appears to have a predictable sequence of utterances.

P: I am getting a yellow sputum.
N: You coughing that up?
P: Yes, sort of.
N: Are you running a fever?
P: 98.8 and it's just right in my throat. I don't know if I'm getting a head cold or what
N: Your throat is still sore?
P: yeah
N: Is this like sinus drainage that you have? Or are you coughing this up?
P: Basically, coughing, it's not sinus. It started with a sore throat
N: let me leave a message with Dr. Smith and I'll see if he wants to put you on something, ok? And I'll get back to you.

Figure 2: Dialogue between a patient and a caregiver

Although ritual is less observed in more informal telephone conversations, we adopt the notion that spoken medical dialogues are made up of sequential activities/segments or composed of sequential topics.

Recurring types of information have been seen to occur in medical data, and specifically in written medical documents. An example is the use of context models to represent medical publications.[37] Typically, medical publications follow a strict format with an introduction, literature review, methodology, results and discussion. A separate context model is built for clinical research articles compared to case reports and review articles as a basis for document representation. By manual context mark-up, the recurring semantic themes in each type of publication are identified manually (e.g. relevant tests, study type, relevant population) by human subjects. They are then used for manually labeling several publications for further clinical to research use.

A method for identifying basic patterns that occur in a medical plan was developed for a language generation system that briefs about a patient who has undergone coronary bypass surgery. Using supervised training on manually annotated narrated events, sequences of semantic tags are identified that occur frequently.[38] This is similar to motif/pattern detection in computational biology. Patterns are identified in a separate training set when they occur above a pre-set threshold of frequency. After obtaining a relatively large number of patterns, clustering is done to group together similar patterns and to allow better visualization of the quality of the patterns that are detected.

Our emphasis on spoken medical discourse sets us apart from the efforts to interpret written medical text.[3,39,40] We attempt to structure spoken dialogue and leverage the sequential nature of topics by using change in semantic type of a dialogue turn as a marker for a corresponding switch in topic of conversation.

2.1.4 Other Relevant Features

The major approaches in natural language processing rely on word and phrase similarities. Term repetition has been known to be a strong indicator of topic structure and lexical cohesion. In fact, word frequencies have been shown to produce successful results for text segmentation, dialogue act classification, and summarization.[26,28] We look at other features that are potentially useful in our thesis.

Word Substitutions

Word substitutions are the use of more general words instead of the actual one used in the data. These are used primarily to remove noise created by very specific numbers or names or to relax the constraints in matching a word with similar words. Text-specific attributes such as proper names, dates and numbers can be replaced by generic tags or placeholders.[41] In this thesis, we will use a placeholder to automatically replace numbers within the text. Another type of word substitution includes the use of word synonyms and word stems to relax word matching constraints. We did not use word stems because upon inspection of relevant features identified by the algorithm, words with similar stems did not appear to be highly predictive features.

Location

Relative location of textual units in the data has been shown to be a useful predictor of important units for summarization. In fact, lead sentences have been shown to produce

understandable and coherent text summaries.[42] This clearly indicates that, especially for written text, substantial amounts of relevant information are given in the first sentence of a paragraph. We will discuss in more detail how we use the relative location of a dialogue turn for our summarization algorithm in section 6.1.

Term-weighting

In order to augment simple word matching and decrease the value of frequently occurring words that are not very discriminatory, term-weighting such as TF*IDF has been used in several studies – from spelling correction to sentence extraction.[35,43] The total frequency (TF) of words in each block can be computed and adjusted according to the number of blocks in the dialogue that the word appears in. The greater the number of blocks containing a particular word, the lesser its weight. Since the number of blocks in which a particular word appears in is commonly referred to document frequency (DF), we inversely adjust the word's TF with its inverse document frequency (IDF). Each word can therefore be represented with its TF*IDF score, calculated as:

$$TF(word_i) \cdot \log(\text{Total number of blocks}/DF(word_i)).$$

Any unit of dialogue can therefore be arbitrarily determined and represented as a vector of word frequencies or TF*IDF.

Using various appropriate features and discourse units, we can induce dialogue structure to enhance our understanding of medical dialogues. The next section deals with text summarization in general and dialogue summarization in particular.

2.2 Text Summarization

A summary is a presentation of the substance of a body of material in a condensed form or by reducing it to its main points.[44] In recent years, a variety of summarization algorithms have been developed for text,[45,46] and are primarily applied for summarizing newspaper articles.[47,48] Identification of salient sentences and important content terms are key contributions of summarization research to date. Our work builds on these approaches in the design of a summarization algorithm for medical dialogues.

2.2.1 Summarization Goal

Independent of the approach to summarization, a summary has to accomplish specific goals – informative versus indicative. We discuss each goal in the next two paragraphs.

Informative versus Indicative

Informative summaries are shorter versions of the original text that act as surrogates to the original. They are meant to fully identify all relevant information in the original text in a condensed form. Examples of these are news summaries especially from multi-document sources.[47,48,49] Aone et al. uses LA Times and Washington Post and selected key sentences to produce an informative summary which preserves as much information as possible from the original text using TF*IDF (See section 2.1.4), durational and positional features of sentences. Sentences are assigned scores based on weighted

features and the highest scoring sentences are selected.[49] Kupiec et al. performed sentence extraction using scientific journal articles and created informative summaries by training a classifier using the following features: fixed-phrase features (e.g. "Summary"), paragraph features (e.g. first 10 paragraphs), thematic word features (e.g. most commonly occurring words in the document), and uppercase word features. In this thesis, we will focus on creating informative summaries so that we preserve the important and relevant contents of a dialogue for use in medical tasks.

Indicative summaries generally point out the main ideas of a text. It is more commonly used for information retrieval and text classification. The output is generally shorter and is not comprehensive in content. This approach is similar to "query-driven" summaries where the output is pre-determined by a user. In the latter, summaries are constrained by the specific interest of a particular user as opposed to a generic topic that is most important to the domain.

2.2.2 Summary Types

The output of a summarization system determines the approach that is taken in designing a summarization system. The types of output – extracts versus abstracts are described in this section.

Extracts versus Abstracts

Most approaches to text summarization involve extracting key sentences from the source text to form a summary.[25,43] These textual materials (whether sentences, phrases or

words) are called extracts. They are glued together in an appropriate manner to form summaries.

Abstracts are often generated from some "deep understanding" of textual material and requires detailed semantic analysis of the source text to enable generation of a shorter summary. The words used may be entirely different from the ones present in the source. Thus, in automatic abstractive summarization, a generation component is essential to create a textual summary from the source data.

2.2.3 Summarization Methods

Approaches to summarization have typically been classified into three major categories: (1) corpus-based approaches, which deal with textual features of the data, (2) discourse structure approaches, which leverage the discourse structure of data and (3) knowledge-rich approaches, which focus heavily on domain knowledge representation. We will elaborate on these approaches below.

Corpus-Based Approaches

Linguistic representation relies on textual features that are readily available from the corpus. Methods are typically based on domain independent machine-learning techniques based on surface-level indicators.[45,50] These features include location of sentences within a text, cue phrases, sentence length, title words and term-based statistics such as TF*IDF

(as described in section 2.1.4). Feature extraction and choosing the best combination of features predict which sentences are typically extracted to form summaries.

Discourse Structure Approaches

The corpus-based approach gradually evolved into exploiting discourse structure that is still predominantly based on linguistic features. Some summarization approaches focus on lexically cohesive text segments that are based on lexical chains.[51] This leverages knowledge from huge lexical resources (e.g. WordNet) to determine semantic relatedness of terms. Scores are computed for lexical chains and this determines extracts for a summary. Other summarization methods based on discourse structure include using rhetorical structure and topic segmentation. [52,53,54]

Another distinct approach for summarizing dialogues is based on Rhetorical Structure Theory, as described by Mann and Thompson.[55] In this method, dialogue units or segments are based on text structure and are thus domain independent. Segments are then related to each other in a hierarchical fashion, each relationship between two segments selected from a predefined set of relationships.[56] An unsupervised approach to recognizing certain types of discourse relations has been developed that was able to distinguish relations between adjoining segments of text as "contrast", "cause-explanation-evidence", "condition" and "elaboration".[56] Based on the relationships and the structure identified, which is typically a tree-based structure, a formula is derived for ordering the units of text and choosing the highest scored textual units.

