Analyzing and Characterizing User Intent in Information-seeking Conversations

Chen Qu
University of Massachusetts Amherst
chenqu@cs.umass.edu

Liu Yang
University of Massachusetts Amherst
lyang@cs.umass.edu

W. Bruce Croft
University of Massachusetts Amherst
croft@cs.umass.edu

Johanne R. Trippas
RMIT University
johanne.trippas@rmit.edu.au

Yongfeng Zhang
Rutgers University
yongfeng.zhang@rutgers.edu

Minghui Qiu
Alibaba Group
minghui.qmh@alibaba-inc.com

ABSTRACT

Understanding and characterizing how people interact in information-seeking conversations is crucial in developing conversational search systems. In this paper, we introduce a new dataset designed for this purpose and use it to analyze information-seeking conversations by user intent distribution, co-occurrence, and flow patterns. The MSDialog dataset is a labeled dialog dataset of question answering (QA) interactions between information seekers and providers from an online forum on Microsoft products. The dataset contains more than 2,000 multi-turn QA dialogs with 10,000 utterances that are annotated with user intent on the utterance level. Annotations were done using crowdsourcing. With MSDialog, we find some highly recurring patterns in user intent during an information-seeking process. They could be useful for designing conversational search systems. We will make our dataset freely available to encourage exploration of information-seeking conversation models.

KEYWORDS

Information-seeking; Conversational Search; User Intent

1 INTRODUCTION

Conversational assistants (CAs) such as Siri and Cortana are becoming increasingly popular. Users can issue simple queries and commands to a CA by voice to conduct single-turn QA or goal-oriented tasks, such as asking for weather and setting timers. However, CAs are not yet capable of handling complicated information-seeking tasks which involve multiple turns of information exchange. These conversations are typically referred to as *information-seeking conversations*, where the information provider (agent) provides answers to a query from an information seeker (user) and the agent modifies the answers based on user feedback.

To build functional and natural CAs that can reply to more complicated tasks we need to understand how users interact in these information-seeking environments. Thus, it is necessary to analyze and characterize user interactions and utterance intent. At CAIR[1] workshop at SIGIR'17, researchers indicated that there is a lack of conversational datasets to conduct studies. Therefore in this paper, we address this issue by collecting conversation data and creating the *MSDialog*[2] dataset. We present an analysis of user intent here, but MSDialog could also be used to conduct other dialog related tasks including response ranking and user intent prediction.

For effective analysis of user intent in an information-seeking process, the data should be multi-turn information-seeking dialogs. To support natural dialogs, conversational systems should be modeled closely to human behavior, thus the data should come from conversation interactions between real humans. As shown in Table 1, we found that most existing dialog datasets are not appropriate for user intent analysis. The most similar data to ours is the Ubuntu Dialog Corpus (UDC), which also contains multi-turn QA conversations in the technical support domain. However, the user intent in this dataset is unlabeled. In addition, UDC dialogs are in IRC (Internet Relay Chat) style. This informal language style contains a significant amount of typos, internet language, and abbreviations. Another dataset, the DSTC 6 Conversation Modeling track data contains knowledge grounded dialogs from Twitter. However, this dataset contains scenarios where users do not request information explicitly, which do not fit the information-seeking narrative. Thus these datasets are not appropriate for user intent analysis.

Table 1: Comparison of related dialog datasets

Dataset	Multi-turn	Human-human	Information-seeking	User intent label
DSTC 1-3 [4]	✓			
DSTC 4-5 [6]	✓	✓		
Switchboard [3]	✓	✓		
Twitter Corpus [12]	✓	✓		
DSTC 6 (2nd Track) [5]	✓	✓	✓	
Ubuntu Dialog Corpus [8]	✓	✓	✓	
MSDialog	✓	✓	✓	✓

For open-domain chatting, it is common practice to train chatbots with social media data such as Twitter [13]. Similarly, real human-human multi-turn QA dialogs are the appropriate data for characterizing user intent in information-seeking conversations. In technical support online forums, a thread is typically initiated by a user-generated question and answered by experienced users (agents). The users may also exchange clarifications with the agents or give feedback based on answer quality. Thus the flow of a technical support thread resembles the information-seeking process if we

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210124

[1] https://sites.google.com/view/cair-ws/ [2] The MSDialog dataset is available at https://ciir.cs.umass.edu/downloads/msdialog

consider threads as dialogs and posts as turns/utterances in dialogs. We created MSDialog by crawling multi-turn QA threads from the Microsoft Community[3] and annotate them with fine-grained user intent types on an utterance level based on crowdsourcing on Amazon Mechanical Turk (MTurk)[4].

