DREAM: Classification Scheme for Dialog Acts in Clinical Research Query Mediation

Julia Hoxha[a], Praveen Chandar[1 a], Zhe He[a], James Cimino[b], David Hanauer[c], Chunhua Weng[a,d]

[a]*Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[b]*Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD, USA*
[c]*Department of Pediatrics, School of Information, University of Michigan, Ann Arbor, MI*
[d]*Corresponding Author, Department of Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20, New York, NY 10032, USA, Email: chunhua@columbia.edu,Tel: 212-305-3317*

Abstract

Clinical data access involves complex, but opaque communication between medical researchers and query analysts. Understanding such communication is indispensable for designing intelligent human-machine dialog systems that automate query formulation. This study investigates email communication and proposes a novel scheme for classifying dialog acts in clinical research query mediation. We analyzed 315 email messages exchanged in the communication for for 20 data requests obtained from three institutions. The messages were segmented into 1333 utterance units. Through a rigorous process, we developed a classification scheme and applied it for dialog act annotation of the extracted utterances. Evaluation results with high inter-annotator agreement demonstrate the reliability of this scheme. This dataset is used to contribute preliminary understanding of dialog acts distribution and conversation flow in this dialog space.

Keywords: clinical research query mediation dialog, dialog act modelling,

[1] Author is currently working at the company x.ai

dialog annotation scheme, human-human conversation analysis

1. Introduction

Leveraging the rich data in electronic health records (EHR) for clinical research, including cohort identification, is promising to accelerate clinical and translational research. However, this process remains difficult, expensive, and time-consuming [1] due to complex data representations and the black-box nature of most clinical databases. In order to translate data requests to executable data queries, medical researchers usually consult with query analysts through a series of email communications, phone calls, and face-to-face meetings. These conversations help to clarify researchers' data needs and formulate feasible and accurate data queries.

Understanding this communication is a necessary building block for the design of structured query negotiation. Clinical researchers would greatly benefit from a mixed-initiative dialogue system that enables human-machine collaboration for query formulation. We envision an intelligent conversational agent to act as a broker between clinical data and clinical researchers, while guiding them step by step through an effective and efficient query optimization. At this point, there exist no contemporary system or automated solutions to assist clinical researchers in their data quest. Furthermore, there is a wide research gap in studies of discourse structure or dialog acts used in query mediation, which are needed to provide useful characterization dialog behavior in human-to-human conversation and, potentially, human-computer dialog systems [2].

Our goal is to bridge this gap by shedding light on the communication involved during the query mediation process, especially with respect to the written communication. We aim to provide a classification scheme of acts involved in this type of dialogs. Dialog acts (DAs) are particularly important for modelling the intent underlying the utterance of each communication party [3, 4, 5].

This work makes two significant contributions. First, we develop an annotation scheme of Dialog acts in clinical REsearch dAta query Mediation (referred to as DREAM taxonomy), which is novel for the characterization of the discourse in this domain. We apply the resulting taxonomy of dialog acts to manually annotate utterances extracted from the email messages, generating a labeled dataset that can be used for innovative methods for automated DA learning. Second, this study contributes an analysis of dialog act distribution in the annotated dataset and knowledge on conversational flow in this dialog space.

2. Background

The query negotiation process between clinical researchers and query analysts, which aims to facilitate access of clinical data, remains complex [6] due in large to two main problems. On one side, it is difficult for query analysts to fully understand data requests because of their limited medical domain knowledge. On the other side, clinical researchers often lack the necessary technical expertise to comprehend and access clinical databases, which are usually characterized by an opaque internal implementation.

Recognizing the importance of the biomedical query mediation process and its inherent difficulties, previous research has investigated the communication modalities between query analysts and medical researchers. One way for clinical researchers to obtain data is by completing a data request form, which aims at conveying the complex data needs in an understandable manner. Hanauer et al. [7] conducted a content analysis of a collection of these forms from different institutions, identifying their over-emphasis on metadata that are not relevant to actual data needs. Their findings help to suggest recommendations on how to improve the quality of this particular means of communicating queries in support of clinical and translational research.

Hruby et al. [8, 9] conducted content analysis of phone conversations between researchers and analysts for a set of data requests. These dialogs were transcribed, annotated with dialog codes, and further analyzed with respect to the overall tasks (e.g. problem statement, clinical process description, design study) involved in the negotiation process. Analysis of the annotated conversations enable the illustration of the dialog progression and negotiation space, pointing to their complexity. The study identifies the significant effort needed to reach an understanding of researchers' data queries, confirming that negotiation is a difficult and iterative process.

While making valuable contributions and useful recommendations for the improvement of structured query negotiation, these works are focused on a higher level of the process and its tasks. The studies are based on different forms of capturing communications, such as the static data request forms and investigative interviews. They are not analyzing records of real, natural language conversations between researchers and analysts at a finer-grained level.

In order to design a framework for the characterization of query mediation dialogs founded on a theoretical basis, we adopt Speech Act Theory (SAT), whose development is credited to Austin [10] and Searle [11].

According to Speech Act Theory, spoken words alone do not have a simple fixed meaning in and of themselves, but their meaning is affected by the situation, the speaker and the listener. SAT distinguishes between the different aspects of dialog acts. The *locutionary act* represents the actual utterance and its ostensible semantic meaning. The *perlocutionary act* is the intended effect on the feelings, thoughts or actions of either the speaker or the listener. Unlike locutionary acts, perlocutionary acts are external to the performance, capturing the power to change minds. The *illocutionary act* is the speaker's intent embodied in the utterance, a true speech act e.g. questioning, informing, ordering, warning.

According to this framework, the agreed smallest unit of speech analysis is the *utterance*, regarded as a discernable segment of speech that conveys only one thought. The representation of an utterance can be captured in various ways in written communications.

The first step for modelling and automatic detection of discourse structure is the identification of dialog acts (DA) at the utterance level. The notion of a dialog act plays a significant role in SAT studies of dialog, particularly in the interpretation of participants' communicative behavior, creation of annotated dialog corpora, and the design of human-computer dialog systems [3, 12]. We adhere to the definition of dialog act provided by Austin [10], according to which a DA represents the meaning of an utterance at the level of illocutionary force. This is also known as the equivalent of the speech act of Searle [11].

Dialog act classification schemes have been a focus of research in linguistics aiming at standardizations of discourse structure annotation systems. In order to preserve as much comparability with well-established systems and previous research as possible, we use as foundation of our work the Dialogue Act Markup in Several Layers (DAMSL) tag set [12]. DAMSL is a rich, multi-layered annotation scheme for dialog acts that is domain- and task-independent.

In the medical domain, studies of communication and dialog acts at the utterance level have been deployed particularly in the analysis of doctor-patient encounters [13, 14, 15]. However, studies of discourse in biomedical query mediation, even though critical and widespread, are limited. Research on the classification of query mediation dialog acts are essential to guide the designs of intelligent query aids for medical researchers. In the rest of this paper, we present the proposed classification scheme and results of the experimental results, followed by the range of applications that can build upon such a classification scheme.

3. Materials and Methods

3.1. Dialog Domain

In our problem domain, a dialog is email communication between the query analyst and clinical researcher. The task of the analyst is to transform data requests into executable clinical database queries. A dialog is comprised of *turns*, in which a single speaker/writer has temporary control of the dialog and writes for some period of time. Each turn consists of an email message represented as unstructured text. Within a turn, the speaker may produce several typed utterance units, whose meaning at the level of illocutionary force is represented through a dialog act [10].

Definition 1. (Dialog Act) *A dialog act represents the meaning of an utterance at the level of illocutionary force.*

We adhere to the representation of dialog act with two components: *semantic content*, which specifies the objects, relations, actions, and events that the dialog act is about, and *communicative function*, which is a specification of the effect that the semantic content has on the addressee for updating his or her information state upon understanding the stretch of dialog [16].

The communication in this domain is task-oriented in nature: the data analyst helps the researcher to refine the query with the clinical research eligibility criteria, which specify the medical, demographic, or social characteristics of eligible research volunteers [6].

This study was approved by Columbia University Medical Center Institutional Review Board (study ID AAAJ8850).

3.2. Data Collection

We analyzed a collection of email messages exchanged in the communication between query experts and clinical researchers for 20 data requests. We refer to the sequence of messages exchanged for each data request as one conversation. The dataset consists of 315 English-language messages (153 from researchers, 162 from query analysts), with an average of 15.8 and standard deviation of 4.9 messages per conversation. Figure 1 illustrates the distribution of emails per conversation.

From the email metadata, we observe that the processing time of the data requests range from a few days (3-4 days) to several months, with one conversation (denoted as $conv_{17}$) taking place from *October 13, 2011* until *January 9, 2013*. The conversations involved team discussion with size ranging from 2 to 7 participants, with a median of 2 participants. The data requests were sampled from three different institutions (5 data requests from the first institution, 4 data requests from the second, and 11 data requests from the third institution).