Knowledge-Rich Approaches

Knowledge-rich representation has been used in limited domains for text summarization.[57,58] Text is typically represented using substantial domain knowledge and every word of an utterance has to be previously identified and encoded. In a research study by Hahn and Reimer, they created a knowledge base for each paragraph in a given text using terminologic logic. Terminologic logic is comprised of concepts and two types of relations: (1) properties, which link concepts to specific strings, and (2) conceptual relations, which denotes relations between two concepts. A text graph is built using generalization operators applied to each paragraph (which is referred to as a topic description after representing it in terminologic logic) by picking common elements and creating a more general node, until no additional new node can be created. Summarization occurs by leveraging the representation structures of the text. For example, a really concise summary is extracted by choosing only the root nodes of the text graph. Clearly, substantial domain knowledge has to be encoded in order to represent every new data that has to be summarized.

2.3 Dialogue Summarization

Spoken language summarization differs from written text summarization in several ways. For example, uppercase words cannot be expected in automatic transcription of spoken dialogues. Second, there are no keywords or titles (e.g. "Result", "Summary") to guide the reader about the contents of a dialogue. Third, the informal nature of dialogue typically does not contain fixed-phrases or cue phrases (e.g. "In summary"), as in text. Fourth, dialogues lack the structure customary in written text – topics switch in content without any visible delineation. Spoken language can further be distinguished by single-speaker, written-to-be-spoken text and unscripted spoken dialogues. In spoken systems, a whole range of phenomena have to be addressed, including interruptions and hesitations, speech recognition errors and disfluencies.[30]

In this thesis, we will focus on the structure of dialogue, without relying on acoustic or speech related features by using manually transcribed spoken dialogue. In previous work, prosodic features and acoustic confidence scores have been used in generating summaries for spoken text.[59] We will concentrate on linguistic and structural features for dialogue summarization in the summarization systems we discuss below.

Approaches to Dialogue Summarization

Several summarization systems have been developed for very limited task-based domains.[28,60] In particular, the MIMI and VERBMOBIL systems deal with travel

planning and reservations. The MIMI system uses finite-state transducers that process each utterance for words, basic phrases, complex phrases and domain patterns.[60] It recognizes the particular combination of subjects, objects and verbs necessary for correctly filling templates for a given information task using a set of predefined rules. It then ignores unknown input and merges redundant information. Summaries are generated from the template, which in this domain contain the current state of travel reservation at a particular point in the conversation.

Verbmobil was initially developed as a translation system for travel planning.[28] In addition to finite state transducers, it uses a knowledge-rich approach for dialogue processing and summary generation. It has a rich hierarchy of dialogue acts which at the leaves are very domain-specific. Each dialogue utterance is classified into a dialogue act using a language-modeling approach. It also identifies the propositional content of an utterance using a cascade of rule-based finite state transducers. It then represents the dialogue act and propositional content using a frame notation including several nested objects and attributes that cover the travel-planning task. A template based approach to summary generation is performed, which can be translated into multiple languages within this limited domain.

Several other summarization systems have been developed for summarizing topics in spoken human-human dialogues.[29,33,61] Gurevych and Strube worked on Switchboard data[34] to create representative extracts from the dialogue using automatic segmentation.[61] Based on semantic similarity of a segment of text to individual utterances within the

segment, relevant textual data are selected for summaries. Waibel, et al., also worked with the Switchboard data using an implementation of an algorithm based on maximum marginal relevance (MMR).[62] Each dialogue turn is weighted using the most common word (stem) that are highly weighted in a given segment while minimizing similarity from previously ranked turns. One dialogue turn is selected at each iteration until a pre-specified summary length is reached. Finally, Zechner used human-human dialogues which considered speech-recognition issues such as speech recognition word-error rate reduction, dysfluency detection and removal, and sentence boundary detection.[63] They then used a topic segmentation algorithm[64] to detect different segments. Also using MMR, they extracted sentences containing the most highly weighted terms while sufficiently dissimilar from previously ranked sentences.

2.4 Summary Evaluation Techniques

Automatic summary evaluation is a challenging task that is crucial for the continued development of summarization systems. It is especially challenging because of the following factors: (1) It requires an output (the summary) for which there is no single correct answer, (2) Humans may need to judge the system's output, which is very expensive, (3) Because of (2), information has to be presented to humans in a manner that is sensitive to the user's needs, and (4) Summarization involves compression and the compression rate may further complicate the evaluation.[65] Methods for evaluating text summarization can be broadly categorized into two categories: intrinsic evaluation and extrinsic evaluation.[66,67] We will discuss both approaches in this section.

Intrinsic Evaluation

Intrinsic evaluation techniques are designed to measure the coherence and informativeness of summaries. Clearly, summaries which are extracts may have readability and coherence issues especially if sentences are extracted out of context. In a particular study, judges assessed the readability of a summary based on general readability criteria, such as good spelling and grammar, a clear indication of the topic of the source document, understandability, and acronyms being presented with expansions.[68]

Several summarization systems were evaluated using intrinsic methods by measuring informativeness using precision and recall.[45,46] The typical approach involves creating a

"gold standard" or ideal summary, made by humans. The output of the summarizer is then compared to this "gold standard" and the system's output is typically scored higher if it contains greater overlap with the "gold standard." This approach also contains several pitfalls that have been identified previously:[67]

1. There is no single correct summary.
2. The length of the summary influences the evaluation results.

In order to create a "gold standard" summary that is reproducible, summaries have been created by multiple experts for the same text and agreement is measured between the human summaries. The "gold standard" is usually selected using majority opinion.

Precision and recall are typically used to evaluate summarization systems' outputs compared to the ideal summary. Another approach extends precision and recall by considering alternate sentences which are chosen by some human summarizers but were eliminated by majority vote.[69] If this alternate sentence is chosen by a system, they are assigned some weighted score for this sentence (instead of 0).

Different summary lengths influence evaluation results as well as human agreement in selecting statements for a "gold standard."[67] Thus, it is important to restrict the length of the summaries so that summary lengths are comparable when evaluating summarizers.

Extrinsic Evaluation

Extrinsic or task-based evaluation follows a key recommendation by Jones and Galliers for evaluating a natural language processing system – evaluations must be designed to address issues relevant to the specific task domain of the system.[66] Several methods have been developed that address certain tasks:

1. Relevance assessment – A person may be presented with a document and a topic and asked to determine how relevant the document is to the topic. Accuracy and time in performing the task are measured and studied. This framework has been used in several task-based evaluations.[67,70,71]

2. Reading comprehension – A person reads a full source or summary and then answers a series of questions (e.g. multiple choice test). The goal is to get the greatest number of correct answers. This framework has been used for various task-based evaluations as well.[72,73]

Extrinsic evaluations have the advantage of being able to assess the summaries when given a specific task. This is important for continued system development and provides a practical feedback especially for the users of the summaries.

Our work builds on these approaches in the design of a summarization algorithm for medical dialogues. However, 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 thus can be used to evaluate other summarization systems for medical dialogues.

3. Data

We first introduce our methods for data collection and describe basic characteristics of the medical dialogue. We then describe some techniques for data representation and clean-up.

3.1 Data Collection

We collected our data from the Lynchburg Nephrology home hemodialysis program, the oldest and largest such 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)[74] 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 manually transcribed, maintaining delineations between calls and speaker turns. The data were then divided chronologically into training and testing sets.

3.2 Data Representation

The transcribed spoken dialogues contained speaker delineations, and thus allows convenient identification of dialogue turns. We use each dialogue turn as our basic dialogue unit. We performed data scrubbing and stop word identification, as described below. We then represent each unit as a vector of words, with corresponding word counts.

3.2.1 Data Scrubbing

We removed all punctuations and extra white spaces in the transcribed text. After noting a substantial number of numerals (e.g. 96, 125.5) in the text, we decided to use a placeholder (a string "Integer") for each occurrence of a distinct real number in the text. This not only reduces the number of words in the feature vector but also augments the representation because of the semantic information. Finally, we removed all proper names of persons in the text to protect the patients' and caregivers' privacy.

3.2.2 Stop Words Identification

Stop words are common words that frequently occur in the text and contain very little additional information. Examples include "the", "a", and "and". We identified stop words by identifying the 20 most common words in the entire dialogue and removed them from the data. The stop words we selected are very similar to those published in the literature[75]

and are shown in Table I. In subsequent experiments, we did not see any improvement in performance by removing stop words. Thus, we decided to keep them in all the final models.

in	that
your	a
know	is
be	it
of	to
at	the
just	ok
on	and
I	you
was	are

Table I: 20 Most Common Words

4. Naïve Approaches to Structure Induction

We initially analyzed the first 25 dialogues that were recorded, containing 8,422 words.[76] Upon manual analysis of the dialogues, a total of 44 topics were identified. The topics, however, appear to be limited to three broad categories: clinical, technical and miscellaneous.[76] In an effort to automatically identify these segments of dialogue that contain the same topic, we performed unsupervised clustering on the data as described in the next section. In addition, we describe segment-based supervised classification at the end of this section.