With this new dataset, we analyze the user intent distribution, co-occurrence patterns and flow patterns of large-scale QA dialogs. We gain insights on human intent dynamics during information-seeking conversations. One of the most interesting findings is the high co-occurrence of negative feedback and further details, which typically occurs after a potential answer is given. This co-occurrence pattern provides feedback about the retrieved answer and critical information about how to improve the previous answer. In addition, negative feedback often leads to another answer response, indicating that co-occurrence and flow patterns associated with negative feedback can be the key to iterative answer finding.

To sum up, our contributions can be summarized as follows. (1) We create a large-scale annotated dataset for multi-turn information-seeking conversations, which is the first of its kind to the best of our knowledge. We will make our dataset freely available to encourage relevant studies. (2) We perform in-depth data analysis and characterization of multi-turn human QA conversations. We analyze the user intent distribution, co-occurrence and flow patterns. Our characterizations also hold in similar data (UDC). Our findings could be useful for designing conversational search systems.

2 RELATED WORK

Early conversational search systems through man-machine dialog include the THOMAS system by Oddy [10]. It allowed users to conduct searches through dialogs. Belkin et al. [1] explored and demonstrated the justifiability of using information interaction dialogs to design the interaction mechanisms in IR systems. Shah and Pomerantz [14] considered community QA as information-seeking processes and built models to predict answer quality. Radlinski and Craswell [11] described a conceptual framework for conversational IR and the major research issues that must be addressed.

Recently, two observational studies captured how participants communicate and conduct searches in a voice-only setting [15, 16]. Both studies attempted to provide initial labeling for each utterance. Trippas et al. [16] analyzed the initial turns for patterns to classify with a qualitative analysis approach. The MISC data [15] came from similar experiments with data release including video, audio, and even emotions. Even though they offered valuable insights on how users conduct searches in a conversation, the data is not sufficient to perform a large-scale analysis and model training.

Also related to conversational search, Marchionini [9] and White and Roth [17] addressed the importance of exploratory search, where the behavior of search is beyond a simple look up and more like learning and investigating. In this setting, the interpretation of user intent would rely heavily on the interactions between human and computer. This highlights the research need to characterize and understand user intent dynamics in information-seeking processes.

3 THE MSDIALOG DATA

Our data collection contains two sets: the complete set and a labeled subset. Both will be publicly available. The complete set could be useful for unsupervised/semi-supervised model training. The data used in the user intent analysis is the labeled subset. In this section, we describe the three stages of generating MSDialog, which are data collection, taxonomy definition, and user intent annotation.

3.1 Data Collection

We crawled over 35,000 dialogs from Microsoft Community, a forum that provides technical support for Microsoft products. This well-moderated forum contains user-generated questions with high-quality answers provided by Microsoft staff and other experienced users including Microsoft Most Valuable Professionals.

To ensure the quality and consistency of the dataset, we selected about 2,400 dialogs that meet the following criteria for annotation: (1) With 3 to 10 turns. (2) With 2 to 4 participants. (3) With at least one correct answer selected by the community. (4) Falls into one of the categories of Windows, Office, Bing, and Skype, which are the major categories of Microsoft products.

We observe that dialogs with a large number of turns or participants can contain too much noise, while dialogs with limited turns and participants are relatively clean. By choosing dialogs with at least one answer, we can use this dataset for other tasks such as answer retrieval. Also, by limiting the categories to several major ones, we can ensure language consistency across different dialogs, which is better for training neural models.

3.2 Taxonomy for User Intent in Conversations

We classify user intent in dialogs into 12 classes shown in Table 2. Seven of the classes (*OQ*, *RQ*, *CQ*, *FD*, *PA*, *PF*, *NF*) were first introduced in FIRE'10[5]. Bhatia et al. [2] added the eighth class of *Junk* as they observed a significant amount of posts with no useful information in their data (200 dialogs labeled with eight classes).

