CLARIFYING USER'S INFORMATION NEED IN CONVERSATIONAL INFORMATION RETRIEVAL

by

SOUMIK MANDAL

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Communication, Information and Media

Written under the direction of

Chirag Shah, Ph.D.

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2021

ABSTRACT OF THE DISSERTATION

Clarifying user's information need in conversational information retrieval

by Soumik Mandal

Dissertation Director: Chirag Shah, Ph.D.

With traditional information retrieval systems, users are expected to express their information need adequately and accurately to get appropriate responses from the system. This setup generally works well for simple tasks. However, with the increase of task complexities, users face difficulties in expressing information need in the form as expected by the system. Therefore, the case of clarifying the user's information need by the system arises. In current search systems, support in such cases is provided in the form of query suggestions or query recommendations.

In contrast, conversational information retrieval systems enable the user to interact with the system in the form of dialogs. The conversational approach to information retrieval enables the system to better support the user's information need by asking clarifying questions. However, current research in both natural language processing and information retrieval systems does limited explaining how to form such questions and at what stage of dialog clarifying questions should be asked. To address the research gap, this dissertation investigates the nature of a user's information-seeking conversations with a dialog agent, where the latter is simulating the role of an intelligent system supporting the user's information need. The goal is to identify the type of questions and their patterns an automated intermediary should ask to negotiate and clarify the user's information need. More specifically, this research explores how an intelligent search system should ask the user questions to clarify her information need in complex task scenarios.

This dissertation used the Taskmaster-1 dataset, which collected prior, simulated written and spoken conversations between the user and a conversational agent from multiple task domains. In this research, a subset of these conversations was qualitatively coded by expert annotators at the utterance level. The coding labels were derived from Taylor's (1967) questions and negotiations in information-seeking conversations. Additionally, the utterances from the selected conversations were also labeled with the speaker's conversational roles in the utterance as per the COR model (1992). A domain-independent typology of clarification questions was established from the analysis of the coded dialogs. Our analysis further revealed the difference in the agent's negotiation plans between the two modalities. In written dialogs, the agent asked most questions on the user's topic of information need compared to the emphasis on understanding the user's motivation observed in spoken conversations. Moreover, the agent mostly used a sequential order of clarification question types while negotiating the need in written dialogs. Thus, the negotiation strategy was more straightforward and without any back and forth transitions between different clarification question types in this modality. In comparison, in spoken dialogs, more complex negotiation strategies were observed in the agent's utterances involving loops between two clarification types. Such loops were observed between clarification questions on the user's preference and anticipating the type of information that the user was after.

Our work on prediction models of clarification questions suggests that prior user's utterance characteristics are important for determining when, within a conversation, a dialog agent should ask a clarification question to the user; however, such characteristics are not so helpful in determining what questions to be asked during a conversation.

Preface

Parts of this dissertation are in various stages of publications by Soumik Mandal.

Acknowledgements

A journey like PhD cannot be completed without taking help from a lot of people. I have been fortunate to meet some of my amazing colleagues, mentors and friends in this journey and I would like to thank everyone of them who guided me during this course and encouraged me to be a better researcher and more importantly, a better person.

First and foremost, I would like to thank my advisor, Dr. Chirag Shah for his constant guidance, patience, mentorship and support for me. His words of wisdom from our first meeting "PhD is a marathon, and not a sprint" still reverberates through me, and writing the final pieces of my thesis, I could not agree more. As my PhD path went through many slopes with its ups and downs, he stood by me in difficult times, pushed me to step out of my comfort zone and encouraged me to take on new challenges which shaped me the person I am today. I consider myself fortunate to have him as my advisor who has been a source of inspiration in my life.

I would like to thank Professor Nick Belkin from my dissertation committee for his invaluable guidance, suggestions and support throughout my PhD. I am grateful to him for offering me opportunities to work in his research projects which introduced me to new research avenues. It has been a privilege to work with him.

I would like to extend my gratitude to my committee members Michel Galley at Microsoft Research and Professor Katherine Ognyanova for their comments and suggestions which was invaluable in my research. They made time to reply to my long emails and provided immense support and advice on my dissertation work despite their busy schedule, that I am truly thankful for. I am also indebted to the Library and Information Science (LIS) department chair, Professor Marie Radford who at times went out of her way to provide support when I needed most.