4.1 Clustering

Clustering algorithms partition a set of objects (in this case, dialogue turns) into clusters.[77] According to the type of structures produced, clusters can be divided into hierarchical and non-hierarchical. Clustering has been referred to as "unsupervised classification" because clusters are created from natural divisions in the data and require no training or labeling.

In order to classify data into various clusters, a measure of similarity is necessary. A standard measure of similarity that has been frequently used is the cosine. We first represent the dialogue turns as a vector of word counts where a vector $\vec{x} \equiv \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}$ has n dimensions, corresponding to the frequency of n words in the turn. Cosine is computed between two vectors as follows:

$$\cos(\vec{x},\vec{y}) \equiv \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}}$$

After computing pair-wise similarities between units of dialogue using cosine, we can use a clustering algorithm to partition the dialogue into different topics.[77] This assumes that dialogue units with similar words have similar topics and thus are clustered together.

4.1.1 Clustering Implementation

We used an agglomerative clustering algorithm to group dialogue turns. A threshold was empirically set to cluster up to a cosine value of 0.5. Figure 3 shows the clustering algorithm we used. The algorithm was implemented in Java.

Given: set $X = \{x_1, \ldots, x_n\}$ of dialogue turns
A similarity function, sim, where $sim(x_a, x_b) \rightarrow$ Real Number
Initialization phase (assigns each dialogue turn to a cluster):
for i := 1 to n
$c_i := \{x_i\}$
$C := \{c_1, \ldots, c_n\}$

Merge clusters:
j := n+1
while C>1 and sim > 0.5
$(c_{n1}, c_{n2}) :=$ argmax $[sim(c_u, c_v)]$, where $c_u \in C$ and $c_v \in C$
$c_j := c_{n1} \cup c_{n2}$
$C := C \setminus \{c_{n1}, c_{n2}\} \cup \{c_j\}$
j := j+1

Figure 3: Clustering algorithm

4.1.2 Clustering Results

That's fine, between 8 8 30 that's fine.

That's a good idea you're putting blood flow up

That's fine thank you.

Good good, I want to let you know, I'll be going out of

town this weekend, on Sunday, I will not be dialyzing.

I'm fine.

Figure 4: Sample of an automatically generated cluster

The partial list of clusters generated by the clustering algorithm is shown in Appendix A. Although the results were able to group together short dialogue turns containing one to two words, it was not as useful for most of the dialogue turns. On manual inspection, dialogue turns are grouped together that belong to various general topics. An example of a cluster is shown in Figure 4. Moreover, similar dialogue turns that belong to the same topic were assigned to various unrelated clusters, as shown in Figure 5.

I wonder what time you could come tomorrow?
When do you want us over there?
What's your schedule Monday?
I couldn't get a doctor's appointment Tuesday evening.

Figure 5: Sample of clusters containing similar dialogue turns

4.2 Segment-based Classification

We decided to organize the dialogues manually into segments. We identified dialogue segments as consecutive dialogue turns that all pertain to the same general topic. After initially segmenting the dialogue manually into dialogue segments, we describe in the next section the algorithm we use to automatically identify the topics of individual dialogue segments in a separate test set.

4.2.1 Expert-derived features

The transcribed dialogues were divided into training and testing sets consisting of 71 segments each. Dialogues were divided manually into dialogue segments, which are composed of consecutive dialogue turns containing the same topic. Segments were then labeled manually according to the following four general headings: (1) Clinical if they pertain to the patient's health, medications, laboratory tests (results) or any concerns regarding the patient's health; (2) Technical if they relate to machine problems, troubleshooting, electrical, plumbing, or any other issues that require technical support; (3) Miscellaneous for all other topics, which are primarily related to scheduling issues and family concerns; and (4) Backchannel for greetings and confirmatory responses. These categories are described in more detail in section 5.1. A domain expert reviewed the training data and manually chose words that appeared predictive for each of the four categories.

Algorithm Development:

During training, each word identified by the expert is assigned a score, corresponding to the weighted count of instances it occurred in the training data for a particular category. The model is therefore composed of four classes containing a weighted set of words, derived from the training data. Examples of features selected for each class are shown in Table II. When a new data segment is presented, each class computes the votes from its list of pertinent words and the class with the highest vote wins. The algorithm is implemented in Common Lisp.

Class	**Features**
Clinical	Integer (167), blood (25), weight (18), pulse (17), pressure (16), low (15), feel (13), target (11), night (10), mean (10), sitting (9), standing (8)
Technical	machine (12), formaldehyde (6), water (6), pressure (6), blood (6), flush (5), green (4), lights (4), syringe (4), arterial (4), started (3)
Miscellaneous	Integer (44), call (8), back (8), going (8), today (7), week (6), night (5), tonight (5), tomorrow (5), time (5), Friday (4), morning (4), Saturday (4), Monday (4)
Backchannel	bye (42), you (23), hello (23), ok (23), Hi (15), thank (11), thanks (5), Good (4)

Table II: Expert-derived features (and scores in parentheses) used for the classification algorithm

Incorporating Semantic Types:

To study the contribution of semantic knowledge in increasing the accuracy of predicting relevant topics in our application, we used the UMLS Semantic Network.[6] For each pertinent word that was identified by the expert, and for each word in the test set, we used MetaMap to represent the word using its semantic type.[78] For our third model, we did a similar substitution using MetaMap for only the nouns in the same text because it achieved better predictive accuracy in our preliminary studies. In the latter case, we used Ratnaparkhi's tagger to identify the nouns in the data.[79] We compared the predictive accuracy of the three models.

The results of the three models in predicting each of the four categories are shown in Table III. T-test was performed to compare the best UMLS model with the base model and showed no significant difference.

Model	**Clinical** (n=25)	**Technical** (n=19)	**Backchannel** (n=10)	**Misc.** (n=17)	**Total** (n=71)
words	80%	37%	100%	35%	61%
UMLS	76%	32%	90%	18%	52%
UMLS nouns	60%	47%	90%	53%	59%

Table III: Accuracy of the models using expert-derived features

The best model achieves an accuracy of 61%, which is better than what we would obtain if we label each segment with the most frequent class (accuracy=35%). Nevertheless, this

unexpectedly low result demonstrates the complexity of semantic annotation for medical dialogues, and justifies the use of machine learning methods, which we will describe in the next section.

5. Structure Induction

We first introduce our annotation scheme followed by a description of the manual data annotation process. Next, we present a basic classification model that uses a shallow dialogue representation. Finally, we present a method for augmenting the basic model with background knowledge.

5.1. 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 IV.

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

Table IV: Examples of dialogue for each semantic type

Dialogue turns are labeled Clinical if they pertain to the patient's health, medications, laboratory tests (results) or any concerns or issue that 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.

5.2 Data Annotation

Two domain experts, specializing in Internal Medicine and Nephrology, independently labeled each dialogue turn in the training and testing data sets with its semantic type. Each annotator was provided with written instructions that define each category and was given multiple examples (see Appendix B). The distribution of semantic types for each set is shown in Table V.

Category	Training (n=1281)	Testing (n=293)	Total (n=1574)
Clinical	33.4%	20.8%	31.1%
Technical	14.6%	18.1%	15.2%
Backchannel	27.2%	34.5%	28.5%
Miscellaneous	24.7%	26.6%	25.1%

Table V: Semantic Type Distribution in Training and Testing Data Set

5.2 1. Kappa agreement

To validate the reliability of the annotation scheme, we computed the percentage of agreement between annotators. In addition, we accounted for chance agreement by using the kappa coefficient.[80] Percentage of agreement is defined as the number of dialogue turns for which both physicians gave the same label, divided by the total number of dialogue turns labeled. We computed the percentage of agreement to be 90%.

Kappa, on the other hand, is a measure of agreement between two observers taking into account agreement that could occur by chance (expected agreement), and is computed by:

$$Kappa \equiv \frac{Observed\,agreement - Expected\,agreeement}{100\% - Expected\,agreement}$$

Agreement increases as kappa approaches 1.0, with complete agreement corresponding to a kappa of 1.0. We computed the kappa to be 0.80, which is "substantial" agreement.[80] This kappa suggests that our dialogue can be reliably annotated using the scheme we developed.

5.3 Semantic-Type Classification

Our goal is to identify features of a dialogue turn that are indicative of its semantic type and effectively combine them. We present a basic model for classification followed by models augmented with background knowledge.

5.3.1 Basic Model

We discuss features selected for our basic model. This 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.`"

Another contextual feature we added is the entire dialogue segment containing the dialogue turn being classified. As we show in the previous paragraph, an utterance may be classified more appropriately if the context of the utterance is known. Thus, the question "`How are you doing?`" is more likely to be classified as clinical if it was mentioned within a segment discussing the patient's clinical condition.