The original emails were not consistently segmented linguistically; therefore we implemented a pipeline of parsing and sentence-level segmentation techniques. The unit of this segmentation is the utterance. Sentence segmentation is a research challenge on its own. In our approach, we apply a technique based on regular ex-

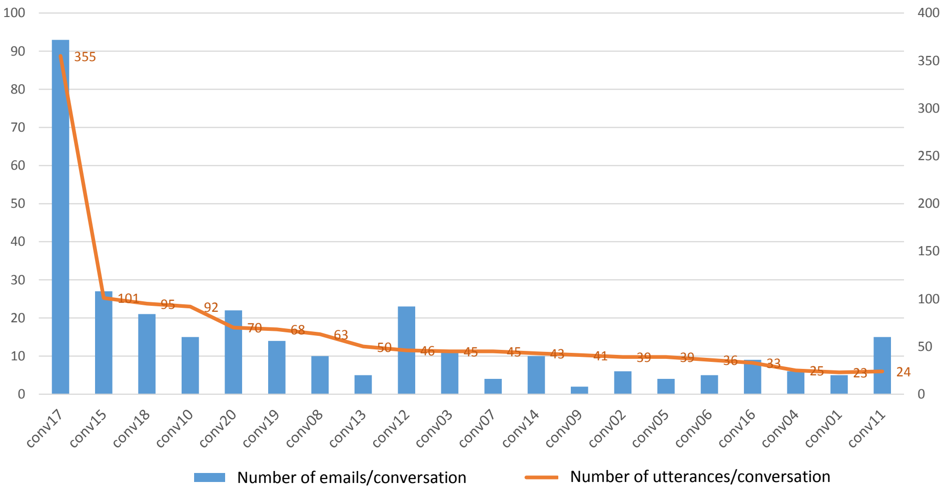

Figure 1: Distribution of email messages and utterances in conversations.

pressions detecting punctuation marks that denote the end of a sentence. Phrases separated by punctuation marks denoting sentence continuation are joined in one segment. As a result of the segmentation process, we identified 1,333 utterances.

We observed that $conv_{17}$ is the longest one in duration, lasting for nearly 15 months, and contains the highest number of utterances (Fig. 1). We manually checked the content of the conversation, observing that this is a complicated data request about an HIV testing project. There is an iterative pattern occurring: the analyst runs a program and sends a set of results, then the researcher identifies problems and sends clarifications, which leads the analysts to re-run the program again and submit other results. Since the utterances of this conversation comprise 27% of the dataset (Figure 1), we considered it as an outlier and excluded its utterances from the annotation. However, this data request manifests the challenges and complexity of the query mediation process.

Table 1 provides an excerpt from one of the annotated conversations for illus-

Turn	Party	Utterance	Dialog Act
t_1	A	(U1.1) *Hi C7: I am serving your request 0387.*	Salutation; Statement-not-opinion
t_1	A	(U1.2) *Please make sure the list of diagnosis you want.*	Medical Condition; Action Directive
t_2	R	(U2.1) *We need to discuss your questions with you.*	Resolution Pending; Statement-opinion
t_2	R	(U2.2) *We did not completely understand your email.*	Resolution Pending; Statement-opinion; Signal-non-understanding
t_2	R	(U2.3) *Do you have any time tomorrow or Friday to discuss?*	Meeting Schedule; Information Request
t_3	A	(U3.1) *Hello C7, I am available most of this week, the week of the 1st.*	Meeting Schedule; Open Option
t_4	R	(U4.1) *We do not need the MRN we only need Age, Race, Gender, and Zip Codes.*	Demographics; Action Directive
t_5	A	(U5.1) *Attached is the result of request 0387.*	Result Submission; Statement-not-opinion
t_6	R	(U6.1) *Can you also identify the patients of Hispanic origin (Black Hispanic and Hispanic or Latino)?*	Demographics; Yes-no-question; Request Clarification
t_6	A	(U6.1) *Attached is the result according to your clarification.*	Result Submission; Statement-not-opinion
t_7	R	(U7.1) *I will analyze it and get back to you soon.*	Resolution Pending; Commit

Table 1: Fragment of an annotated conversation between a Query Analyst (A) and a Reseacher (R)

trative purposes. For each turn, it shows the communication party, i.e. Analyst (A) or Researcher (R), and the utterances issued by that party. Each utterance is assigned one or more DA labels (shown in column 4) from the DREAM classification scheme proposed in this work.

Hence, DAs can be considered as a *tag* set, which classifies utterances based on a combination of semantic content-based type (e.g. Medical Condition; Patient Demographics; Meeting Schedule) and their communicative function (i.e. if the utterance is a statement expressing opinion, question, request) in the dialog. In the next section, we explain the dialog acts assigned to the utterances in Table 1, as well as the other acts pertaining to the proposed DREAM classification scheme.

3.3. DREAM Taxonomy

To preserve comparability with existing systems, we extended a well-known standard for discourse structure annotation, the Dialogue Act Markup in Several

Layers (DAMSL) tag set [5]. We refined our selection of specific tags after consulting other cases of DAMLS extension for dialogs [17, 18]. We further extended the taxonomy with relevant information on the domain of discourse, increasing its granularity with respect to content characterization tags. We particularly augmented the task-specific acts for semantic categorization of the patient's characteristics in the written query.

A distinguishing feature inherited from DAMSL is allowing multiple tags to be applied to an utterance. The rationale is that a particular utterance might simultaneously serve the purposes of responding to a question, confirming understanding, promising to perform an action, or giving information. For each utterance, the annotation involves making choices along the following four dimensions, each one describing a different orthogonal aspect:

- **Communicative Status** - defines whether the utterance is interpretable and belonging to the domain of discourse.
- **Information Level** - characterizes the semantic content of the utterance.
- **Forward-looking Function** - encodes how the current utterance constrains the future actions of the participants and affects the discourse.
- **Backward-looking Function** - captures how the current utterance relates to the previous units of discourse.

3.3.1. Dimension Communicative Status

The dimension Communicative Status defines the cases in which an utterance has no effect on the dialog, because it is either distorted beyond recognition or does not pertain to the domain of discourse. This dimension is composed of two dialog acts: Uninterpretable and Miscellaneous. The tag Uninterpretable defines an utterance that is not intelligible due to bad grammar, typing, or semantically ill-formed presentation. In this dimension, we introduce the new tag Miscellaneous to encode utterances that do not fall in the domain of discourse, e.g. *"-Dashboard info sent"*, or *"Begin forwarded message"*. When none of the two tags qualifies, the utterance is considered interpretable and is assigned the appropriate dialog acts from the other dimensions.

3.3.2. Dimension Information Level

The dimension Information Level provides a characterization of the content and semantics of the utterance. The acts in this dimension are illustrated in Figure 2. In addition to the abstract classification of whether the utterance deals with the description of the specific Task (in our case this is Cohort Identification), management process of how to solve the task (Task Management), or the communication process (Communication Management), we augment this dimension with additional task-oriented tags.

We subdivide the category Task into acts that describe the specification of the query with respect to the Patient Characteristics. Furthermore, Data Source denotes utterances related to sources or warehouses of clinical data (e.g. *"BTRIS has loads of data"*). Another act added to the taxonomy under Task is Data Format, which describes the format/template of exchanging data. An example of an utterance annotated with this act is *"If we could parse the data by month"*. Result Submission is another act under Task that signals sending or obtaining results, namely the retrieved data.

An important part of the dialog in this domain involves discussion on Patient Characteristics. We particularly focus on this aspect of the task and provide a

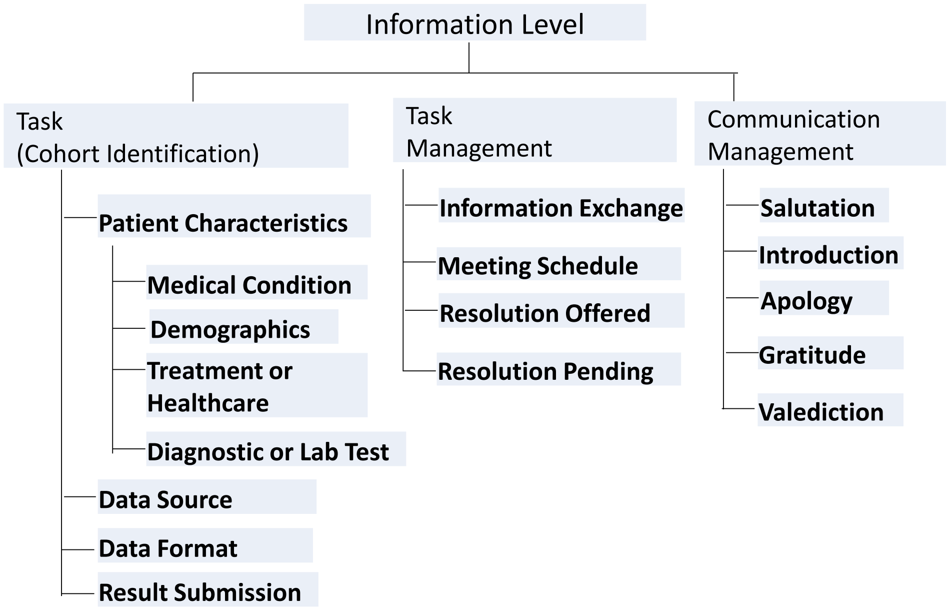

Figure 2: Dimension Information Level in of DREAM-taxonomy, the annotation scheme of dialog acts in clinical research query mediation. Bold-faced are the new acts, not included in DAMSL.

fine-grained specification by extending it with additional dialog acts. This augmentation is based on the groups of most frequent semantic classes of clinical research eligibility criteria that define patient characteristics, as proposed in [19]. The new dialog acts encode the semantics of utterances expressing characteristics on the patient's Medical Condition, Demographics, Treatment or Healthcare, Diagnostic or Lab Test.