I would like to extend my thanks to the professors who advised me in numerous ways throughout out my study at Rutgers; Professor Charles Senteio, Professor Kathryn Greene and Michael L. Hecht for giving me opportunities to collaborate in cutting-edge health information

research, Professor Vivek Singh, Professor Kaitlin Costello, and Professor Nina Wacholder whose inputs helped me to grow as a researcher. I am also thankful to PhD Program Director Jennifer Thesis, student coordinators Allison Machiaverna and Danielle Lopez and other support staff at the School of Communication and Information (SC&I) for providing crucial administrative support at all times.

I would like to thank various members of the InfoSeeking laboratory and SC&I PhD cohorts for their assistance and friendship throughout the years. My apologies if I am missing anyone: Dr. Dongho Choi, Dr. Long T. Le, Dr. Matthew Mitsui, Dr. Jiqun Liu, Dr. Souvick Ghosh, Dr. Yiwei Wang, Dr. Jonathan Pulliza, Dr. Manasa Rath, Dr. Ruoyuan Gao, Diana Floegel, Dr. Isha Ghosh, Kevin Albertson, Jiho An, Shannon Taber, Shawon Sarkar, Liz Smith and Catherine Mcgowan. Special thanks to Bruce Duboff, Laura Costello and other colleagues our PhD program for helping me in various steps of dissertation data collection. I am also thankful to my friends and collaborators from Jadavpur University, specially Dr. Braja Gopal Patra, and Dr. Reshma Kar for their friendship and support throughout the past years.

This dissertation is dedicated to my incredible family who always supports and prays for me. I am extremely grateful to my parents for their unconditional love and tremendous support over the years that helped me to navigate through the most difficult times. I am also thankful to my parents-in-law who always supported my decisions, believed in me and encouraged me to do what I do. Last but the most, I owe my deepest gratitude to my wife, Amritapa Banerjee, who was by my side throughout this journey and never let me to give up on my dreams. Without her, this dissertation would not be possible. I thank you for being a loving wife, my best friend and my partner in life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

With the increase in the number of computer users and the use of computers for more and more diverse situations and tasks, the need for tailored information to specific users in specific use contexts arises. Almost routinely, users come across situations where they find their knowledge to be incomplete to achieve some goal or to perform a task, defined as problematic situations (Belkin, 1980). Information retrieval (IR) systems are designed to support such situations, to fulfill the information requirement of the user in need by retrieving appropriate information from information resources (e.g., specific webpages from the internet) which can help her in problem-management. Consequently, IR activity can be defined as obtaining *necessary* information objects *in relevance* to a user's information need from a collection of information resources.

IR systems, in terms of design and functionalities, have experienced remarkable success in recent years, especially in supporting fact-finding and navigational tasks (White, 2016). The prevalent use of commercial search engines, e.g., Google, Bing, Yahoo! Search, and their large user base, is a testament to that success. However, current search systems, retrieval models, and the underlying algorithms still have many limitations when applied in supporting complex tasks involving multiple iterations of search interactions (e.g., planning a trip, comparing and evaluating movie choices) (Hassan Awadallah, Gurrin, Sanderson, & White, 2019). Most of the current search systems rely on the user to express her information need in the form of a keyword or a combination of keywords, known as a query to support the user's information requirement. The query, in turn, is used by the system to separate potentially relevant information objects from the non-relevant ones and, finally, to present the relevant ones to the user. Thus, the current search systems rely on the user's expressed information need, even though at times,

the query expression may represent at best only a compromised form of user's information requirement (Taylor, 1962). In other words, a query-based search system works best when the user is able to express her information needs closely and accurately in a form that the system can understand. This arrangement works quite well for simple tasks; however, for an average user, attaining such stringent requirements becomes problematic with the increase in task complexity. To understand the difficulties of using query-based search systems in complex tasks, let's consider the following two hypothetical task scenarios:

- **Scenario 1:** I am looking for a place to stay in Kolkata between 10th December to 15th December for a conference, and I would like a place that is nearby the conference venue, ITC Sonar.
- **Scenario 2:** I would like to travel to India from the US on 30th Nov for a conference to be held from 10th to 15th December. I want some help in planning my travel itinerary and recommendations for places to visit, as I would like to explore places beyond the conference venue to get the taste of Indian culture and art. However, due to the inclement weather and hazardous air quality, I would like to avoid traveling through and out of north India.