Feature weighting and combination: We learn the weights of the rules in the supervised framework using Boostexter,[81] 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.

5.3.2 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. In this section we describe methods that bridge this gap by leveraging semantic knowledge from readily available data sources. These methods identify the semantic category for each word, and use this information to predict the semantic type of a dialogue turn. To show the advantages of this approach, 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.

Original:
`Uhumm, what you can do is during the treatment a couple of times, take your blood pressure and pulse and if it's high, like if it's gone up into the 100s, give yourself 100 of saline.`
UMLS Semantic Type:
`Uhumm what you can do is during the T169` **[Functional Concept]** `a T099` **[Family Group]** `of T079` **[Temporal Concept]** `take your T040` **[Organism Function]** `and T060 [Diagnostic Procedure] and if it's high like if it's gone up into the Integer give yourself Integer of T121` **[Pharmacologic Substance]** `T197` **[Inorganic Chemical]** `.`

Cluster Identifier:
1110111110110 111110100 1111111111 11100111101101110
1111011100000 111110100 1111010111101101 11111101010
1111010001001 1001 110111101000 11001111101101 11010100101
111110100 11110100110110 111110100 11110100110111 110011100011
111110100 11110100110110 10111111 11011101101101 11110111001100
11001111111111 1111001101110 11101110 111110100 11110100110110
1111101110 11111101010 1111010001000 1111001000110.

Table VI. Dialogue turn represented in its original form, augmented with UMLS semantic type and cluster identifiers. Terms in square brackets are included for illustrative purposes only.

We explored two orthogonal ways to add lexico-semantic knowledge into our system – (1) Unified Medical Language System (UMLS), a manually-crafted, large-scale domain ontology and (2) clusters of semantically-related words automatically computed from a large text corpus.

5.3.2.1 UMLS Semantic Types

Our first approach builds on a large-scale human crafted resource, UMLS. The UMLS Metathesaurus is the largest thesaurus in the biomedical domain.[6] Among other things, it provides a representation of biomedical concepts that are classified by semantic types, with hierarchical relationships among certain concepts. The UMLS contains information about over 1 million biomedical concepts and 4.3 million concept names from more than 100 controlled vocabularies and classifications (some in multiple languages) used in patient records, administrative health data, bibliographic and full-text databases and

expert systems. It includes vocabularies and coding systems designated as U.S. standards for the exchange of administrative and clinical data.[6] This resource is widely used in medical informatics, and has been shown to be beneficial in a variety of applications.[3,4,82]

Several applications have been developed to map natural language terms into these UMLS concepts. One particular project developed at the National Library of Medicine is the MetaMap Program.[78] This program maps natural language biomedical text to concepts in the UMLS. Noun phrases are extracted, variants are generated, and mapping to the set of UMLS concepts that covers the entire text is done using an exhaustive algorithm. This algorithm assigns a score to the (set of) UMLS concepts with the best match to a given amount of text (i.e. sentence) based on centrality (involvement of the head), coverage, cohesiveness and variation.[78] The output can be the UMLS concept(s) or the corresponding semantic type(s).

Using UMLS for Data Augmentation

For our experiments, we used the 2003 version of UMLS which consists of 135 semantic types. Each term that is listed in UMLS is substituted with its corresponding semantic category. An example of such substitution is shown in the second row of Table VI. To implement this approach, we first employ Ratnaparkhi's tagger[79] to identify all the nouns in the transcript. Then, using MetaMap, we extract the corresponding semantic type and replace the noun with the corresponding semantic type from the UMLS.[83] An utterance with the UMLS substitutions is added to the feature space of the basic model.

When we encounter a term that is not listed in UMLS, we do not perform any substitution and the term is left " as is." A term can have multiple matches in UMLS, and we pick the one with the highest MetaMap score. The UMLS semantic types can be combined using predefined relationships to generate more types. This can leverage the multiple matches that are sometimes encountered for a particular term. We chose to rely solely on single semantic types and leave this task for future research.

5.3.2.2 Automatically Constructed Word Clusters

Our second approach derives background knowledge from clusters of semantically-related words automatically computed from a large text corpus. An example of a cluster is shown in Figure 6. Being automatically constructed, clusters are noisier than UMLS, but at the same time have several potential advantages. 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, which may not be covered by UMLS. In addition, similarity based clustering has been successfully used in statistical natural language processing for such tasks as name entity recognition and language modeling.[84,85]

headaches	cramps	swelling
cramping	fluids	itching
radiation	saline	patience
pain		

Figure 6: An Example of a Cluster (1111000111110)

To construct word classes, we employ a clustering algorithm that groups together words with similar distributional properties.[85] The algorithm takes as an input a corpus of (unannotated) text, and outputs a hierarchy of words that reflects their semantic distance. The key idea behind the algorithm is that words that appear in similar contexts have similar semantic meaning. The algorithm computes mutual information between pairs of words in a corpus, and iteratively constructs a word hierarchy using a binary tree. It terminates when it has clustered all unique words into a pre-specified number of clusters identified by the user. Once clustering is completed, each word has a binary identifier that reflects the cluster where it belongs, and its position in the hierarchy. We use these identifiers to represent the semantic class of a word. The third row of Table VI shows an example of a dialogue turn where all the words are substituted with their corresponding identifiers. We add cluster-based substitutions to the feature space of the basic model.

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 inspected word clusters that were generated for arbitrary numbers of cluster sizes. Partial results of word clusters when the algorithm was terminated after achieving 1000, 1500, 2000 and 5000 clusters are shown in Appendix C. We empirically determined that the best classification results are achieved for 2000 clusters.

5.3.3 Results of Semantic Type Classification

Table VII 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, the UMLS augmented model and the cluster based 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, durational and contextual features, and is augmented with background information, obtained through statistical clustering.

Models	**Accuracy** (n=1281)
Dialogue turn	69%
Dialogue turn with length	70%
Dialogue turn with previous turn	68%
Dialogue turn with corresponding dialogue segment	61%
Basic Model (Dialogue turn with length and previous turn)	**70%**
Basic Model + UMLS	**71%**
Basic Model + 2000 clusters	**73%**

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

The first four rows of Table VII 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%. Adding the dialogue

segment as an additional feature worsens the model's performance. We therefore decided to omit this feature in succeeding models. Table VIII shows the most predictive features for each category.

The last two rows in Table VII demonstrate that adding background knowledge improves the performance of the model, although not significantly. The model based on statistical clustering outperforms the basic model by 3%, compared to UMLS augmentation which improves the performance by 1%. 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, as we will describe in the next section.

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 VIII: Examples of predictive features

An interesting finding of this research is that background knowledge did not improve the performance of the base model profoundly. We explain this finding by the markedly different vocabulary used in written and spoken discourse and the significantly lesser term coverage of consumer terms within UMLS.[86] We examined this phenomenon further

and found that MetaMap was only able to extract semantic types for 1503 of 2020 (74.3%) noun phrases that were identified in the data. Moreover, a significant fraction of nouns are mapped to the wrong category. For instance, the word "kidneys" is labeled as a "body part", while in our corpus "kidney" always refers to a dialyzer. This problem, in particular, is difficult to address when one is building a huge generic vocabulary for the entire medical domain. Clearly, vocabularies have to be tailored to appropriate users and specialties of medicine, especially when choosing the correct meaning of a given term. We encounter problems both with discrepancies in word usage and lay terminology that are not present in UMLS. The discrepancies between word usage in spoken and written language as well as differences in lay and expert terminology present a distinct problem in using UMLS for processing spoken medical dialogue. A corpus-based acquisition of semantic knowledge provides a promising solution for this problem.

5.3.4 Using a Sequential Model for Semantic Type Classification

We explored the potential contribution of the sequential nature of medical dialogue to semantic type classification. First, we determined whether the distribution of the semantic types of succeeding dialogue turns given the semantic type of the current dialogue turn is uniform. Second, we used the label of the previous dialogue turn as a feature in predicting a dialogue turn's semantic type to determine its possible utility. Finally, we utilized a sequential classification model that would explicitly model the label of the previous dialogue turn in predicting the succeeding one.

5.3.4.1 Semantic Type Transition

We determine the distribution of semantic types following a current dialogue turn's semantic type by counting the number of dialogue turns that belong to each semantic type following each turn in our dialogue. Table IX shows the distribution of semantic type transition from one dialogue turn to the next.