We added four more classes to Bhatia et al. [2]'s taxonomy: *Information Request*, *Follow Up Question*, *Greetings/Gratitude*, and *Others*. We observed that agents' inquiries about user's version of software or model of computer is common in this technical support data and does not necessarily overlap with *Clarifying Question*. *Follow Up Question* is another utterance class in MSDialog as users sometimes expect agents to walk them step-by-step through the technical problem. *Greetings/Gratitude* is quite common in the data. Finally, the *Others* class is for utterances that cannot be classified with other classes. Note, each utterance can be assigned multiple labels because an utterance can cover multiple intent (e.g. *GG+FQ*).

3.3 User Intent Annotation with MTurk

3.3.1 Procedure. We employed crowdsourcing workers through MTurk to label user intent of each utterance using a set of 12 labels that is described in Section 3.2. The workers are required to have a HIT (Human Intelligence Task) approval rate of 97% or higher, a minimum of 1,000 approved HITs, and be located in US, Canada, Australia or Great Britain. The workers are paid $0.3/dialog.

In this annotation task, the workers are provided with a complete dialog. They are instructed to go through a table of labels with descriptions and examples before they proceed. For each utterance, the workers are tasked to choose all applicable labels that represent the user intent of the utterance and leave a comment if they choose the *Others* label.

[3] https://answers.microsoft.com [4] https://www.mturk.com/

[5] https://www.isical.ac.in/~fire/2010/task-guideline.html

Table 2: Descriptions and examples of user intent classes

Code	Label	Description	Example
OQ	Original Question	The first question by a user that initiates the QA dialog.	If a computer is purchased with win 10 can it be downgraded to win 7?
RQ	Repeat Question	Posters other than the user repeat a previous question.	I am experiencing the same problem ...
CQ	Clarifying Question	Users or agents ask for clarification to get more details.	Your advice is not detailed enough. I'm not sure what you mean by ...
FD	Further Details	Users or agents provide more details.	Hi. Sorry for taking so long to reply. The information you need is ...
FQ	Follow Up Question	Users ask follow up questions about relevant issues.	Thanks. I really have one simple question – if I ...
IR	Information Request	Agents ask for information of users.	What is the make and model of the computer? Have you tried installing ...
PA	Potential Answer	A potential answer or solution provided by agents.	Hi. To change your PIN in Windows 10, you may follow the steps below: ...
PF	Positive Feedback	Users provide positive feedback for working solutions.	Hi. That was exactly the right fix. All set now. Tx!
NF	Negative Feedback	Users provide negative feedback for useless solutions.	Thank you for your help, but the steps below did not resolve the problem ...
GG	Greetings/Gratitude	Users or agents greet each others or express gratitude.	Thank you all for your responses to my question ...
JK	Junk	There is no useful information in the post.	Emojis. Sigh .... Thread closed by moderator ...
O	Others	Posts that cannot be categorized using other classes.	N/A

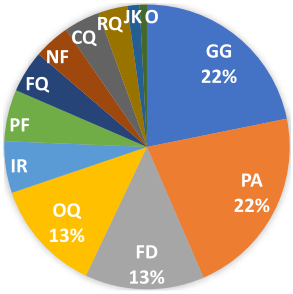

Figure 1: Distribution

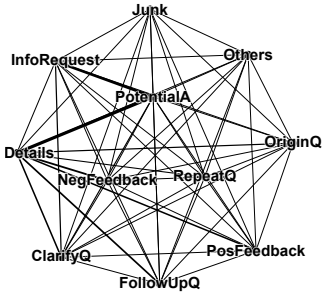

Figure 2: Co-occurrence

3.3.2 Quality Assurance. To ensure the annotation quality, we employed two workers on each dialog. We calculated the inter-rater agreement using Fuzzy Kappa [7] for this one-to-many classification task. We applied the threshold of 0.18 to filter the dialogs with too small Kappa scores, which reduced the number of dialogs by 9%.

4 DATA ANALYSIS & CHARACTERIZATION

4.1 Data Statistics

The annotated dataset contains 2,199 multi-turn dialogs with 10,020 utterances. Table 3 summarizes the properties of MSDialog. Each utterance has 1.83 labels on average.

Table 3: Statistics of MSDialog

Items	Min	Max	Mean	Median
# Turns Per Dialog	3	10	4.56	4
# Participants Per Dialog	2	4	2.79	3
Dialog Length (Words)	27	1,467	296.90	241
Utterance Length (Words)	1	939	65.16	47

4.2 User Intent Distribution

Figure 1 shows the user intent distribution. Labels without a percentage are under 10%. *Greetings/Gratitude* and *Potential Answer* are the most frequent labels. This suggests that good manners and answers are at the center of human QA conversations. *Repeat Question* is the most infrequent label except for *Junk* and *Others*, which is because the number of participants is limited to four.