In current search systems, a user with the first task scenario can closely express the information requirement for the task by using the query "*hotels near ITC Sonar Kolkata for 10th to 15th December*" or some variations of it based on similar keywords. Articulating the information need accurately in the second scenario is, however, not so simple or straightforward as the first scenario. First of all, it is unreasonable to expect any ordinary user to describe the complete information need in one search interaction in such complex task scenarios. As per Gricean maxim of quantity (Grice, 1961), in a conversation, the speaker is supposed to contribute *as is required for the current purpose of the exchange* and not supposed to make her contribution more informative than is required. In other words, the user is expected to avoid stating information before the needs for it arises in the conversation. If the user chooses to describe the information required completely in a single search query and manages to do so, the conversation will violate both maxims of quantity and manner (Grice, 1961). Additionally, the query will be too long for any reasonable search system to process for retrieval purposes. Evidence supporting

the ambiguity in user's initial queries offered by researchers, such as Shenouda (1990), who suggested that most information seekers developed their initial search queries by selecting for "more general terms".

Moreover, most of the current search systems lack support for the user to express the information need iteratively in multiple search interactions as required in scenario 2. Thus, in the absence of adequate support for complex tasks as scenario 2, the plausible workaround for a user is to transform the complex task in hand into multiple simpler subtasks, followed by issuing queries for these subtasks individually in some sequence as per the context and finally arrive at some conclusion by piecing together the information acquired from each task. Failure to complete any of these steps may lead to a communication breakdown between the user and the system, which can further cause an inability to express the information need adequately, unresolved information requirement, and failure to complete the task. The system, unaware of these complications, wrongly assumes that the presented information need or the query represents the ground truth of the user's information need. Consequently, the traditional IR systems, e.g., search engines, rely on the user to perform at least two major functions to provide support in complex task scenarios, which are as follows:

- The user must be able to articulate information need closely and accurately in all conditions regardless of the complexity of the task.
- The user must develop a strategy to disentangle the complexity of the task in hand and, if needed, transform the task into a form that can be easily communicated to the system.

However, expecting a user to complete the last function is unrealistic unless the user is an expert in communicating with the system. Thus, being an expert with the system is a prerequisite for getting adequate support in complex task scenarios from current search systems. However, in reality, users of search systems are ordinary humans who are not necessarily experts in the system, who do not always have well-specified needs and often do not have a great desire to learn the conventional means to communicate with the system (Belkin, 1982). Consequently, the best way to enable support for any user's information needs, irrespective of the task nature, is to ensure the interaction with the system as familiar as possible to the user. And one of the most familiar forms of interaction for any ordinary user is a conversation. In other

words, the search system should empower the user to express information needs through conversations. This arrangement, on the one hand, would demand minimal effort from the user to grasp the possibilities and restrictions of the system (familiarity reduces the technical barriers of effective communication between the user and machine). Secondly, such a system can be the best bet to make discovery and representation of need effective. Hence, the dialog model of IR was suggested as a viable alternative of query-based search system in past information seeking and IR research. Evidence supporting the fruitfulness of a conversational approach to the IR process is offered by researchers such as Yerbury and Parker (1998).

In contrast to a query-based IR system that forces the users to express information need in a form the system can understand, a true conversational agent for IR permits mixed-initiative back and forth between the user and the system based on naturally spoken with or without typed interaction, more particularly in the form of well-formulated questions and commands. Questions or requests for information by a user is an element within a dialog-based approach to modeling user-intermediary system interaction (Saracevic, Spink, & Wu, 1997). The system's actions are chosen in response to a model of user's needs within the current conversation, using both short- and long-term knowledge of the user (Radlinski & Craswell, 2017). A proper conversational IR system is expected to be able to perform at least the following functionalities:

- The agent must have the capability to understand and engage in conversations with the user in natural language.
- It must know user preferences, at the very least, for the task in hand.
- Through conversation, the agent must be able to elicit the user's information need.