Current Dialogue Turn's Semantic Type	**Succeeding Dialogue Turn's Semantic Type**			
	Clinical	**Technical**	**Backchannel**	**Miscellaneous**
Clinical (488)	287	12	142	30
Technical (240)	9	130	77	16
Backchannel (449)	132	63	111	100
Miscellaneous (395)	27	11	101	236

Table IX: Semantic Type Transition from Current Dialogue Turn to the Next

We observe from Table IX that a technical dialogue turn is likely to be followed by another technical turn. A technical turn is usually preceded by another technical turn as well (60% of the time). The distribution is clearly not uniform except for backchannels, where the preceding turn's semantic type is more uniformly distributed. Thus, we anticipate that leveraging the sequential nature of dialogue would augment the classification accuracy for the remaining three categories – clinical, technical and miscellaneous. This is quite desirable because classifying turns into the backchannel category is a relatively easier task.

5.3.4.2 Utilizing the Previous Turn's Label

To determine whether the previous turn's semantic type would augment our classifier, we added the actual semantic type of the immediately preceding dialogue turn in the current dialogue turn's feature vector. We did the same using the two previous turns' semantic types. For leading dialogue turns, which have no preceding turns, we added a label called "none." The results are shown in Table X.

Semantic Types Added	**Classification Accuracy**
One previous turn	75%
Two previous turns	77%

Table X: Classification Accuracy Using the Semantic Types of Previous Turns

As expected, we show that there is improvement in the model's accuracy when we added the semantic types of the previous dialogue turns. However, when performing semantic type classification on test data, we do not know the preceding turns' labels beforehand. Thus, we hope to model the semantic type(s) of the preceding turn(s) with the rest of our features, as we discuss in the next section.

5.3.4.3 Conditional Random Fields

In this section, we present conditional random fields or CRF – a framework for building probabilistic models for sequential data.[87] In this framework, the relationship between adjacent pairs of labels is modeled as a Markov random field, solely conditioned on the observed inputs. We therefore model how adjacent labels influence each other through the input features. For example, we model the conditional probability of a possible semantic type sequence $\mathbf{t} = t_1, \ldots t_n$ given input data $\mathbf{o} = o_1, \ldots, o_n$ and maximize the probability $p(\mathbf{t} \mid \mathbf{o})$. We used the MALLET implementation of CRF.[88] Table XI shows the results of using CRF to predict the semantic types of dialogue turns on our test data.

Order	**Accuracy**
0	62%
0, 1	56%
0, 1, 2	53%

Table XI: Accuracy of CRF Model

In Table XI, we used various setting of MALLET to determine whether previous turns' semantic types will positively affect our model. In this implementation of CRF, setting the order to "0" directs the classifier to create predictions without regard for the previous dialogue turns' semantic type. Setting the order to "1" models the immediately preceding dialogue turn's semantic type and setting the order to "2" models the two previous turns' semantic types. We find that the classification accuracy actually decreases compared to

our current models. Perhaps, the low accuracy exacerbates the model's poor performance when taking into account the previous turns' semantic types.

6. Summarization

In the next section, we describe our method for automatically summarizing our dialogues. We create informative summaries using extraction of key dialogue turns based on the induced semantic structure of the dialogues.

6.1 Summarization Method

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.[76] 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 segment type. An example of such segmentation is shown in Figure 7.

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.[89] 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 7 shows one run of the algorithm. The summarizer compresses a conversation of 14 into five key dialogue turns.

Segmented Dialogue	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: You don't have any pumps open where? On your catheter? P: I have pressures a little bit there. N: I can hear the warning. Does it flush ok? *P: Yeah* N: I will try switching the ports. Start the pump	Technical

	and clamp off your lines and try switching the ports. And then turn it on and see what happens 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: How are you feeling? P: I feel fine. 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: And how's your weight now P: 129.2 N: Your blood pressure medicine, I'll have you finish that. P: I finished taking that on Friday N: Oh, so you finished taking that Friday, and the diarrhea and nausea, all that stopped. P: Yuh.	Clinical
	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?	Miscellaneous
Summarized Version	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	Technical Technical

	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?	 Clinical Clinical Miscellaneous

Figure 7: Segmented Dialogue and the Summarized Version (P: patient, N: nurse)

6.2 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 Chapter 5 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.

7. Evaluation

In this section, we detail our evaluation protocol. First, we describe two alternative summaries that we use for comparison with our system – a gold standard summary and a baseline naïve summary. Second, we introduce two evaluation frameworks for testing our automatically-generated summaries.

7.1 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 (see Appendix D for instructions). For each dialogue, we restricted the number of turns that the human subject could select, from one to a third of the dialogue's original size. This way, we obtain 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 0.50, which is "substantial" agreement.[80] Although kappa was not impressive, 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. The odds ratio is 10.8. It indicates that the odds of Subject 2 making a positive

decision increases 10.8:1 for cases where Subject 1 makes a positive decision ($p<0.0003$, log odds ratio) These two measurements indicate that dialogue turn extraction can be reliably performed by humans in our domain.

7.2 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.[63,90] 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 chose not to do a "lead summary" baseline because initial utterances in dialogues are typically backchannels and are not very informative. We expect that this would perform worse than a random baseline.

We therefore have the complete dialogue and four types of summaries for each dialogue: the "gold standard", a randomly generated baseline, and two semantic type based summaries. Appendix E shows a sample of all four summaries with the original complete dialogue.

7.3 Intrinsic vs. Extrinsic Evaluation Techniques

Our evaluation is composed of two parts – intrinsic and extrinsic.[66,67] 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:

$$precision \equiv \frac{\#\ Documents\ Correctly\ Chosen}{\#\ Documents\ Chosen}$$

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

$$recall \equiv \frac{\#\ Documents\ Correctly\ \mathrm{Re}cognized}{\#\ Documents\ Should\ Have\ Been\ \mathrm{Re}cognized}$$

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.[91,92] (see Table XII) The questions address relevant issues in clinical assessment, technical support and overall delivery of quality patient care.

We distributed 360 dialogues, comprised of the complete version of 80 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. There were only 40 summaries based on the "predicted semantic types" of the dialogue turns because the rest of the data was used for

training. We had nine licensed physicians (who did not participate in the selection of questions or in the manual summarization process) answer each of the six questions using each of 40 dialogues. They received written instructions prior to performing their task (see Appendix F). 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 XIII 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.

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 XII: Questions used in task based evaluation

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 XIII: Answer distribution across the six questions

7.4 Evaluation Results

In this section, we report the results of the intrinsic and extrinsic evaluations of our automatically-generated summaries.

A. Intrinsic Evaluation

The precision, recall and F-measure for the random baseline and the computer-generated summaries are shown in Table XIV. The results indicate that machine-generated summaries outperform random summaries by a wide margin. The results of a 2-tailed Fisher's Exact test comparing various summaries are shown in Table XV. 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 as we mentioned earlier. It is more significant 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 clearly demonstrate the contribution of structural information to text summarization.

	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
# of dialogue turns	183/516 (35.47%)	199/516 (38.57%)	277/516 (53.68%)

Table XIV: Precision, Recall and F-measure for 40 Dialogues

	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	1.38×10^{-4}	7.94×10^{-4}	0.4580
Recall	2.53×10^{-6}	1.03×10^{-16}	3.23×10^{-4}

Table XV: p-values using 2-tailed Fisher's Exact Test comparing precision and recall ($p<0.05$ is statistically significant)

B. Extrinsic Evaluation

The total correct responses, compared to the responses for the complete dialogue, are shown in Table XVI below for the first three summaries. The computer-generated summaries based on the true labels of the dialogue turns had higher scores across all questions, compared to the random summaries. The "gold standard" summaries also performed better than the randomly generated summaries, except for question 5 (Did the patient need to make, verify, cancel or reschedule an appointment?).

	Q1	Q2	Q3	Q4	Q5	Q6	Total
Random	58 (72.5%)	59 (73.8%)	70 (87.5%)	50 (62.5%)	59 (73.8%)	74 (92.5%)	370 (77.1%)
Manual	66 (82.5%)	70 (87.5%)	71 (88.8%)	53 (66.2%)	50 (62.5%)	74 (92.5%)	384 (80.0%)
Computer-generated using true-label	68 (85.0%)	67 (83.8%)	73 (91.2%)	54 (67.5%)	68 (85.0%)	76 (95.0%)	406 (84.6%)

Table XVI: Correct responses comparing three summaries

We report the results of physicians' answers to each of our six questions when given various summaries, including ones based on predicted semantic types 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 XVII 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.

	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 XVII: Correct responses comparing four summaries

Statistical significance was measured using one-tailed Sign test as shown in Table XVIII. This test is applicable for our evaluation: we want to measure the degree of improvement our method has over the baseline. 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.[93] The automatically generated summaries outperform random summaries on 5 questions, with a tie for the sixth. More importantly, there is no significant difference between computer-generated summaries and manually-generated summaries.