Table 1 shows examples of utterances assigned to Medical Condition. Furthermore, the act Treatment describes utterances with information on the procedure, medication, or therapy. When an utterance describes several criteria simultaneously, we summarize the annotation with the tag Patient Characteristics.

An important category that describes utterances involving the coordination of participants' activities is Task Management. It is essential to distinguish it from the Task category. While Task addresses the specific aspects of the data request, the category Task Management captures utterances that deal with the overall process of solving the problem and talk about coordination activities. We have extended the Task Management category with dialog acts that are specific to the conversation domain. It is divided into acts that encode the means and characteristics of Information Exchange (e.g. *"Do you have an encrypted USB?"*), and discussion related to Meeting Schedule such as ask/propose/confirm a meeting. Furthermore, it encodes utterances where there is Resolution Offered by the speaker (e.g. *"I will be running your request now"*), and utterances describing Resolution Pending (e.g. *"We expect to have follow-on queries"*). In the conversation fragment of Table 1, we show examples of utterances annotated with Meeting Schedule, e.g. (U2.3), and utterances that show a Resolution Pending, e.g. (U2.1)*"We need to discuss your questions with you."*.

The third main category in this dimension is Communication Management. The acts under this category do not make a direct contribution to solving the task, but rather address social behavior in conversation. The category is augmented with dialog acts, which describe conventional utterances that maintain the communication process: Salutation, Introduction, Apology, Gratitude, Valediction.

3.3.3. Dimension Forward-looking Function

A very interesting aspect inherited from DAMSL is the use of two complementary dimensions: Forward-looking Function, which includes traditional speech acts (statements, directives, requests, etc.), and Backward-looking Function that indicates how the current utterance relates to the previous discourse to signal agreement, understanding, or provide answers.

The dimension Forward-looking Function characterizes the effect of an utterance on the subsequent dialog (Figure 3). It defines whether, as a result of the utterance, the speaker is making a claim, or committing to certain beliefs or particular future actions. Forward-looking functions are divided into three categories: Representative, Directive, and Commissive. Representative, also referred to as Statements, are utterances that make claims about the world, whose content can be evaluated as true or false.

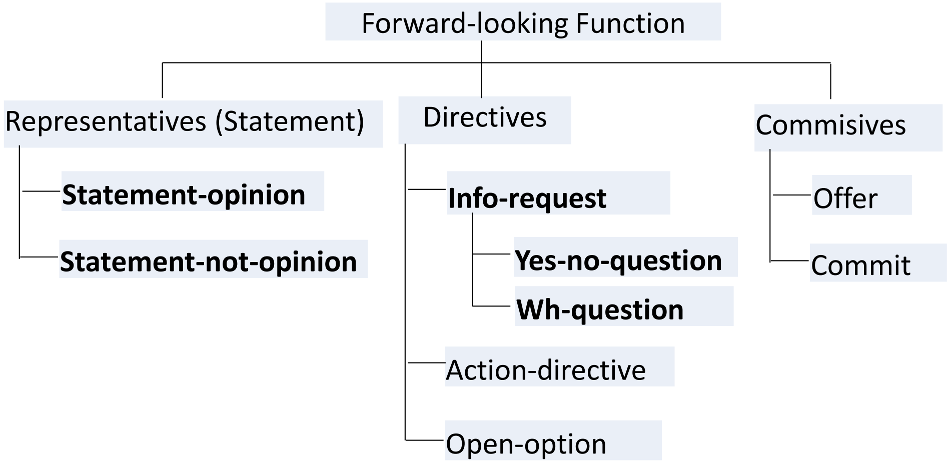

Figure 3: Dimension Forward-looking Function in DREAM-taxonomy. The extended acts are bold-faced.

We have extended this category with two dialog acts: Statement-non-opinion and Statement-opinion. In the conversation fragment in Table 1, we observe several utterances annotated as factual statements, such as (U6.1)*"Attached is the result according to your clarification."*, and statements expressing opinion e.g. (U6.1)*"We did not completely understand your email."*.

The category Directive, also referred to as Influencing-addressee-future-action, aims to classify those utterances that affect the listener's actions, as in the case of requests. Our annotation scheme makes the distinction between Information Request, which characterizes a form of obligation to provide an answer, and Action Directive that requires the addressee to either perform the requested action or communicate a refusal to perform the action. Table 1 shows as part of the conversation fragment an example of an utterance (U6.1)*"Please make sure the list of diagnosis you want."*, which expresses a directive for action. Note how this utterance in the Information Level dimension is assigned to Medical Condition, whereas in the dimension of Forward-looking Function it qualifies as Action Directive. Another important act is the Open Option, which suggests a course of action, but makes no obligation on the addressee. An example utterance of open option from the fragment is (U3.1) *"Hello C7, I am available most of this week, the week of the 1st."*, which content-wise is also assigned to Meeting Schedule.

The category Commissives, also referred to as Committing-speaker-future-action, encodes utterances that potentially engage the speaker to some future course of action. Within this aspect, we keep the distinction as to whether the commitment is conditional on the listener's agreement or not (Offer), or the typical case of a promise (Commit). Example of a promise is shown in Table 1 with the utterance (U7.1) *"I will analyze it and get back to you soon."*

3.3.4. Dimension Backward-looking Function

The dimension Backward-looking Function indicates how the utterance responds to a previous dialog act. For instance giving an answer, accept, reject, or trying to correct some previous utterance (referred to as antecedent). This dimension is illustrated in Figure 4. It is composed of three main categories: Agreement, Understanding, Answer, and Information Relation.

The acts under the category Agreement are kept the same as in the DAMSL scheme. The category Understanding also inherits the two dialog acts from

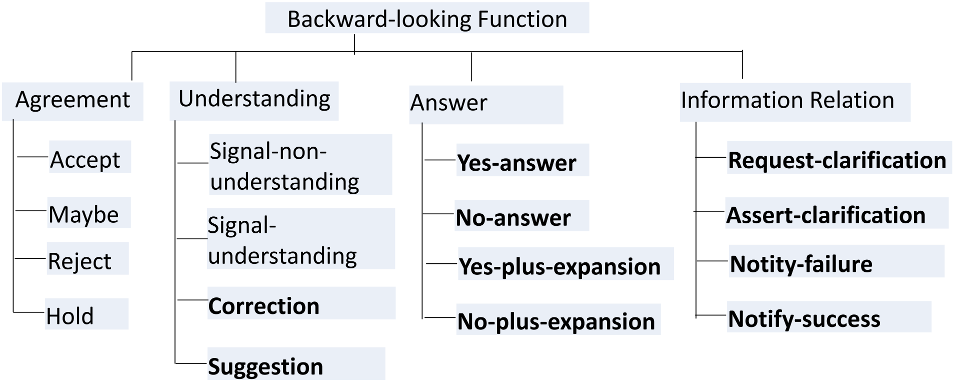

Figure 4: Dimension Backward-looking Function in DREAM-taxonomy. The extended acts are bold-faced.

DAMSL: Signal-non-understanding and Signal-understanding. The former describes utterances, which indicates explicitly a problem in understanding the antecedent (e.g. (U2.2) *"We did not completely understand your email."* in Table 1. Signal-understanding describes utterances that express comprehension of utterances in previous turns of the dialog.

However, the category Understanding has an important augmentation with the acts Correction and Suggestion. These acts were added as part of the iterative schema refinement process during dataset annotation. Furthermore, the category Answer has also been specialized with acts that distinguish different forms of answering.

The highlight in this dimension is the extension with acts under Information Relation, which captures how the content of the current utterance relates to antecedents in the discourse. This category was not elaborated in DAMSL. We have extended it with tags that are semantically relevant to the cohort identification task.

As such, Request Clarification indicates that the speaker is making a request for clarification related to a previous utterance about the task. For example, the utterance (U6.1) *"Can you also identify the patients of Hispanic origin (Black Hispanic and Hispanic or Latino)?"* of the conversation fragment shown in Table 1.