Recent advancements in the field of AI made it possible to design dialog agents that can mimic human responses. Systems or natural language interfaces that can engage humans in open-domain non-goal-oriented conversations (H. Chen, Liu, Yin, & Tang, 2017), are more commonly known as chatbots, e.g., ELIZA (Weizenbaum, 1966). Development of task or goal-oriented dialog systems designed to assist users with specific tasks, such as recommending products in the e-commerce systems (Sun & Zhang, 2018), are also seeing significant attention from the research community. Task-oriented dialog systems of this kind can consider user preference, however often restricted within narrow domains (Serban, Lowe, Henderson, Charlin,

& Pineau, 2018). Recent advancements in neural network architecture has prompted further research on moving away from the pipeline architecture towards building end-to-end dialog system (Sordoni et al., 2015; Serban, Sordoni, Bengio, Courville, & Pineau, 2016). Such end-to-end dialog systems has been useful for both language models for chatbots and task-oriented dialog systems. Slotted in between these two types are commercial personal assistants, e.g., Apple's SIRI, Amazon's Alexa, Google Assistant etc. These systems are hybrids of task-based and open-domain conversational assistants, meant for fulfilling the information requirement of ordinary users through conversation using some smart devices. Yet they lack capabilities to elicit the users' information need through conversation in complex task scenarios. The lack of support often causes the interaction between the system and the user to appear relatively unnatural (Luger & Sellen, 2016) to the user. Unnatural dialogs or unnatural responses from the conversational agent defeat the primary purpose of modeling the interaction between user and system as a conversation. In conclusion, all three types of systems discussed above fall short in some aspects to be considered as a true conversational IR system.

Furthermore, current research on conversational systems does not fully explain the extent to which systems can rely on users' expressed information need for effective retrieval or recommendation purposes. For example, in scenario 1 as described above, the expressed information need, or the following query, "hotels near ITC Sonar Kolkata for 10^{th} to 15^{th} December", may closely reflect the actual need. However, in scenario 2, where the nature of the task is amorphous and open-ended, the user's expressed information need of recommendations for places to travel in India for the specified dates without traveling through north India may obscure various implicit needs. For example, the user can be looking for recommendations of only those places that have not been visited by him or her before; recommendation of historical places, forts, and monuments instead of any art galleries, etc. Thus, in any complex task cases as scenario 2, the expressed information need may reveal only part(s) of the actual need. Moreover, a user's information need often goes through a transformation while interacting with the system and may change depending on the nature of the task. Therefore, to realize the user's information need, the agent must take steps proactively to clarify the user's information need when the user faces difficulty in articulating it. It is unclear in the existing literature on how the system can identify when the expressed information need underrepresents the user's "actual" need. Further

investigation is needed on how the system can help the user to clarifyherneed.

To address the research gap discussed above, this dissertation attempts to investigate the nature of the dialog between the user and an ideal conversational agent to identify when the expert intermediary (agent) should proactively clarify about the searcher's information "need"; and, how this should be accomplished. The rest of this dissertation is organized as follows. The next section of Chapter 1 provides a statement of the problem of interest, followed by the significance of the proposed work. Chapter 2 reviews relevant literature on information retrieval, information need, previous experiments on simulation of Conversational IR systems, and some of the very recent work on designing dialog systems that can ask clarifying questions to the user. Chapter 3 proposes a methodology to study the problem of interest. Next, a detailed description of the data collection and preparation process is described in Chapter 4. Analysis of these data are presented in Chapter 5. Finally, Chapter 6 discusses the implications of the result followed by limitations of this work and how some of the limitations can be addressed in future research direction.

1.2 Problem Statement

Recent thrust in development of conversational artificial intelligence primarily focuses on the naturalness of the conversation (Ritter, Cherry, & Dolan, 2011), context-sensitive response generation (Sordoni et al., 2015), or developing end-to-end (Sordoni et al., 2015) task-oriented systems (Bordes, Boureau, & Weston, 2017), however, does not provide as much attention on the user side, especially on users information need in different task scenarios.

The few studies in the past that showed interest to the user side had done so more from the personalized recommendation perspective (Sun & Zhang, 2018; J. Li et al., 2016; Al-Rfou, Pickett, Snaider, hsuan Sung, & Strope, 2016). Complex task scenarios are particularly challenging and of interest, as they are expected to require faceted elicitation (Radlinski & Craswell, 2017), and therefore may drive longer and more engaging conversations between the user and the system. To have a sustainable and meaningful discourse in such cases, it is important that the system does not rely solely on what is being said by the user, but also considers what is being not said, or more specifically what aspects of the information need the user is finding

difficult to express. In this regard, the primary goal of the proposed research is here to understand how a conversational IR system can support the user's information need when it is difficult to express accurately. This is achieved by analyzing prior information seeking conversations between users and a simulated dialog agent in complex task scenarios, where the users likely to face difficulties in expressing their actual needs. Specifically, this research analyzes the utterances in conversations between the users and an expert intermediary to examine the the type of clarification questions the expert asks towards the understanding and attainment of the user's information requirement. The typology identified in the process in turn forms the basis of what clarifying questions the system may ask in the future when the user's information need is underrepresented in the expressed form. And lastly, the final part of this research comprises of classification of agent's utterances that are clarification questions and their type in a user and a intermediary pair's conversations in information seeking dialogs.