	Sign Test (One-tailed, n=5)
Computer-generated using true semantic type vs. Random	p=0.031
Computer-generated using predicted semantic type vs. Random	p=0.031
Computer-generated using true semantic type vs. Manual	NS (not significant)
Computer-generated using predicted semantic type vs. Manual	NS (not significant)

Table XVIII: Comparison of the accuracy of the summaries using Sign Test ($p<0.05$ is statistically significant)

The importance and complexity of the task require substantial participation from physicians. These doctors are able to make intelligent presumptions even when given simple randomly-generated summaries. In spite of this limitation, we still demonstrated that physicians perform significantly better in answering very important questions related to patient care when given our summaries compared to a simple baseline. More significantly, we demonstrate that although not statistically significant, our summarizers provided physicians with summaries that allowed them to answer pertinent questions more accurately than when they were using manually generated summaries. The physicians were only able to answer two questions more accurately using the manual

summaries compared to the more conservative automatic summarizer, which relies on predicted semantic types. These questions are as follows:

(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?

The rest of the time, the automatically generated summaries were better able to provide sufficient information for the physicians to answer more accurately. Expectedly, manually generated summaries created by physicians invariably focus more on issues regarding clinical care. Thus, they may not have included as much information about technical and scheduling concerns. When we used the "true semantic types" of the dialogue turns in the automatic summarizer, the summaries actually allowed physicians to answers more questions correctly than when they used the manual summaries ($p=0.10$, Sign test). Although not statistically significant, the automatic summarizer using "true semantic types" provided summaries that allowed physicians to answer more questions correctly for five of the six questions we provided.

The framework we developed for extrinsic evaluation emphasizes the importance of selecting appropriate tasks for a summarization system while in the development phase. The questions we identified are broad in coverage and were selected independent of the summarization methods. This addresses our summarization goal – to create informative summaries that capture as much information content as possible. We show that this framework can be used in separating summaries generated using simple random summarizers from automatically generated summaries using our methods.

8. 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 develop and evaluate semantic categories that are relevant to our specific domain. Although the categories are broad in coverage, they capture major topics in segments of our dialogues and are practical distinctions for identifying relevant topics for specific care providers. The categories are sufficiently distinct and two physicians are able to perform manual annotation with reasonable agreement, illustrating that the annotation scheme is stable.

We demonstrate how we can improve the performance of our method for structure induction by augmenting our data using two orthogonal sources of information – UMLS semantic types and automatically-induced word clusters. We recognize the importance of identifying words that are semantically related (e.g. headache and pain) or have similar meanings within the domain (e.g. hurts and pains). We achieve modest improvement when we incorporated these knowledge sources into our feature set.

We demonstrate the utility of our structural abstraction by incorporating it into an automatic dialogue summarizer. By eliminating backchannels, we are able to condense the dialogues and remove unnecessary information. Using the rest of the semantic categories, we are able to select dialogue turns that contain relevant information for our summaries by leveraging the finding that a change in semantic category signals the beginning of a new topic. Clearly, the first utterance when a new topic is discussed contains important information for an informative summary.

Our evaluation results indicate that automatically generated summaries exhibit high resemblance to summaries written by humans. More importantly, we show that the summaries are potentially useful in a medical setting. We develop a framework for evaluating our summaries based on a task that clinicians are expected to perform in delivering quality health services. Our task-based evaluation shows that physicians can accurately answer questions related to patient care by looking at the summaries alone, without reading a full transcript of a dialogue. This is a significant result because it spares the physician from the need to wade through irrelevant material ample in dialogue transcripts.

Although we analyzed transcribed spoken dialogue in the home hemodialysis domain, the methods for structure induction and summarization can potentially be applied to other medical specialties. The semantic categories may need to be tailored for specific domains and specific user goals. However, the techniques for acquiring and incorporating additional knowledge sources are not limited to this domain. Furthermore, the evaluation

framework that we used can be applied to evaluating summaries in other medical specialties. The extrinsic evaluation, in particular, contains questions that are sufficiently broad and applicable in other clinical areas. More importantly, the task is able to distinguish our automatically-generated summaries from summaries created using simpler methods.

8.1 Future Work

In the future, this work can be extended in three main directions. I will discuss each of these three approaches below:

1. Our method can be applied to automatically recognized conversations. The use of automatic speech recognition is a logical next step in dialogue analysis. This will allow a completely automated process from spoken conversations to summarization. Clearly, automatic speech recognition will introduce mistakes in a transcript. In addition, one needs to address sentence boundary detection, dysfluency repair and speaker identification. At the same time, however, it will provide 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, the use of acoustic features can be used to compensate for recognition errors in the transcript.

2. The annotation scheme can be refined to include more semantic categories. We can develop a hierarchical annotation scheme, which would contain more specific categories within the same domain. Another approach would be to augment the categories with more generic semantic labels (e.g. "greeting", "accept", "reject") and use a hierarchy of dialogue acts which at the roots are domain-independent and at the leaves are very domain-specific, similar to the approach taken for Verbmobil.[28] This would support a deeper analysis of medical dialogue. To achieve this goal and to

further enhance the accuracy of the present model we described, several approaches can be taken.

a. Immediate improvement can possibly be gained from performing dialogue turn segmentation. A dialogue turn can be segmented further into utterances (phrases or sentences that constitute a speech act). While a dialogue turn can have multiple utterances, an utterance may also contain multiple dialogue turns, especially when a dialogue turn is comprised of a backchannel. Automatic segmentation of dialogue is still a challenging problem.[94]

b. We observed modest improvement from using knowledge augmentation of our base model with both automatically generated word clusters and UMLS semantic types. Although this improvement was not significant, further improvements in performance can probably be attained from fine-tuning either approach. For generating word clusters, the best approach would be to find a conversational data set that has similar content to the system being developed. If unlabelled transcribed data for a similar domain can be obtained for word clustering, this may improve the model's performance significantly. In addition, distributional clustering algorithms are typically trained on corpora with 100 million words.[85] Our corpus is two orders of magnitude smaller because so much data is difficult to obtain in the dialysis domain. Once more data become available, word clustering using this larger data set would be a logical next step.

In addition to augmenting the word clustering approach, we also believe that substantial improvement may be attained from augmenting UMLS with lay terminology.[86] More significantly, vocabularies have to be tailored for appropriate users, tasks and specialties of medicine. In our case, we need vocabularies tailored towards patients and ones that are applicable to dialogue in the hemodialysis domain. We encountered substantial discrepancies in word usage, which are not typically used in this sense in any other context (e.g. kidneys referring to dialyzers). While significant research has been developed for word sense disambiguation, this relies heavily on predefined meanings for specific words. New meanings for terms and new domain-specific terms may have to be identified and added to existing vocabularies to maximize their benefit. These new terms will have to be added in a principled manner to existing vocabularies (e.g. UMLS) so that they do not further complicate this already huge metathesaurus. It is clear, however, that our methods for structure induction can benefit from additional relevant knowledge resources.

c. More expressive statistical models may be used to capture the structure of medical dialogue. Possible modeling methods include hidden Markov models and conditional random fields.[87,95] Although we did not see any improvement in our models using the latter, it was clear that knowing the category of the immediately preceding dialogue turn helps predict the next one. Furthermore, we see that transition from one category to the next is not uniformly distributed. Perhaps

augmenting the current classification accuracy of our model can boost the performance of these sequential modeling techniques.

More innovatively, further research may be done on inducing the structure of a model that would better represent medical dialogue, which may not turn out to be simply a sequential chain of utterances, but may capture a more elaborate structure such as a stack-based model.[96] In a stack, one might represent dialogue segments, which are composed of sequences of utterances focusing on a specific topic. These segments may, in turn, contain sub-segments with more specific subtopics that are still related to that of the current segment. Methods that would explicitly represent these more complex structures in addition to local features for dialogue act labeling might greatly enhance medical dialogue analysis.

3. Query-based summarization may be performed as opposed to generic summarization.[97] 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 as well as patients, summaries may be personalized towards their needs. Further research needs to be designed to understand what specific goals are relevant in different medical domains and for different users in an unobtrusive manner. If the primary goal for a specific care provider is to detect and summarize scheduling utterances, for example, we can train our machine-learning techniques with more utterances in the miscellaneous category. In addition, further

elaboration of this broad category can be performed to create more specific semantic types for scheduling appointments.

The methods described in the preceding paragraphs can be used to address either a deeper or a more focused analysis of spoken medical dialogue.

