# Automatic Processing of Spoken Dialogue in the Home Hemodialysis Domain

**Ronilda Lacson, MD, MS and Regina Barzilay, PhD**
**Massachusetts Institute of Technology (MIT), Cambridge, MA**

**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 ($p<0.01$). This work demonstrates the feasibility of automatically processing spoken medical dialogue.*

## I. Introduction

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

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

Figure 1: Transcribed segment of a phone dialogue

| | |
|---|---|
| (1) | Umm, I'm out of kidneys |
| (2) | out of kidneys, ok |
| (3) | give me a box of them |
| (4) | a box of them, ok, I'll pick up, I'll give you a box of them |
| (5) | ok |
| (6) | and I'll leave them in the room, do you know where the coolers are? |
| (7) | yeah |
| (8) | ok, I'll leave them in there with your name on it |
| (9) | ok |
| (10) | ok, how's the Vioxx helping your shoulder? |
| (11) | oh, now I haven't actually tried to do anything, I haven't lifted weights for 2 weeks |

In this paper, we describe a technique for automatically acquiring the structure of a transcribed medical dialogue by identifying the semantic type of its turns. 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 analysis. The Clinical category represents the patient's health, the Technical category encompasses problems with operating dialysis machines, the Miscellaneous category includes mostly scheduling and social concerns, while Backchannels capture greetings and acknowledgments. This classification allows a provider to distill the portions of the dialogue that support medical reasoning and are of primary interest to clinicians, as opposed to technical or scheduling concerns which are typically routed elsewhere. In the long run, knowing the distribution of patient requests can improve the allocation of resources, and ultimately provide better quality of health care.

We present a machine learning algorithm for semantic type classification for medical dialogue. 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. We use a manually

annotated set of transcripts for training the algorithm. A separate test corpus is used to evaluate its accuracy. The basic model presented in section IIC follows the traditional design of a dialogue act classifier:[3,4] it predicts the semantic type of an utterance based on a shallow meaning representation encoded as simple lexical and contextual features. The lack of world knowledge in this representation is compensated for by a large number of manually annotated training examples.

In Section IID, we show how to enhance our machine learning algorithm with background knowledge. Prior to text classification, we employ a feature generator that maps words of a transcript into semantic concepts augmenting our initial shallow utterance representation with new, more informative features. We explore two sources of background knowledge: a manually crafted, large-scale domain ontology, UMLS, and word clusters automatically computed from raw text using hierarchical distributional clustering. The experimental evaluation, described in Section III, confirms that adding semantic knowledge brings some improvement to dialogue turn classification.

The contributions of this paper are threefold. First, we propose a framework for rendering transcripts of patient-caregiver consultations into a structured representation, amenable to automatic processing. We show that the annotation scheme we propose can be reliably annotated by humans, and thus forms a solid basis for training the learning algorithm. Second, we present a fully-implemented machine-learning method that accurately identifies the semantic type of each utterance. Our emphasis on spoken medical discourse sets us apart from the efforts to interpret written medical text.[5-7] Third, we explore a novel way to automatically incorporate medical knowledge into a dialogue classification algorithm.

## II. Methods

We first introduce our methods for data collection and the annotation scheme. 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.

### A. *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. All patients and nurses whose questions and answers were recorded read and signed an informed consent form approved by the MIT Committee on the Use of Humans as Experimental Subjects. At the end of the study period, we received a total of six cassette tapes, consisting of 118 phone calls, containing 1,574 dialogue turns with 17, 384 words. The conversations were transcribed, maintaining delineations between calls and speaker turns.

Two domain experts, specializing in Internal Medicine and Nephrology, independently labeled each dialogue turn with its semantic type. The annotators were provided with written instructions that defined every category, and provided corresponding examples.[8] To validate the reliability of the annotation scheme, we computed agreement using the kappa coefficient.[9] Complete agreement would correspond to a kappa of 1.0; while kappa is 0.0 if there is no agreement. We computed the kappa to be 0.80, which is "near-perfect" agreement.[9] This kappa suggests that an automatic dialogue turn labeling might be feasible in an initial attempt to do medical dialogue representation.