The act Assert Clarification describes utterances that make clarifying statements with respect to a previous request for task clarification. Furthermore, the acts Notify Failure and Notify Success encode statements that describe, respectively, failure or success of the task.

3.4. Dialog Act Annotation Process

Hand-labeling of the extracted utterances involved three annotators (authors JH, PC, ZH). All annotators conduct research in the field of query refinement for enabling data access for clinical researchers. The annotators were trained through in-person meetings and a written annotation manual.

Each utterance could have multiple tags in this aspect depending on how many functions it simultaneously performs. The annotation process itself consisted of four main steps that an annotator had to follow for each utterance:

1. Check along the first dimension Communicative Status if the utterance is Uninterpretable or Miscellaneous.If yes, mark it with one of the two acts and move to the next utterance.

2. If the utterance is interpretable, check its content and label it with a tag from the dimension Information Level.

3. Considering the effect that the utterance imposes on the receiver and the future actions, assign to it the best fitting tag from the dimension Forward-looking Function.

4. Considering how the utterance relates to a previous turn in the dialog, locate its antecedent and then assign to it a tag the best fitting tag from the dimension Backward-looking Function.

In the first phase, we conducted a pilot study on the set of utterances extracted from two conversations. The pilot study identified important aspects for extension of the annotation manual, enforcement of particular annotation guidelines to increase inter-annotator tagging consistency, and suggestion of new dialog acts (e.g. Correction and Suggestion). The taxonomy was refined accordingly, resulting in the current version of the DREAM-taxonomy proposed in this work.

The second phase comprised the annotation of the entire dataset from each annotator independently. The selection of a tag for each dimension is done with majority agreement. The final phase consisted of disagreement resolution for the cases when there were conflicts in all four dimensions.

We applied the Delphi method [20], which assures anonymity of annotations and reduced bias. After each annotator had independently completed the annotation task, we performed one round of the Delphi method on the utterances where there was no agreement in any of the four dimensions. In Delphi, experts checked their labels along with those of other experts in an anonymous fashion. They then had the option to change their annotation in light of other annotators' labels.

At the end of this process, we were able to have a fully annotated set of utterances, which allowed us to summarize the unstructured text of email messages in the conversations into a sequential set of DA triples. For example, as shown in the fragment of conversation of Table 1, the utterance *(U2.2) "We did not completely understand your email"* is represented with the tags Resolution-Pending, Statement-opinion, and Signal-non-understanding. The utterance *(U2.2) "Can* *you also identify the patients of Hispanic origin (Black Hispanic and Hispanic or Latino)?"* is encoded with tags $\mathrm{Demographics}$, $\mathrm{Yes\text{-}no\text{-}question}$, and $\mathrm{Request\text{-}Clarification}$.

4. Evaluation and Results

In order to evaluate various aspects of the proposed scheme, we performed four different experiments: 1) reliability study based on majority agreement, 2) reliability study with kappa statistics, 3) dialog act distribution analysis, and 4) conversation analysis with transition graph. In the following sections, we describe the experimental setup and results of the evaluation.

4.1. Reliability with Majority Agreement

A key requirement for an annotation scheme is that it can be used reliably by trained annotators. To assess this requirement, we performed experiments that measure agreement among the annotators on the basis of majority vote.

Experimental Setup and Metrics. As explained in section 3.4, the dataset of utterances obtained after the segmentation of conversations is coded by three annotators with the dialog acts proposed in DREAM. The total number of annotated utterances is 978 (after having excluded one conversation as explained earlier). There are 523 utterances provided by the agents in the role of Query Analyst, and 455 by the agents in the capacity of Researcher. Initially, we calculate majority agreement in the following way.

Definition 2. (Majority-based Agreement) *Given the set of annotated utterances $\mathcal{U}$, s.t. each $u_i \in \mathcal{U}$ is annotated with 3 labels (i.e. dialog acts from DREAM scheme) from 3 independent annotators, we define $\mathcal{U}_m \subset \mathcal{U}$ as the set of utterances with majority-based agreement, s.t. each utterance fulfills the following criteria of annotations in at least one dimension:*

C.1- more than half of the annotations per dimension are the same.

C.2 - otherwise, more than half of the annotations per dimension belong to the same parents in the scheme.

Based on Definition 2, we calculate the majority agreement score P_m as:

$$P_m = \frac{|\mathcal{U}_m|}{|\mathcal{U}|} \quad (1)$$

Hence, the score P_m measures the frequency of utterances belonging to the set $\mathcal{U}_m$. The formation of this set is done as follows: for each utterance, we look at the labels provided by the annotators for each dimension. We then select one label for that dimension based on majority agreement, i.e. if more than half of the annotations match (C.1 in Definition 2). We perform this check and label selection for each of the four dimensions. If the criterion is fulfilled for *at least one* dimension, the utterance is added to set $\mathcal{U}_m$.

For cases with no majority agreement of annotations in one dimension, we roll one level up in the hierarchy and check if labels belong to same parent (C.2 in Definition 2). We execute this step for each dimension. Again, if the criterion is fulfilled for at least one dimension, the utterance is added to set $\mathcal{U}_m$. We apply the second criterion C.2 motivated by the high degree of granularity provided by the scheme.

Results. The annotation results of this reliability experiment with respect to majority agreement are illustrated in Table 2. We report the agreement values before resolution and after resolution, i.e. after the annotators saw other annotations and changed theirs.

We observe very high agreement in the classification of utterances to tags in at least one dimension, precisely 91.1% before resolution and 100% post resolution.

Unit	Quantity	
Total Num of annotated utterances	978	
Number of utterances from Analysts	523	
Number of utterances from Researcher	455	
	Pre-Delphi	Post-Delphi
P_m in at least one dimension	891 91.1%	978 100%
P_m in all four dimensions	650 66.4%	757 77.4%

Table 2: Results of utterance annotation with majority agreement before and after resolution.

This means that the three annotators have tagged 91.1% of the utterances with the same dialog act in one or more of the four available dimensions. After applying the Delphi method, at post resolution all the utterances have been consistently annotated with the same label from the three annotators in at least one dimensions.

We also calculate majority agreement not only in at least one dimension, but also in all four dimensions. We observe a P_m score of 66.6% before resolution, reaching a high agreement of 77.4% after applying Delphi resolution.

4.2. Reliability with Kappa Statistics

Experimental Setup and Metrics. In this experiment, we use the annotations described in Table 2 and estimate inter-annotator reliability with the well-known statistical metrics of pairwise agreement and kappa. We apply Fleiss kappa as a statistical measure used to evaluate concordance or agreements between multiple annotators [21]. This measure is interpreted as expressing the extent to which the observed amount of agreement among annotators (pairwise agreement P_a) exceeds what would be expected if all annotators made their choices completely randomly (expected agreement P_e). Fleiss kappa κ is defined as:

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{2}$$

Kappa ranges between -1 and 1, where higher values denote better agreement. The factor $1 - P_e$ expresses the degree of agreement attainable above chance. Whereas, $P_a - P_e$ gives the degree of agreement actually achieved above chance. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators (other than what would be expected by chance) then $\kappa \leq 0$.

In order to judge the absolute values of kappa, a few guidelines have been introduced in the literature. However, they are not universally accepted because the criteria of interpreting kappa depend on the inherent difficulty of the task. To enable a general interpretation of the agreement level, we follow the guidelines of Landis and Koch [22]. We provide an additional discussion on the sensitivity of these values to the annotation task at hand in Section 5.

Initially, we calculate pairwise agreement and kappa for annotations at the DA dimension level. This means that for each utterance, we take the tag provided by each annotator and map it to its parent in the taxonomy corresponding to one of the four dimensions, i.e. $\mathrm{Communicative\ Status}$ (CS), $\mathrm{Information\ Level}$ (IL), $\mathrm{Forward\text{-}looking\ Function}$ (FLF), and $\mathrm{Backward\text{-}looking\ Function}$ (BLF). The evaluation measures are estimated for the annotations mapped at this level.

Results. The results of this reliability experiment are illustrated in Table 3.

DA Dimension	P_a	P_e	**Kappa**
Communicative Status (CS)	0.98	0.97	0.14
Information Level (IL)	0.76	0.38	0.61
Forward-looking Function (FLF)	0.88	0.66	0.64
Backward-looking Function (BLF)	0.5	0.35	0.22

Table 3: Reliability of annotations for the dimensions in DREAM-taxonomy.

There is high observed pairwise agreement P_a for CS, IL, and FLF (0.98, 0.76, and 0.88, respectively). We note lower agreement (0.5) between annotators on BLF tags. In terms of kappa, there is substantial agreement (above 0.6) for IL and FLF, and slight to fair agreement for CS and BLF.