1.3 Significance of this Study

The research is one of the very few studies the researcher is aware of that explores explicit elicitation of information needs of a human user by a simulated non-human intermediary through conversation with the user. Previous studies that explored the user's information need through user studies had done so with human intermediaries, often an expert with an information resource, e.g., a librarian in library information systems. As per the best of our knowledge, at present, the only other works that explored clarification of user's information need by a non-human intermediary in open-domain conversational information seeking system is by Zamani, Dumais, et al. (2020); Aliannejadi et al. (2019). Aliannejadi et al. (2019) studied the task of asking clarifying questions by using human annotators to generate different clarifying questions for a given query and focused on retrieving a good clarifying question from the human-generated question set, while, Zamani, Dumais, et al. (2020) focused on generating clarifying questions for IR systems. In this work, we investigate the type of clarifying questions for open domain information seeking tasks and identifying patterns of their occurrences in information seeking conversations.

Secondly, one of the significant aspects of the work here is to identify the basic functions

an intelligent intermediary performs to clarify the user's information needs. The goal here is to identify these functions regardless of the task in hand, which can provide a template for designing an intelligent assistant's responses in an open-domain information retrieval system. To this end, conversations from six different task types are considered, and the negotiation of user's information need in dialogs from all task types is analyzed.

Next, one of the crucial aspects that may affect the nature of any conversation is modality. There has been very limited research in the literature on how the modality of conversation affects the negotiation process between the user and the assistant on the latter's information needs (Du & Crestani, 2004). This study compares the functions of both the user's and the agent's utterances between written and spoken dialogs in complex task scenarios, which is a significant contribution.

Finally, this dissertation presents a dialog annotation scheme to analyze conversations of a human-agent pair on information-seeking dialogs. The annotation scheme is domain-independent and captures the major themes of negotiation between the two parties, along with the exchange in conversational roles (Stein & Maier, 1995) that happens during the negotiation process. The annotation process The annotation scheme is based on the previous work of information filters (Taylor, 1967) and the COnversational Role (COR) model by Stein and Maier (1995). The first annotation scheme helps us to understand the aspect of the user's information need and the task in hand is being negotiated in each utterance. On the other hand, the second annotation scheme identifies the utterances where the agent steps out of its usual role of *offer*-ing information and responds with a *request* for more information to the user.

1.4 Summary

This chapter presented the problem statement for this research and explained the motivations. The research problem being addressed is why clarifying user's information need is of importance in conversational informational retrieval systems. The fundamental issues behind the problem are the artificial nature of the interaction, in the form of queries in modern information retrieval systems (e.g., search engines), and lack of capability in eliciting user's information

needs through negotiation in current conversational assistants. Users face difficulties in expressing information needs, especially in complex task scenarios, where negotiation with the conversational agent can help to resolve ambiguities in initial expressions of the needs. This chapter also provides an overview of the proposed approach that will be discussed in detail in the methodology section (Chapter 3).

The next chapter presents a comprehensive literature review focusing on previous research on user's information needs. Next, current advancement in the field of goal-driven dialog systems, and non-goal oriented social dialog agents, or chatbots is discussed. Finally, a summary explaining how the literature informed the design decisions of experiments conducted in this work is provided at the end of the chapter. The design was also influenced by some recent work in conversational IR frameworks that explained various components and capabilities an ideal system should have.

Chapter 2

Literature Review

This chapter provides a review of previous research in the area in which this dissertation is situated. The primary relevant research areas include information needs, interactive information retrieval, dialog models of information retrieval, and research in conversational agents. This literature review in this chapter attempts to summarize 1) the information need and related discussion in the literature, suggested from primarily studying the conversations of a human user and a human intermediary pairs, next 2) the conversational or dialog models of IR proposed in the literature of interactive information retrieval, followed by 3) advancement in the development of goal-driven dialog systems designed for specific tasks (e.g., ticket booking), and open-domain dialog agents, and how that informed the research direction in conversational information retrieval systems, finally 4) some recent experiments conducted in conversational IR framework that explained various components and capabilities an ideal system should have. Additional details on how the previous work informed the experiment designs in this dissertation is provided. Figure 2.1 further illustrates where the proposed dissertation study is situated in relevant research domains.