9. Appendices

Appendix A: Partial Results of the Agglomerative Clustering Algorithm

```
**************************************************
765
 hello
277
 hello this
528
 hello
268
 hello
637
 hello
537
 hello this
549
 hello this
231
 hello
661
 hello
527
 hello
548
 hello
483
 hello
10
 hello
481
 hello
577
**************************************************
**************************************************
 uhumm
353
 uhumm
579
**************************************************
**************************************************
 Bye
182
 Bye
52
 Bye
135
 Bye
230
**************************************************
**************************************************
 No 10 hours fine
212
 No
```

235
No didn't call anybody
65
No didn't
233
****
****
thanks
364
thanks
325
****
****
what weight
96
Pulse 92 Now what weight
292
****
****
can do today if that's
491
if doesn't give headache can do
632
****
****
Alright I'm off tomorrow night
125
I'm all off
108
****
****
Bubble up through
656
let bubble up through
655
****
****
Good
87
Good
18
Good
78
Good
56
****
****
Right
721
Right take another one
388
Right
418
Right
31
Right

1
Right
196
Right
138
****
****
let me look number
618
well let me call me tomorrow sometime let me what they are after have done rinse formaldehyde
222
Ok let me call him
273
well let me call then will have him call
262
said want me call let how do
552
let me look think real pause good
599
give me call let me how you're feeling then we'll decide then
449
****
****
want make sure What's his target weight
404
mmm he's not much over his target weight
413
67 over his target weight I'm sure
555
****
****
Thank feel today
264
Thank
180
Thank
228
Thank
240
****
****
So mean flush back forth
383
flush back forth get blood moving back forth through catheter
386
****
****
Monday
141
what schedule Monday
572
****
****
aha
624
aha

532
aha
510
aha
534
aha
470
aha
538
****
****
right
564
right
420
You're right
754
right
553
****
****
yuh
601
yuh
718
yuh
593
yuh
403
yuh
724
yuh
609
yuh
730
1:00 yuh
545
****
****
Thanks lot
326
Thanks
8
Thanks for calling
132
****
****
Ok well I'll ask her I'll call back let
7
I'll ask coz I'm not sure honest with but I'll ask I'll call back let
5
****
****
will
340
will fine

153
will fine
540
will
743
****
****
Yuh
728
Yuh he 200
566
Yuh
359
Yuh That's her problem
712
Yuh
194
Yuh
658
200
567
****
****
OK me too
227
OK
257
****
****
hi I'm calling back see laughs
639
hi
551
hi I'm
766
hi this
638
hi
550
****
****
Alright
45
Alright
131
Alright
47
Alright
55
Alright
21
Alright
238
Alright
29
****

Appendix B: Request for Annotation

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

B.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.

B.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 C: Samples of word clusters for various cluster sizes

C.1 Results using 1000 clusters

Below are samples of word clusters that were obtained when the clustering algorithm was terminated upon reaching 1000 clusters.

00101001000	olives
00101001000	meats
00101001000	administrators
00101001000	businesses
00101001000	corporation
00101001000	profits
00101001000	eggs
00101001000	corporations
00101001000	techs
00101001000	regulations
00101001001	Remove
00101001001	substitue
00101001001	tecs
00101001001	staffs
00101001001	latter
00101001001	beaches
00101001001	churches
00101001001	docs
00101001001	Australian
00101001001	employees
00101001001	surgeons
00101001001	physicians
00101001001	docs
00101001001	doctors
00101001001	Doctors

C.2 Results using 1500 clusters

Below are samples of word clusters that were obtained when the clustering algorithm was terminated upon reaching 1500 clusters.

001111101101 Lawyers
001111101101 Caregivers
001111101101 Recirculation
001111101101 Consumers
001111101101 Attorneys
001111101101 Others
001111101101 Lord
001111101101 officials
001111101101 People
001111101101 Doctors

110001011001 parathormone
110001011001 triglyceride
110001011001 unexplained
110001011001 testosterone
110001011001 TSH
110001011001 PRA
110001011001 hematocrit
110001011001 hgb
110001011001 urea
110001011001 hct
110001011001 phos
110001011001 hemoglobin
110001011001 PTH

C.3 Results using 2000 clusters

Below are samples of word clusters that were obtained when the clustering algorithm was terminated upon reaching 2000 clusters.

10110100011011	mytral
10110100011011	aortic
10110100011011	newmitral
10110100011011	mytrial
10110100011011	Thyroid
10110100011011	mitral
10110100011011	thyroid
10110100011011	parathyroid
10110100011011	pth
101101000111001	Decrease
101101000111001	diff
101101000111001	creainine
101101000111001	phosphous
101101000111001	creatinine
101101000111001	creatnine
10110100101001	Opthamologist
10110100101001	Docs
10110100101001	obgyn
10110100101001	neph
101101011100111110	Dietician
101101011100111110	nefrologist
101101011100111110	Nephro
101101011100111110	Nephrologist
101101011100111110	dr
101101011100111110	doc

C.4 Results using 5000 clusters

Below are samples of word clusters that were obtained when the clustering algorithm was terminated upon reaching 5000 clusters.

111000100 staff

1110001010 doctors

11100010110 nurses

11100010111 techs

1110100111000 doc

11101001110010 dr
11101001110010 abilities

11101001110011 Doc

1110100111010 surgeon

11101001110110 RN

11101001110111 Doctor

111010011110 doctor

1110100111110 neph

11101001111110 nephrologist

111010011111110 dietician

111010011111111 hemos
111010011111111 Neph
111010011111111 heme

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

D.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.

D.2. Example of dialogue turn selection (underlined text)

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 E: Complete dialogue with four summaries (P: patient, N: nurse)

Complete Dialogue	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: You don't have any pumps open where? On your catheter? P: I have pressures a little bit there. 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 P: ok 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. P: Mmm

	N: How are you feeling? P: I feel fine. 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: And how's your weight now P: 129.2 N: Your blood pressure medicine, I'll have you finish that. P: I finished taking that on Friday N: Oh, so you finished taking that Friday, and the diarrhea and nausea, all that stopped. P: Yuh 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?
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 F: Instructions given to evaluators

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?

10. 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] 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.
[4] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo. 2004: 565-72.
[5] Hsieh Y, Hardardottir GA, Brennan PF. Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses. Medinfo. 2004: 511-5.
[6] McCray A. Miller R. Making the Conceptual Connections: The UMLS after a Decade of Research and Development. J Am Med Inform Assoc. 1998 Jan-Feb; 5(1):129-30.
[7] Document Understanding Workshop. HLT/NAACL Annual Meeting. Boston, MA. May, 2004. In: http://duc.nist.gov/. Accessed on: June 16, 2005.
[8] Gorin A. Processing of semantic information in fluently spoken language. Proceeding of Intl. Conf. on Spoken Language Processing (ICSLP). 1996; 2: 1001-1004.
[9] Chu-Carroll J, B Carpenter. Vector-based natural language call routing. Computational Linguistics. 1999; 25(3): 361-388.
[10] Reithinger N, Maier E. Utilizing statistical dialogue act processing in Verbmobil. Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, MA. 1995: 116-121.
[11] Stolcke A, Coccaro N, Bates R, Taylor P, Ess-Dykema C, Ries K, Shriberg E, Jurafsky D, Martin R, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics. 2000; 26(3): 339-373.
[12] Samuel K, Carberry S, Vijay-Shanker K. Dialogue act tagging with transformation-based learning. Proceedings of COLING/ACL. 1998: 1150-1156.
[13] Austin, J. L. How to do Things with Words. Clarendon Press, Oxford. 1962.
[14] Searle, J. R. Speech Acts. An Essay in the Philosophy of Language. University Press, Cambridge. 1969.
[15] Hobbs J. On the coherence and structure of discourse. CSLI Technical Report 85-37. Stanford, CA. 1985.
[16] Grosz B, Sidner C. Attention, intentions and the structure of discourse. Computational Linguistics. 1986; 12(3): 175-204.
[17] Polanyi L. A formal model of the structure of discourse. Journal of Pragmatics. 1988; 12: 601-638.
[18] Mann W, Thompson S. Rhetorical structure theory: Toward a functional theory of text organization. Text. 1988; 8(3): 243-281.
[19] Carletta J, Isard A, Isard S, Kowtko J, Doherty-Sneddon G, Anderson A. The reliability of a dialogue structure coding scheme. Computational Linguistics. 1997; 23: 13-31.