The dimension of $\mathrm{Information\ Level}$ (IL) plays a significant role in defining the semantics of the utterances with respect to the specification of a clinical trial data request. Therefore, in the second step of this experiment, we stratify the analysis of inter-annotator reliability for the categories Task, $\mathrm{Task\ Management}$, and $\mathrm{Communication\ Management}$ under this dimension. We illustrate the results of this experiment in Table 4. We observe high pairwise agreement for the three categories under IL, particularly for the category $\mathrm{Communication\ Management}$ where P_a is 0.93. The annotations in this category are also characterized by a high kappa agreement (0.81).

DA Category	P_a	P_e	**Kappa**
Task	0.79	0.64	0.40
Task-management	0.83	0.71	0.41
Communication-management	0.93	0.63	0.81

Table 4: Reliability for Information Level dimension.

For the category Task and $\mathrm{Task\ Management}$, we observe moderate reliability values of kappa ($\geq$ 0.4). We also note a high value of expected agreement P_e, especially for the category $\mathrm{Task\ Management}$. In the discussion section, we provide a more detailed analysis on the impact of P_e in these values of kappa. Both categories reach high pairwise agreement, respectively 0.79 for Task and 0.83 for $\mathrm{Task\ Management}$.

4.3. Dialog Act Distribution Analysis

Experimental Setup and Metrics. We report on the *frequency distribution* of the most frequent dialog acts in the dataset of annotated utterances. We also report on statistical significance values p using Chi-Square test [23] with one degree of freedom and 0.05 level of significance.

Results. As illustrated in Table 5, the most frequent act in utterances of both research and analysts is Patient Characteristics, occurring in 17.7% of the utterances.

Dialog Act	**Freq (R+A)**	**Freq (R)**	**Freq (A)**
Task	92 (9.4%)	34 (7.5%)	58 (11.1%)
Patient Characteristics	173 (17.7%)	96 (21.1%)	77 (14.7%)
Medical Condition	69 (7.1%)	28 (6.2%)	41 (7.8%)
Patients-Demographics	35 (3.6%)	16 (3.5%)	19(3.6%)
Result Submission	38 (3.9%)	10 (2.2%)	28 (5.4%)
Task Management	153 (15.6%)	62 (13.6%)	91 (17.4%)
Action Directive	107 (10.9%)	74 (16.3%)	33 (6.31%)
Meeting Schedule	56 (5.7%)	33 (7.3%)	23 (4.4%)
Gratitude	93 (9.5%)	69 (15.2%)	24 (4.6%)
Yes-no-question	90 (9.2%)	43 (9.5%)	47 (8.9%)
Info-Request	67 (6.9%)	31 (6.8%)	36 (6.9%)
Statement-not-opinion	90 (9.2%)	26 (5.7%)	64 (12.2%)
Statement-opinion	50 (5.1%)	26 (5.7%)	24 (4.6%)
Request Clarification	53 (5.4%)	12 (2.6%)	41 (7.8%)
Assert Clarification	48 (4.9%)	34 (7.5%)	14 (2.7%)

Table 5: Distribution of most frequent dialog acts, ordered by dimension and frequency, in the utterances of Researcher (R), Query Analyst (A), or both (R+A).

We also analyzed the frequency of utterances annotated with dialog acts in the category Patient Characteristics. The majority of utterances express characteristics related to Medical Condition (7.1%), Demographics (3.6%), Laboratory Tests (1.8%), and Treatment (1.2%). The last two acts are not shown in Table 5, which illustrates the most frequent acts with $Freq(R + A)$ values above 3.5%.

Results in Table 5 further help to identify differences in communication style between researchers and analysts. In comparison to researchers, analysts express more utterances related generally to Task (11.1% vs. 7.5%, p=0.0123) and $\mathrm{Task\ Management}$ (17.4% vs. 13.6%, p=0.019). However, researchers carry more conversation about patient characteristics (21.1% vs. 14.7% for analysts, p=0.148). As can be expected, analysts express more utterances related to result submission (5.4% vs. 2.2% for researchers, p=0.0035).

With respect to the style of the communication, researchers give many more directives (16.3 % vs. 6.3% for analysts, p<0.001). However, both parties ask questions and make requests with similar frequencies. Another interesting observation, is the high frequency of factual information (statement-non-opinion) expressed by the analysts (12.2% vs. 5.7% for researchers, p<0.001).

We also observe a difference in the communication style with respect to expressing clarifications. Analysts tend to request more clarifications (7.8% vs. 2.6%, p<0.001), whereas researchers are the ones to assert clarifications more often (7.5% vs. 2.7% for analysts, p=0.0038). The low p-values indicate that these differences are statistically significant.

4.3.1. Cross-site Conversation Analysis

We further stratified our conversation analysis by comparing communication patterns among the three different institutions (referred here as *sites*) where the emails were originally gathered. The goal is to observe how the model used for annotating such dialogs is sufficiently accurate to reveal insights in interactions and help us draw comparisons among sites.

We look at the frequency distribution of dialog acts in the utterances of each site separately. Overall, the distribution of emails and annotated utterances for each site is the followings: s_1 has 32 emails and 171 utterances, s_2 has 17 emails and 140 utterances, and site s_3 has 161 emails and 622 utterances. We illustrate the most frequent acts in Table 6. For each site, we have highlighted the frequencies of those labels that particularly differentiate the communication behavior.

First of all, we observe that in all the three sites, there is similar frequency of $\mathrm{Patient\ Characteristics}$ and $\mathrm{Demographic}$ acts. This once again confirms the task-oriented nature of the conversation, focused on specification of the eligibility criteria for clinical trial patient recruitment.

However, there are differences observed after a comparative analysis of these sites. There is low frequency of $\mathrm{Medical\ Condition}$ in s_1, but higher frequency in s_2. At both s_1 and for s_2 there is high frequency of $\mathrm{Task\ Management}$ itself rather than its children nodes. This is explained by a higher ambiguity in this site's conversations, as such we have merged the agreed annotation at the $\mathrm{Task\ Management}$ parent node. This reflects higher difficulty of annotators to classify task management tags in s_2.

Meanwhile, site s_1 tends to assign more meeting schedules, rather than discuss about task in email. Site s_2 has very low frequency of $\mathrm{Meeting\ Schedule}$ acts. This is an interesting observation, since from experience we know that at certain institutions the analysts avoid personal meetings, and query mediation is solely conducted online.

Dialog Act	Site s_1			Site s_2			Site s_3		
	Freq (R+A)	**Freq (R)**	**Freq (A)**	**Freq (R+A)**	**Freq (R)**	**Freq (A)**	**Freq (R+A)**	**Freq (R)**	**Freq (A)**
Task	15(8.8%)	3(4.2%)	12(12.1%)	26(14.1%)	10(14.5%)	16(13.8%)	48(7.7%)	20(6.4%)	28(9.1%)
Patients Characteristics	26(15.2%)	10(13.9%)	16(16.2%)	33(17.8%)	14(20.3%)	19(16.4%)	114(18.3%)	72(22.9%)	42(13.6%)
Demographics	5(2.9%)	2(2.8%)	3(3%)	5(2.7%)	1(1.4%)	4(3.4%)	25(4%)	13(4.1%)	12(3.9%)
Medical Condition	3(1.8%)	-	3(3%)	19(10.3%)	8(11.6%)	11(9.5%)	47(7.6%)	20(6.4%)	27(8.8%)
Result Submission	1(0.6%)	-	1(1%)	2(1.1%)	-	2(1.7%)	35(5.6%)	10(3.2%)	25(8.1%)
Task Management	44(25.7%)	15(20.8%)	29(29.3%)	33(17.8%)	10(14.5%)	23(19.8%)	76(12.2%)	37(11.8%)	39(12.7%)
Meeting Schedule	13(7.6%)	6(8.3%)	7(7.1%)	1(0.5%)	1(1.4%)	-	42(6.8%)	26(8.3%)	16(5.2%)
Gratitude	17(9.9%)	15(20.8%)	2(2%)	10(5.4%)	6(8.7%)	4(3.4%)	66(10.6%)	48(15.3%)	18(5.8%)
Action Directive	19(11.1%)	14(19.4%)	5(5.1%)	8(4.3%)	4(5.8%)	4(3.4%)	80(12.9%)	56(17.8%)	24(7.8%)
Yes-no-question	12(7%)	5(6.9%)	7(7.1%)	30(16.2%)	12(17.4%)	18(15.5%)	48(7.7%)	26(8.3%)	22(7.1%)
Wh-question	7(4.1%)	4(5.6%)	3(3%)	1(0.5%)	-	1(0.9%)	4(0.6%)	3(1%)	1(0.3%)
Information Request	7(4.1%)	4(5.6%)	3(3%)	19(10.3%)	-	19(16.4%)	41(6.6%)	27(8.6%)	14(4.5%)
Statement-not-opinion	17(9.9%)	3(4.2%)	14(14.1%)	23(12.4%)	10(14.5%)	13(11.2%)	50(8%)	13(4.1%)	37(12%)
Statement-opinion	12(7%)	7(9.7%)	5(5.1%)	23(12.4%)	9(13%)	14(12.1%)	15(2.4%)	10(3.2%)	5(1.6%)
Request Clarification	-	-	-	8(4.3%)	-	8(6.9%)	45(7.2%)	12(3.8%)	33(10.7%)
Assert Clarification	6(3.5%)	1(1.4%)	5(5.1%)	22(11.9%)	20(29%)	2(1.7%)	20(2.3%)	13(4.1%)	7(2.3%)

Table 6: Distribution of most frequent dialog acts, ordered by dimension and frequency, in the utterances of Researcher (R), Query Analyst (A), or both (R+A) for each site s_1, s_2, s_3.