4.3 User Intent Co-occurrence

Label co-occurrence in the same utterance can be useful for understanding user intent. Preliminary results indicate that the most frequent co-occurrence is between *Greetings/Gratitude* and another label, suggesting good manners of forum users. Nevertheless, we removed *GG* for the analysis later to emphasize more on crucial user intent of information-seeking interactions.

The user intent co-occurrence graph with undirected edges weighted by co-occurrence count is presented in Figure 2. We observe that *Potential Answer* often co-occurs with *Further Details* or *Information Request*. This indicates that agents tend to enrich possible solutions with details, or send *Information Requests* in case the solutions do not work. Also, users tend to give *Negative Feedback* with *Further Details* to explain how the suggested answer is not working. In addition, *Further Details* is observed to co-occur with *Follow Up Question* or *Clarifying Question*, suggesting that when people raise a relevant question, they tend to add details to them.

4.4 User Intent Flow Pattern

We use a Markov Model to analyze the flow patterns in the dialogs as shown in Figure 3. Because of the complexity and diversity of human conversations, many utterances are labeled with multiple user intent. We preprocess the traces (complete user intent flow in a dialog) with multiple labels by only using one label each time. For example, if we have a trace of "*OQ→PA+FD→PF*", we transfer it into two separate traces. The first one is "*OQ→PA→PF*", and the second one is "*OQ→FD→PF*". This preprocessing step can lead to a more concise model compared with using the original multi-labels as nodes. However, it does magnify some user intent nonproportionally. We alleviate the issue by only using dialogs that generate no more than 100 traces. This only filtered 30 dialogs.

In addition, we remove *Greetings/Gratitude* because of the same reason described in Section 4.3. Instead of simply hiding the *GG* node from the final graph, we remove the occurrences of *Greetings/Gratitude* if the utterance has multiple labels or change *GG* to *JK* if the utterance only has one label.

The flow pattern with a Markov model is presented in Figure 3. As highlighted in the graph, a typical user intent transition path of MSDialog is "*INITIAL→OQ→PA→FD→PA→PF→TERMINAL*". This represents the frequent user intent transition pattern in an information seeking process. We can make some observations from the graph : (1) In most cases, dialogs begin with an *Original Question*, sometimes accompanied by *Further Details*. (2) *Original Question* tends to lead to *Potential Answer* and *Information Request*. (3) *Information Request* and *Clarifying Question* tend to lead to *Further Details*. (4) *Positive Feedback* tends to terminate the dialog while

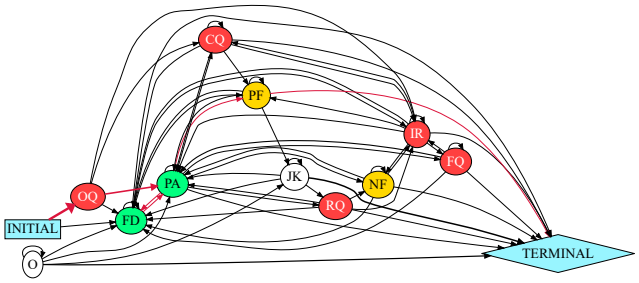

Figure 3: Flow pattern with a Markov model. Node colors: red (questions), green (answer related), yellow (feedback). Edges are directed and weighted by transition probability.

Negative Feedback tends to lead to *Potential Answer* or *Further Details*. (5) Dialogs tend to end after *Others* or *Junk*.

Besides the Markov transition graph, we use a different perspective to inspect the flow pattern by focusing on the user intent transition between turns in each dialog. We find that a quite significant flow path across turns is "*INITIAL*→*OQ*→(*PA*→*FD*)×3→*PA*→*PF*→*TERMINAL*". The "*PA*↔*FD*" circle pattern is typically caused by the "*PA+IR*", "*PA+CQ*", "*NF+FD*" co-occurrences described in Section 4.3 and the "*IR*→*FD*", "*CQ*→*FD*", "*NF*→*PA*" sequential relationship suggested in Figure 3.

4.5 Comparison with Ubuntu Dialog Corpus

Although UDC is less suitable for user intent analysis due to the informal language style, we investigate the characterizations of UDC and compare them to MSDialog since they are both in the technical support domain. We sampled 200 UDC dialogs and annotated user intent with MTurk using the same method with MSDialog. The informal language style of UDC may impact the annotation quality.