[20] Cohen PR, Perrault CR. Elements of a plan-based theory of speech acts. Cognitive Science. 1979: 3; 177-212.
[21] Traum D. Speech acts for dialogue agents. In: Wooldride and Rao (eds.). Foundations of Rational Agency. Kluwer. 1999.
[22] Ang J. Liu Y. Shriberg E. Automatic dialogue act segmentation and classification in multiparty meetings. Proc. ICASSP, Philadelphia. 2005.
[23] Brown G, Yule G. Discourse Analysis, Cambridge University Press. 1983.
[24] Levinson, SC. Pragmatics. Cambridge, England: Cambridge University. 1983.
[25] Waibel A. Bett M, Finke M. Meeting browser: Tracking and summarizing meetings. Proceedings of the DARPA Broadcast News Workshop. 1998.
[26] Gorin A, B Parker, R Sachs and J Wilpon. How may I help you? Proc. of IVTTA, 1996: 57-61.
[27] Brill E. Transformation-based error driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics. 1995; 21(4): 543-566.
[28] Wahlster W. Verbmobil : Translation of face-to-face dialogues. Proc of MT Summit IV. Kobe, Japan. 1993.
[29] Serafin R, Di Eugenio B. FLSA: Extending latent semantic analysis with features for dialogue act classification. 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain. July 2004.
[30] Levin L, Thyme-Gobel A, Lavie A, Ries K, Zechner K. A discourse coding scheme for conversational Spanish. Proceedings ICSLP. 1998.
[31] Anderson A, Bader M, Bard E, Boyle E, Doherty GM, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, Sotillo C, Thompson HS and Weinert R. The HCRC Map Task Corpus. Language and Speech. 1991; 34: 351-366.
[32] Douglas M, Towne. Approximate reasoning techniques for intelligent diagnostic instruction. International Journal of Artificial Intelligence in Education. 1997.
[33] Core M, Allen J. Coding dialogs with the DAMSL annotation scheme. Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines. Cambridge, MA. November, 1997: 28-35.
[34] Godfrey JJ, Holliman C. Switchboard: Telephone speech corpus for research and development. Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing. San Francisco, CA. March, 1992; 1: 517-520.
[35] McInnes BT, Pakhomov S, Pedersen T, Chute C. Incorporating bigram statistics into spelling correction tools. Medinfo. 2004.
[36] Ainsworth-Vaughn N. The discourse of medical encounters. In: The Handbook of Discourse Analysis (Deborah S. Schiffrin, ed.). 2003: 453-469.
[37] Purcell G, Rennels G, Shortliffe E. Development and evaluation of a context-based document representation for searching the medical literature. International Journal on Digital Libraries. 1997; 1(3): 288-296.
[38] Duboue P and McKeown K. Empirically Estimating Order Constraints for Content Planning in Generation. Proceeding of the ACL/EACL. 2001: 172-179.
[39] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo. 2004: 565-72.

[40] Hsieh Y, Hardardottir GA, Brennan PF. Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses. Medinfo. 2004: 511-5.
[41] Barzilay R, Elhadad N. Sentence alignment for monolingual comparable corpora. EMNLP, Sapporo, Japan. 2003.
[42] Ishikawa K, Ando S, Okumura A. Hybrid text summarization method based on the TF method and the Lead method. Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. Tokyo, Japan. 2001.
[43] Teufel S, Moens M. Sentence extraction as a classification task. Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization. 1997: 58-65.
[44] The American Heritage® Dictionary of the English Language, Fourth Edition Copyright © 2000 by Houghton Mifflin Company.
[45] 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.
[46] Edmundson HP. New methods in automatic extracting. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 23-42.
[47] McKeown K, Radeev DR. Generating summaries of multiple news articles. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 381-390.
[48] 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.
[49] Barzilay R, Elhadad N, McKeown K. Sentence ordering in multidocument summarization. Proc. of HLT. San Diego, CA. 2001.
[50] Aone, Chinatsu, Okurowski ME, Gorlinsky J. Trainable, scalable summarization using robust NLP and machine learning. ACL/EACL Workshop on Intelligent and Scalable Text Representation. Madrid, Spain. 1997.
[51] Barzilay R, Elhadad M. Using lexical chains for text summarization. Proceedings of the Intelligent Scalable Text Summarization Workshop, 1997.
[52] Marcu D. Discourse trees are good indicators of importance in text. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 123-136.
[53] 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.
[54] 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.
[55] Mann W, Thompson S. Rhetorical structure theory: Toward a functional theory of text organization. Text. 1988; 8(3): 243-281.
[56] Marcu D, A Echihabi. An unsupervised approach to recognizing discourse relations. Proceedings of the ACL/NAACL. 2002.
[57] Norbert Reithinger, Robust Information Extraction in a Speech Translation System. Proceedings of EuroSpeech-99. 1999: 2427-2430.

[58] Hahn U. Reimer U. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 215-232.
[59] Valenza R, Robinson T, Hickey M, Tucker R. Summarization of spoken audio through information extraction. Proceedings of the ESCA Workshop. Cambridge, UK. 1999: 111-116.
[60] Kameyama M, Megumi, Kawai G, Arima I. A real-time system for summarizing human-human spontaneous spoken dialogues. Proc ICSLP. 1996: 681-684.
[61] Gurevych I, Strube M. Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland. August 2004: 764-770.
[62] Carbonell J, Geng Y, Goldstein J. Automated query-relevant summarization and diversity-based reranking. IJCAI-97 Workshop on AI and Digital Libraries. 1997.
[63] Zechner K. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. SIGIR. New Orleans, LA. September 2001: 199-207.
[64] Hearst M. TextTiling: A quantitative approach to discourse segmentation. Technical Report 93/24, U. of Californa, Berkeley. 1993. In: http://citeseer.ist.psu.edu/hearst93texttiling.html. (Last accessed July 19, 2005).
[65] Mani I. Summarization evaluation: An overview. Proceedings of the NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo National Institute of Informatics. 2001.
[66] Jones KS, Galliers JR. Evaluating natural language processing systems: an analysis and review. New York, Springer (eds). 1996.
[67] 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.
[68] Saggion H, Lapalme G. Concept identification and presentation in the context of technical text summarization. Proceedings of the Workshop on Automatic Summarization. 2000: 1-10.
[69] Hatzivassiloglou V, McKeown K. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. 1993: 172-182.
[70] Brandow R, Mitze K, Rau L. Automatic condensation of electronic publications by sentence selection. Information Processing and Management; 31(5): 675-685. Reprinted in: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999: 293-303.
[71] Mani I, Bloedorn E. Summarizing similarities and differences among related documents. Information Retrieval. 1999; 1: 35-67.
[72] Maybury M. Generating summaries from event data. Information Processing and Management. 1995; 31(5): 735-751.
[73] Morris A, Kasper G, Adams D. The effects and limitations of automatic text condensing on reading comprehension performance. Information Systems Research. 1992; 3(1): 17-35. Reprinted in: Advances in Automatic Text Summarization (eds: Mani and Maybury). 1999:305-323.
[74] QuickTap. Telephone Handset Tap. JKAudio, Inc. Sandwich, IL. 2000.

[75] Fox C. A stop list for general text. SIGIR Forum. 1990; 24(1-2): 19-35.
[76] Lacson R, Lacson E, Szolovits P. Discourse structure of medical dialogue. Proceedings of MEDINFO. 2004: 1703.
[77] Manning C and Schutze H. Clustering. In: Foundations of Statistical Natural Language Processing. The MIT Press. 2000: 495-528.
[78] Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proc. AMIA Symposium, 2001: 17-21.
[79] Ratnaparkhi A. A maximum entropy part-of-speech tagger. EMNLP Conference. 1996; 133-142.
[80] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-174.
[81] Schapire R, Singer Y. Boostexter: A boosting-based system for text categorization. Machine Learning. 2000; 39(2/3):135-168.
[82] Chapman W, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Medinfo. 2004: 487-491.
[83] Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proc. AMIA Symposium; 2001: 17-21.
[84] Miller S, Guinness J, Zamanian A. Name tagging with word clusters and discriminative training. HLT-NAACL. 2004: 337-342.
[85] Brown PF, Della Pietra VJ, DeSouza PV, Lai JC, Mercer RL. Class-based n-gram models of natural language. Computational Linguistics. 1990; 18(4): 467-479.
[86] Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? Medinfo. 2001;10(Pt 1):399-403.
[87] 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.
[88] McCallum, A. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.
[89] 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.
[90] Allan J, Gupta R, Khandelwal V. Temporal summaries of news topics. Proceedings of SIGIR. 2001: 10-18.
[91] 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.
[92] Lacson R, Lacson E, Szolovits P. Home Hemodialysis Queries. Proceedings of Medinfo; 2004(CD): 1702.
[93] 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.
[94] Shriberg E, Stolcke A, Hakkani-Tur D, Tur G. Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication. 2000; 32(1-2): 127-154.

[95] 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.
[96] Flammia G. Discourse segmentation of spoken dialogue: An empirical approach [PhD Thesis].Cambridge, MA:MIT;1998.
[97] 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.