Researchers at s_1 give more action directives, and ask few questions. Whereas at s_2, both analysts and researchers give few action directives and exchange many more questions. However, they compensate by higher frequency of clarification assertions. At s_1, there is no occurrence of clarification requests from the $\mathrm{Backward\text{-}looking\ Function}$ dimension. One explanation could be that the participants catch up on the previously discussed issues in live meetings.

At s_3, we observe the occurrence of many result submissions, showing good signs of solving the task. There is also high frequency of meeting schedules requested by the researchers. It is interesting to see that researchers pose more questions than analysts, and give more action directives.

As expected, analysts request more clarifications than researchers, and researchers assert more clarifications than analysts. Similar to s_2, there is high frequency of clarification requests, showing discussion on issues raised previously in the dialog.

4.4. Conversation Analysis with Transition Graph

It may not be straightforward to interpret the above metrics in terms of their implications for analyzing query analysts-research communication. As such, one aspect worth investigating is whether the DREAM-based annotations allow to reveal meaningful high-level patterns in interactions. This work can facilitate conversation flow analysis, by representing utterances with tags that capture semantics and communicative actions.

Experimental Setup. In order to demonstrate this capability, we performed an analysis on the annotated dataset with respect to the most frequent pairs of dialog acts in consecutive turns. We generate a directed graph (Figure 5) to illustrate the frequency of transitioning from one dialog act to another between two consecutive turns in the conversations.

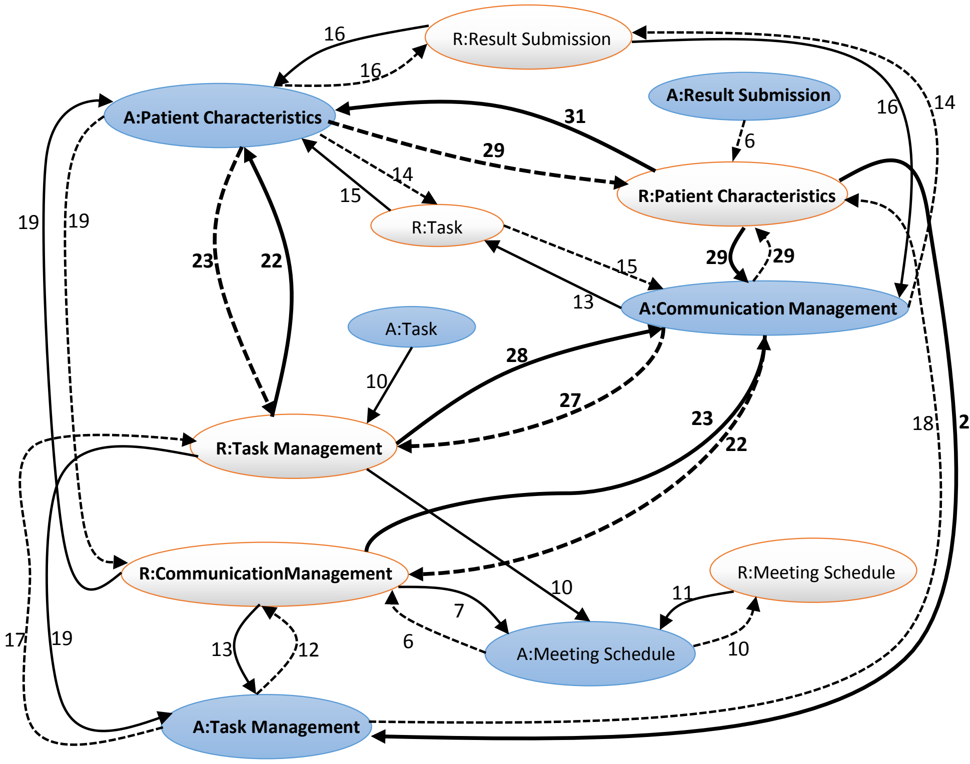

Figure 5: Graph shows the transitioning from one dialog act to another between two consecutive turns in the conversations. We decode two types of nodes (also presented with different colors): R-nodes for acts of utterances by Researchers, and A-nodes for those by query Analysts. A directed edge from an A-node to an R-node is shown with dotted line.

Each unique act in Information Level is represented by a node in the graph. We distinguish between two types (presented with different colors): *R-nodes* for tags of utterances by Researchers, *A-nodes* for those by query Analysts. Each directed edge represents the transition of tags from one turn to the subsequent turn. The edge weight denotes the sum of occurrences of node pairs in consecutive turns.

Results. This graph representation yields several interesting results. First, it helps us outline the nodes with the highest centrality, being Communication Management, Task Management, and Patient Characteristics. Centrality, typically used as a measure of how many connections one node has to other nodes, here serves as an indicator to identify the most frequently occurring dialog acts in the transition graph.

We observe that Communication Management attributes to frequent utterances, which help maintain social obligations in a dialog. The high centrality of Task Management and Patient Characteristics indicates the frequent discussions about patient cohort specification, as is also expected in this domain of discourse.

Furthermore, it is interesting to note that utterances related to Meeting Schedule are often followed by conversation on Patient Characteristics. This indicates the need during the query mediation process for repetitive refinement of the cohort criteria even after personal meetings.

We note the high occurrence of loops in the interaction, such as between A:Patient Characteristics and R:Patient Characteristics, also between R:Task Management and A:Patient Characteristics. These loops capture potential points of bottlenecks in the dialog where the complexity of the mediation increases, leading to repetitive refinements between the parties that last up to several months.

We also observe that Result Submission does not necessarily lead to task completion, rather it is more often followed by Patient Characteristics or Communication Management. This important observation indicates the complexity of negotiation driven by the need for further dialog even after result submission.

5. Discussion

5.1. Implications of Findings

There are various interesting findings drawn from this study. First, we note that it is difficult to distinguish interpretations of utterances related to Task and Task Management, attributing to lower kappa scores.

Classification of acts is also difficult for Backward-looking Function. This problem is related to the difficulty of properly recognizing the antecedent (previous utterance unit or set of units to which the current utterance responds). For example, it is usually hard to decide between Accept and Yes-answer, since this requires finding and correctly interpreting the antecedent as Action Directive or Information Request. This is particularly difficult in this setting, where antecedents are located in turns that consist of long and bulky text messages.

We observe that similar issues are also raised in the work of Core and Allen [5], where DAMSL is originally introduced. The results of their annotation experiments with test dialogs from a collection of discussions between humans on train-related transportation problems. The lowest value of kappa scores (0.15) of the annotations occur in Committing-Speaker-Future-Action, which is the equivalent of our Commissive label. The low scores of kappa are argued to be the result of annotators' difficulty of properly distinguishing if the speaker is making an agreement or acknowledgment. Agreement present commitment done at the Task level, whereas acknowledgment is performed at the Communication Management level. This subtle distinction is not trivial for the annotators.

Similar to our results, Core and Allen also report high values of pairwise agreement (0.99) with low kappa values (0.14) in the first dimension Communicative Status, particularly in the label Unintelligable label. Since kappa is adjusted by the measure of expected agreement (P_e), which is sensitive to the variance of variables, it penalizes reliability even for very high values of observed pairwise agreement (P_a). However, this measure fails to capture the ambiguity of the utterances communicative act. For example, although $\mathrm{Information\ Level}$ has low variance when considering it has three possible categories Task, $\mathrm{Task\ Management}$, and $\mathrm{Communication\ Management}$, we still observe that it is usually difficult to distinguish interpretations of the first two.

The findings play a significant role in the advancement of automated DA classification techniques, which have major limitations in the domain of clinical research. We are planning to investigate these techniques in our future work. The design of such techniques should be guided accordingly to calibrate parameters and metrics in the training and evaluation phases, acknowledging the high ambiguity in certain categories even by human annotators.

In cases where it is nontrivial to distinguish interpretations of utterances pertaining to different dialog acts, the strategies that could be followed to improve annotations include strengthening the familiarity of the annotators with extended description and examples of dialog acts, considering to merge dialog acts based on the individual context or task, as well as applying iterative disagreement resolution for the annotations.