4.5.1 Statistics. For this section, we present the statistics for UDC (complete set) and MSDialog (complete set) instead of the dialogs we sampled. As shown in Table 4, UDC dialogs have shorter utterances because of the informal language style.

Table 4: Statistics of UDC & MSDialog (both complete sets)

Items	Ubuntu Dialog Corpus	MSDialog
# Dialogs	930,000	35,000
# Utterances	7,100,000	300,000
# Words (in total)	100,000,000	24,000,000
Avg. # Participants	2	3.18
Avg. # Turns Per Dialog	7.71	8.94
Avg. # Words Per Utterance	10.34	75.91

4.5.2 Data Characterization. *Potential Answer* and *Further Details* are the most significant user intent in UDC, which is consistent with MSDialog. Interestingly, the most common user intent in MSDialog, *Greetings/Gratitude*, is quite rare in UDC. In addition, we observe the exact same top 5 label co-occurrences in UDC as described in Section 4.3. Note that they are not necessarily in the same order. Finally, we found that the flow patterns observed in MSDialog also hold in UDC, except for the tendency from *Positive Feedback* to *TERMINAL*. This can be explained by the scarcity of *Positive Feedback* in UDC. Although the UDC dialogs with informal language style are drastically different from the formal written style of MSDialog, the resemblance in user intent characterizations indicates that human QA conversations, regardless of the communication medium, follow similar patterns.

5 DISCUSSION

In this section we discuss the limitation of our findings. The patterns we discovered are closely related to several design choices, including using dialogs from a well moderated forum in a specific domain. These choices were made to keep the setting as clean as possible as the research community is at an initial stage of this study. Although MSDialog does not cover every aspect of the highly diverse information-seeking conversations, it should be a first step to analyze and predict user intent in an information-seeking setting.

6 CONCLUSIONS

In this paper, we create and annotate a large multi-turn question answering data for research in conversational search. We perform in-depth characterization and analysis of this data to gain insights on the distribution, co-occurrence and flow pattern of user intent in information-seeking conversations. We will make our dataset freely available to inspire future research. Future work will consider using neural architectures for user intent prediction tasks.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1419693 and NSF grant #IIS-1715095. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems. 1970.
[2] S. Bhatia, P. Biyani, and P. Mitra. Classifying User Messages For Managing Web Forum Data. In *WebDB '12*, 2012.
[3] J. J. Godfrey and E. Holliman. Switchboard-1 Release 2. *Jour. of Env. Health*, 1997.
[4] M. Henderson, B. Thomson, and J. D. Williams. The Third Dialog State Tracking Challenge. In *SLT '15*, 2015.
[5] C. Hori and T. Hori. End-to-end Conversation Modeling Track in DSTC6. *arXiv:1706.07440*, 2017.
[6] S. Kim, L. F. D'Haro, R. E. Banchs, J. D Williams, M. Henderson, and K. Yoshino. The Fifth Dialog State Tracking Challenge. In *SLT '16*, 2016.
[7] A. P. Kirilenko and S. Stepchenkova. Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. *Plos One*, 2016.
[8] R. Lowe, N. Pow, I. Serban, and J. Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL '15*, 2015.
[9] G. Marchionini. Exploratory Search: From Finding to Understanding. *Commun. ACM*, 2006.
[10] R. N. Oddy. *Information Retrieval through Man-Machine Dialogue.* 1977.
[11] F. Radlinski and N. Craswell. A Theoretical Framework for Conversational Search. In *CHIIR'17*, 2017.
[12] A. Ritter, C. Cherry, and W. B. Dolan. Unsupervised Modeling of Twitter Conversations. In *ACL '10*, 2010.
[13] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven Response Generation in Social Media. In *EMNLP '11*, 2011.
[14] C. Shah and J. Pomerantz. Evaluating and Predicting Answer Quality in Community QA. In *SIGIR '10*, 2010.
[15] P. Thomas, D. McDuff, M. Czerwinski, and N. Craswell. MISC: A Data Set of Information-seeking Conversations. In *CAIR '17*, 2017.
[16] J. R. Trippas, D. Spina, L. Cavedon, and M. Sanderson. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *CHIIR '17*, 2017.
[17] R. W. White and R. A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm.* Synthesis Lectures on Information Concepts, Retrieval, and Services. 2009.