The data were then divided into training and testing sets according to the chronological order in which they were received. The distribution of semantic types for each set is shown in Table I.

Table I. Semantic Type Distribution in Training and Testing Data Sets

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

### *B. Semantic Taxonomy*

Our annotation scheme was motivated by the nature of our application – analysis of phone consultations between nurses and dialysis patients. It consists of four semantic types – Clinical, Technical, Backchannel and Miscellaneous. Examples of utterances in each semantic type are shown in Table II.

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

Table II. Examples of dialogue for each semantic type

| |
|---|
| **Clinical**: You see, his pressure is dropping during his treatments.<br>Is he eating a lot? Did he gain weight? |
| **Technical**: The machine is stuck.<br>That's where you spike it; the second port is the one where you draw from. |
| **Backchannel**: Hello. How are you doing?<br>Yeah. |
| **Miscellaneous:** M wants me to remind you of your appointment today at 8:30.<br>I'm just helping out 'til they get back from vacation. |

### C. *Basic Model*

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

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

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

**Feature weighting and combination:** We learn the weights of the rules in the supervised framework using Boostexter,[10] 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.

### D. *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.

1. Our first approach builds on a large-scale human crafted resource, UMLS. This resource is widely used in medical informatics, and has been shown beneficial in a variety of applications.[7,11] The degree of generalization we can achieve is determined by the size and the structure of the ontology. 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 type. An example of such substitution is shown in the second row of Table III. To implement this approach, we first employ Ratnaparkhi's tagger to identify all the nouns in the transcript.[12] Then, using MetaMap, we extract the corresponding semantic type and replace the noun with the corresponding semantic type from the UMLS.[13] An utterance with the UMLS substitutions is added to the feature space of the basic model.

2. As a source of automatically computed background knowledge, our second approach uses clusters of words with similar semantic properties. An example of a cluster is shown in Figure 2 below. Being automatically constructed, clusters are noisier

than UMLS, but at the same time have several potential advantages. First, we can control the degree of abstraction by changing the desired number of clusters, while in UMLS the abstraction granularity is predetermined. Second, 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.[14,15]

Figure 2: An Example of a Cluster (1111000111110)

| | | |
|---|---|---|
| headaches | cramps | patience |
| cramping | pain | saline |
| radiation | itching | fluids |
| swelling | | |

To construct word classes, we employ a clustering algorithm that groups together words with similar distributional properties.[14] 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. Once the clustering process 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 III 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.

Table III. Data in various representations

| |
|---|
| **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. |

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 patients undergoing hemodialysis, available in the following url: http://health.groups.yahoo.com/group/dialysis_support. Our corpus contains more than 1 million words corresponding to discussions within a ten month period.

## III. Results

Table IV 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 that combines lexical, turn length and contextual features and is augmented with background information obtained through statistical clustering achieves an accuracy of 73%.

Table IV. Models and their Accuracy

| Models | Accuracy (n=1281) |
|---|---|
| Dialogue turn | 69% |
| Dialogue turn with length | 70% |
| Dialogue turn with previous turn | 68% |
| **Basic Model (Dialogue turn with length and previous turn)** | **70%** |
| **Basic Model + UMLS** | **71%** |
| **Basic Model + 1500 clusters** | **73%** |

Table V. Examples of predictive features

| Category | Current Dialogue Turn | Previous Dialogue Turn |
|---|---|---|
| Clinical | weight, blood, low, feel, feeling, pulse | weight, take Integer, you feeling |
| Technical | filter, box, leaking | Machine, a little |
| Backchannel | Thank, ok, and umm | Hi, make sure, lab |
| Miscellaneous | Appointment, hold, phone | Can, o clock, what time |

The first four rows of Table IV show the contribution of different features of the basic model. While the words of the dialogue turn alone are strong predictors of its type (Table IV, row 2), both length of the turn and the words of the previous utterance improve the performance of the basic model, thus achieving an accuracy of 71%. Table V shows the most predictive features for each category.