5.2. Limitations

While novel in its two-folded contribution and the analyzed data, the most significant limitation of our study is the small sample size of cases. Although we analyzed only 20 cases, we included 1333 utterances for schema development. Other seminal works for dialog act classification, such as DAMSL, were designed based on a comparable sample sizes (600 utterances). In addition, it is very difficult to collect such data in this domain. Therefore, despite the small sample size, this work is one of the first to shed light on this communication space and is valuable for setting the stage for additional research in this area. Furthermore, the hierarchical nature of the proposed scheme accommodates appropriate extension of both shallow discourse and task-oriented acts. It is worth noting the variability of the emails contained in these cases. Most of the variability in case size comes from Site 3, which also makes up four-fifths of the email messages. Cases from Sites 1 and 2 average 5 to 6 emails per case with a standard deviation of 2 emails. Cases from Site 3 average to 23 emails per case with a standard deviation of 22. The study is also limited to only one language, English, in which the emails are provided. Future work can look at communication in other languages.

When investigating query negotiation processes, we were aware of the frequent need for in-person meetings or phone conversations for supplementing email communications during the negotiation process. Therefore, a limitation of this study is that it does not account for the complexities in the in-person meetings. One of our future works is to triangulate and analyze data from email, phone conversation, and in-person communication.

In order to improve data query mediation in these difficult cases, we believe that clinical researchers would greatly benefit from a mixed-initiative dialogue system that enables human-machine collaboration. We envision an intelligent conversational agent to act as a broker between the clinical data and clinical researchers, while guiding them step-by-step through effective and efficient query optimization. The results of this work can be used in the development of the Natural Language Understanding (NLU) module in an intelligent dialogue-based system. More precisely, the identified dialog acts can be used in the design of the NLU module, and the annotated utterances can be used for the training and evaluation of classification techniques that automatically map sentences to dialog acts. The implementation of dialogue system is the target of our future work.

5.3. Comparison to Related Works

To the best of our knowledge, this work is the first to investigate the discourse carried out via email during clinical research query mediation and contributes a novel taxonomy for classifying communicative actions in this domain. Dialog act studies have been prominent in other fields, particularly contributing with annotation schemes for task-oriented human-to-human or human-computer interaction in specific domains.

Physician-patient communication is a crucial element of clinical practice, hence it has constantly been the focus of research on discourse annotation and analysis. A major stream of works has produced useful coding systems for the study of general medical encounters [24, 14] and oncology visits [13]. The closest in features to our proposed scheme, with a consistent theoretical basis in Speech Act Theory, is Generalized Medical Interaction Analysis System (GMIAS) [15]. It is designed to study physician-patient communication about the adherence to antiretroviral (ARV) treatment[2]. GMIAS also assigns separate tags for annotating each utterance with the communicative function and content. However, in DREAM we propose the annotation of an utterance with more than two tags, additionally capturing the backward-looking function.

The design of classification schemes for dialog act annotation has been prevalent in other domains outside healthcare. The TRAINS project [12] illustrates a

[2] Medication treatment that prevent the growth of HIV

case study in building a conversational planning agent. The goal of the agent is to enable human and system dialog-based interaction for managing a railway transportation system, i.e. finding the best way to realize the transportation by train on a map. This dialog approach is task-oriented and built upon simulated human-computer interactions. It exploits speech acts to aid agents move through different modalities until they achieve a shared plan. Speech acts are further executed to generate natural language utterances, which constitute the system's output to the human user.

As part of Verbmobil-2 project, Alexandersson et al. [25] present a dialog act scheme for annotations of dialogs that enable negotiation on travel planning appointment scheduling between participants. The set of dialog acts is structured in the form of a hierarchy, whose leaves have growing specificity. Similar to our work, the dialog acts are modeled based on task-oriented, human-to-human dialogues.

Another dialog act coding scheme based on task-oriented dialogues, human-to-human conversations is proposed by Carletta et al. [26] in the scope of HCRC Map Task project. The communication is targeted at the reconstruction of a route on slightly different maps belonging to the participants. The coding systems defines categories of conversational moves structured in a tree representation. These schemes were designed with a particular task as a target and a specific application domain. They also contain overlapping sets of communicate functions.

DAMSL [5] framework, designed as part of the Discourse Research Initiative, marked an important step forward in dialog act classification through its domain-independence and multi-dimensionality features. Variations and extensions of DAMSL are used to construct other schemes, such as Switchboard-DAMSL [18], designed for specific purposes. Bunt et al. [17] uses the foundations of DAMSL in combination with tags of alternative models in a comprehensive schema named DIT++. This schema keeps the multi-dimensionality features of DAMSL, extending it with tags about turn allocation (turn management information) and dialogue structuring (topic and dialogue structure information).

We were motivated to use DAMSL as a basis of our work, because it is not only a well-known schema for DA annotation, but it is also easier to apply and has adequate granularity. The rationale behind this choice is to facilitate the reuse of our proposed DREAM framework, keeping it as simple as possible for future training or implementations.

Compared to DA annotation schemes of other domains, a significant extension that characterizes DREAM taxonomy is the detailed specification of content-related acts in the category Task, which capture the different features of the clinical researchers' data needs. Another major extension is performed in the category Information Relation of the Backward-looking function, which captures how the content of the current utterance relates to antecedents in the dialog. These characteristics particularly differentiate DREAM scheme from schemes of other domains. Whereas for the domain at hand, this is the first attempt at classifying dialog acts of the query mediation discourse.

5.4. Application

We highlight several applications for which a DA classification scheme is important, grouping them into two main classes: *dialog systems* that enable human-computer conversational mechanisms, and *automatic analysis*, which aids the interpretation of human-human communication.

5.4.1. Dialog Systems

The first and foremost application of a DA classification scheme is seen in the discipline of dialog systems, which act as participants in a conversation with human users for task completion or problem solving. A core component of such systems is Dialog Management (DM), which operates the communication between humans and computer-based systems using natural language. DM technologies bridge the gap by modeling users' intentionality, making predictions and decisions of next steps in the negotiation process, and resolving conflicts.

A significant and established element of research in the Natural Language Processing (NLP) approach to DM is the annotation of utterances with dialog acts, referred to as DA classification. An example of the value of classifying DAs is in the detection and use of questions, assertions and instructions to communicate with machines. However, the classification task is time-consuming and requires highly-trained personnel. DA classification has been the focus of many works outside the medical domain. These works investigate supervised [27, 3, 28] and unsupervised techniques [29, 30, 31, 32, 33, 34]. Their goal is to enable large-scale automated DA annotation in order to reduce the costs and increase the coverage of training data.

5.4.2. Human Communication Analysis

Rigorous knowledge and classification of dialog acts is useful for automated analysis of human-human communications. Besides providing important insights and better understanding, the findings of such analysis can be exploited for the evaluation of human-machine dialogue systems. An interesting direction in assessing the performance of these systems is their comparison with human-to-human dialogues. Identification of similarities and differences in structure between machine-human and human-to-human dialogs can advance the development of automated systems.

DA tagging schemes particularly aid the analysis of such communications through their use in the various techniques of discourse summarization [35, 36], induction of discourse structure [37, 38], automatic topic detection and its use in the comparative analysis of communications [39]. In the latter, it is important to study the variability in the content and structure of communications at different institutions. For clinical research query mediation, this is particularly important given the heterogeneous representations and implementations of EHR repositories at clinical institutions.

In medicine, these techniques have been consistently investigated to advance the design of dialog systems and analysis of communications between patients and caregivers [40, 36, 41, 42, 43, 39]. However, their application to clinical research communication is remarkably missing. In-depth investigation of the conversations carried during clinical research query mediation helps to gain better insights into the process, and accordingly react to improve the access of complex data from institutional databases in support of clinical and translational research.

6. Conclusions

The proposed DREAM-taxonomy is a novel scheme for analyzing dialog in email-based clinical research query mediation and its annotation with information on the dialog acts performed by dialog segments. We have demonstrated that this scheme is reliable for labeling query negotiation conversations. Furthermore, it helps to summarize and identify high-level patterns of conversation in this negotiation space, as well as draw comparison between the communication patterns across different clinical institutions.

The introduction of this model plays an important role in advancing research in dialog systems for automated query optimization. This line of research is promising for alleviating the challenges of the biomedical query mediation process, but unfortunately still very limited presently. We plan to use the identified dialog acts in the future development of an intelligent human-machine dialog agent that assists clinical query refinement.

Acknowledgments

This work was supported by The National Library of Medicine Grants R01LM009886 (PI: Weng). We thank Dr. Julia Hirschberg for the precious feedback.

Competing Interests

None.

Contributors

Dr. Hoxha conceived the concept of this submission, implemented the processing modules, developed the scheme, performed annotations, conducted the evaluations, performed literature research, drafted and revised the manuscript. Dr. Hoxha is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Dr. Chandar was very active in discussions of the methodology and evaluation, performed annotations, and revised the manuscript.

Dr. He helped in discussions of the evaluation, performed annotations, and revised the manuscript.

Dr. Cimino provided the data from one institution, helped in discussions of the analysis and results, and revised the manuscript.

Dr. Hanauer provided the data from one institution, helped in discussions of the analysis and results, and revised the manuscript.

Dr. Weng initiated the research, collected the data, supervised the research, contributed significantly to methodology design, and edited and approved the submission. As corresponding author, Dr. Weng takes primary responsibility for the research reported in this manuscript.

References

[1] T. Hao, A. Rusanov, C. Weng, Extracting and normalizing temporal expressions in clinical data requests from researchers, in: Proceedings of the 2013 International Conference on Smart Health, ICSH'13, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 41–51.

[2] M. Walker, R. Passonneau, Date: A dialogue act tagging scheme for evaluation of spoken dialogue systems, in: Proceedings of the First International Conference on Human Language Technology Research, HLT '01, Association for Computational Linguistics, Stroudsburg, PA, USA, 2001, pp. 1–8. doi:10.3115/1072133.1072148.

[3] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, M. Meteer, Dialogue act modeling for

automatic tagging and recognition of conversational speech, Computational Linguistics 26 (3) (2000) 339–373. doi:10.1162/089120100561737.

[4] A. Ezen-Can, K. Elizabeth Boyer, Unsupervised classification of student dialogue acts with query-likelihood clustering, in: Proceedings of the International Conference on Educational Data Mining, EDM'13, 2013, pp. 20–27.

[5] M. G. Core, J. F. Allen, Coding dialogues with the damsl annotation scheme, in: Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines, 1997.

[6] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, S. B. Johnson, Elixr:an approach to eligibility criteria extraction and representation, Journal of the American Medical Informatics Association 18. doi:10.1136/amiajnl-2011-000321.

[7] D. A. Hanauer, G. W. Hruby, D. G. Fort, E. A. M. Luke V. Rasmussen, C. Weng, What is asked in clinical data request forms? a multi-site thematic analysis of forms towards better data access support, in: AMIA 2014, American Medical Informatics Association Annual Symposium, Washington DC, USA, November 15-19, 2014, 2014, pp. 616–625.

[8] G. W. Hruby, A. B. Wilcox, C. Weng, Analysis of query negotiation between a researcher and a query expert, in: AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, Illinois, USA, November 3-7, 2012, AMIA, 2012.
URL http://knowledge.amia.org/amia-55142-a2012a-1.636547

[9] G. Hruby, M. B. MR, J. Cimino, J. Gao, A. Wilcox, J. Hirschberg, C. Weng, Characterization of the biomedical query mediation process, in: AMIA 2013,Clinical Research Informatics Summit, San Francisco, CA, 18-22 March 2013, 2013, pp. 89–93.

[10] J. Austin, How to do things with words, Clarendon Press, Oxford, 1962, 1962.

[11] J. Searle, Speech acts: An essay in the philosophy of language, 1969.

[12] J. F. Allen, L. K. Schubert, P. H. George Ferguson, C. H. Hwang, M. L. Tsuneaki Kato, N. Martin, B. Miller, M. Poesio, The trains project: A case study in defining a conversational planning agent (1994).

[13] S. Ford, A. Hall, D. Ratcliffe, L. Fallowfield, The medical interaction process system (mips): an instrument for analysing interviews of oncologists and patients with cancer, Soc.Sci.Med. 50 (4) (2000) 553–566.

[14] D. Roter, The roter method of interaction process analysis, in: RIAS Manual, Johns Hopkins University, 2013.

[15] M. Laws, GMIAS coding manual (2009).
URL `https://sites.google.com/a/brown.edu/m-barton-laws/home/gmias`

[16] H. Bunt, J. Alex, J. Carletta, J. woong Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-belis, L. Romary, C. Soria, D. Traum, Towards an iso standard for dialogue act annotation, in: Proceedings of the International Conference on Language Resources and Evaluation, LREC'13, 2010.

[17] H. Bunt, The DIT++ taxonomy for functional dialogue markup, in: Proceedings of the AMAAS Workshop - Towards a Standard Markup Language for Embodied Dialogue Acts, 2009, pp. 13–24.

[18] D. Jurafsky, E. Schriberg, D. Biasca, Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual (1997).
URL http://web.stanford.edu/~jurafsky/ws97/manual.august1.html

[19] Z. Luo, S. Johnson, C. Weng, Semi-automatic induction of semantic classes from free-text clinical research eligibility criteria using umls, in: AMIA 2010, American Medical Informatics Association Annual Symposium, 2010, pp. 487–491.

[20] H. Linstone, M. Turoff, The Delphi method: techniques and applications (1975).

[21] J. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382.

[22] J. Landis, G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. doi:10.2307/2529310.

[23] K. J. Preacher, Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence (2001).
URL http://quantpsy.org

[24] S. H. Kaplan, S. Greenfield, J. Ware, J. E., Assessing the effects of physician-patient interactions on the outcomes of chronic disease, Medical Care 27 (1989) 110–127.

[25] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, M. Siegel, Dialogue acts in verbmobil-2, Verbmobil Report 226. Saarbruecken: DFKI.

[26] J. A. Carletta, S. Isard, J. Kowtko, G. Doherty-Sneddon, HCRC dialogue structure coding manual (1996).

[27] O. Ferschke, I. Gurevych, Y. Chebotar, Behind the article: Recognizing dialog acts in wikipedia talk pages, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 777–786.

[28] M. Tavafi, Y. Mehdad, S. Joty, G. Carenini, R. Ng, Dialogue act recognition in synchronous and asynchronous conversations, in: Proceedings of the SIGDIAL 2013 Conference, Association for Computational Linguistics, Metz, France, 2013, pp. 117–121.
URL `http://www.aclweb.org/anthology/W/W13/W13-4017`

[29] N. Crook, R. Granell, S. Pulman, Unsupervised classification of dialogue acts using a dirichlet process mixture model, in: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 341–348.

[30] A. Ritter, C. Cherry, B. Dolan, Unsupervised modeling of twitter conversations, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguis-

tics, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 172–180.

[31] D. Lee, M. Jeong, K. Kim, S. Ryu, G. Lee, Unsupervised spoken language understanding for a multi-domain dialog system, Audio, Speech, and Language Processing, IEEE Transactions on 21 (11) (2013) 2451–2464. doi:10.1109/TASL.2013.2280212.

[32] A. Ezen-Can, K. Boyer, Combining task and dialogue streams in unsupervised dialogue act models, in: Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Association for Computational Linguistics, Philadelphia, PA, U.S.A., 2014, pp. 113–122.

[33] B. D. Eugenio, Z. Xie, R. Serafin, Dialogue act classification, higher order dialogue structure, and instance-based learning, Dialogue and Discourse 1 (2) (2010) 1–24.

[34] J. OŚhea, Z. Bandar, K. Crockett, A multi-classifier approach to dialogue act classification using function words, in: N. Nguyen (Ed.), Transactions on Computational Collective Intelligence VII, Vol. 7270 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 119–143.

[35] M. Joshi, C. P. Rosé, Using transactivity in conversation for summarization of educational dialogue, SLaTE Workshp on Speech and Language Technology in Education.

[36] R. C. Lacson, R. Barzilay, W. J. Long, Automatic analysis of medical dia-

logue in the homehemodialysis domain: structure induction and summarization, Journal of Biomedical Informatics 39 (2006) 541–555.

[37] M. Finke, M. Lapata, A. Lavie, L. Levin, L. M. Tomokiyo, T. Polzin, K. Ries, A. Waibel, K. Zechner, Clarity: Inferring Discourse Structure from Speech, in: AAAI Spring Symposium Series, 1998, pp. 25–32.

[38] M. Rotaru, D. J. Litman, Discourse structure and performance analysis: Beyond the correlation, in: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 178–187.

[39] B. C. Wallace, M. B. Laws, K. Small, I. B. Wilson, T. A. Trikalinos, Automatically annotating topics in transcripts of patient-provider interactions via machine learning, Medical Decision Making 34 (4) (2014) 503–512. doi:10.1177/0272989X13514777.

[40] L. Rojas-Barahona, T. Giorgino, Adaptable dialog architecture and runtime engine (adarte): a framework for rapid prototyping of health dialog systems, International Journal of Medical Informatics (2009) 56–68.

[41] T. Bickmore, T. Giorgino, Health dialog systems for patients and consumers, Journal of Biomedical Informatics 39 (5) (2006) 556–571. doi:10.1016/j.jbi.2005.12.004.

[42] T. W. Bickmore, D. Schulman, C. L. Sidner, A reusable framework for health counseling dialogue systems based on a behavioral medicine

ontology, Journal of Biomedical Informatics 44 (2) (2011) 183–197. doi:10.1016/j.jbi.2010.12.006.

[43] G. Ferguson, J. Quinn, C. Horwitz, M. Swift, J. Allen, L. Galescu, Towards a personal health management assistant, Journal of Biomedical Informatics 43 (5) (2010) 13–16.