The last three rows in Table IV demonstrate that adding background knowledge improves the performance of the model. The model based on statistical clustering outperforms the basic model by 3%, compared to UMLS augmentation which improves the performance by 1%.

## IV. Discussion

We describe a method for automatically computing the semantic type of a dialogue turn in phone conversations between nurses and dialysis patients. The current model does not yield perfect performance. However, the output of the system still provides a meaningful and useful abstraction of medical dialogue, currently underutilized in the health care process. Several tasks such as resource allocation and data archiving could benefit from this technique, even at the current level of performance. To the best of our knowledge, the presented work is a first step towards automatic analysis of spoken medical dialogue. While this work focuses on medical dialogue related to a home hemodialysis program, the techniques we develop are sufficiently general and can be applied to analysis of medical dialogue in other applications.

The main contributions of this paper include the identification of features that characterize each semantic category and an implementation of a dialogue classification algorithm based on those features. We also show that augmenting the algorithm with background medical knowledge brings some performance improvement. While we hoped to gain more substantial improvement, our method is limited by the size of the corpus used for cluster induction. Typically, distributional clustering algorithms are trained on corpora with 100 million words.[15] Our corpus is an order of magnitude smaller because such amount of data is difficult to obtain in the dialysis domain. In the future, we will explore how non-medical text can be used to improve the quality of our clusters.

An interesting finding of this research is that noisy, statistically constructed clusters are more useful for our application than UMLS, a human-constructed source of medical knowledge. We explain this finding by the markedly different vocabulary used in written and spoken discourse. 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. 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.

In the future, we plan to extend this work in two main directions. First, we will apply our method to automatically recognized conversations. To maintain the classification accuracy, we will explore the use of acoustic features to compensate for recognition errors in the transcript. Second, we will refine our annotation scheme to include more semantic categories to support a deeper analysis of medical dialogue. We will experiment with more expressive statistical models for representing the sequential structure of medical dialogues.


### Acknowledgements

*The authors acknowledge the support of the National Science Foundation (Barzilay; CAREER grant IIS-0448168 and grant IIS-0415865). We also thank Dr. Robert Lockridge and the nurses at the Lynchburg Nephrology nightly home hemodialysis program, Mary Pipken, Viola Craft, and Maureen Spencer for their diligence in recording the conversations. We thank Percy Liang for letting us use his word clustering software; Dr. Eduardo Lacson for the annotations; and Mirella Lapata, Peter Szolovits, and three anonymous reviewers for their helpful comments and suggestions. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation.*


### *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 Dec; 13(12):16, 19, 23-6.
3. Gorin A. Processing of semantic information in fluently spoken language. Proceeding of Intl. Conf. on Spoken Language Processing (ICSLP). 1996; 2: 1001-1004.
4. Chu-Carroll J, Carpenter B. Vector-based natural language call routing. Computational Linguistics. 1999; 25(3): 361-388.
5. 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.
6. Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo. 2004:565-72.
7. Hsieh Y, Hardardottir GA, Brennan PF. Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses. Medinfo. 2004:511-5.
8. http://people.csail.mit.edu/rlacson/RequestAnnotation.doc.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159-174.
10. Schapire R, Singer Y. Boostexter: A boosting-based system for text categorization. Machine Learning. 2000; 39(2/3):135-168.
11. 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.
12. Ratnaparkhi A. A maximum entropy part-of-speech tagger. EMNLP Conference. 1996; 133-142.
13. Aronson A. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. Proc. AMIA Symposium, 2001: 17-21.
14. Brown PF, DeSouza PV, Mercer R, Della Pietra VJ, Lai JC. Class-based n-gram models of natural language. Computational Linguistics. 1992; 18:467-479.
15. Miller S, Guinness J, Zamanian A. Name tagging with word clusters and discriminative training. HLT-NAACL. 2004; 337-342.