2.1 Information Need

Information need represents the start state for someone seeking information, which involves information search using an IR system. There are two dominant perspectives on information need found on the literature (Cole, 2011). The computer science and system side focused perspective is that the user needs to find an answer to a well-defined question which could be formulated into a query to the system. In this approach, an IR system is an information or answer-retrieval system, designed for the user to find a suitable answer. Whereas, user-oriented theory and

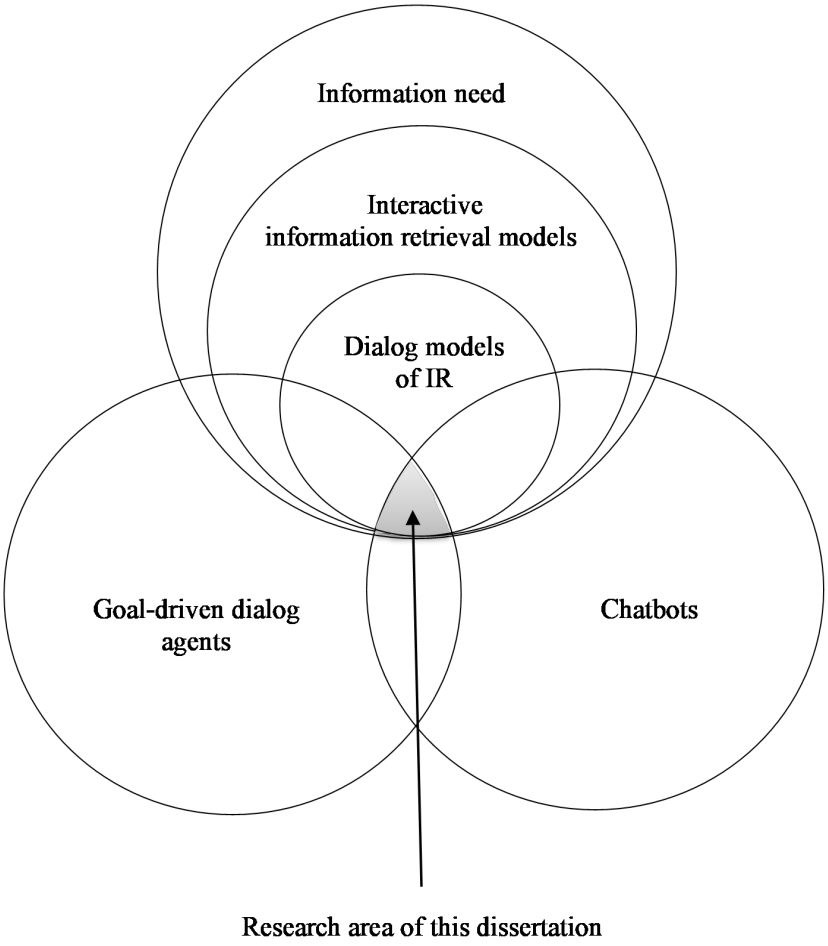


Figure 2.1: Position of the proposed dissertation study in research areas

information science's model of information need (Taylor, 1967) suggested it to be a "black-box", as unknowable and non specifiable by the user and thus, unable to represent as a query to the information system. In this direction, Taylor (1962) explicated the nature of "information need" in his attempt to provide a unified framework that focused on user needs and preferences in evaluating and designing information systems, coined as Value-added model (Taylor, 1982). Analyzing the transcripts of conversations between human users and a human intermediary (e.g., librarian or "information specialist"), Taylor suggested how an inquirer's information need could go through a sequential transformation in various levels. This transformation of

information need happened either consciously or unconsciously through negotiation with the information system. The transformation of information need of user happens in four levels as per Taylor.

1. **Visceral need or *Q1*:** At the very first stage, the user realizes conscious and unconscious need for information not existing in his remembered experience. Visceral need represents the actual yet unspecifiable need for information of the user. In terms of query, this level might be expressed by "ideal question" (Taylor, 1962) — the question which would bring from the ideal system exactly what the inquirer needs, if the need could be stated.

2. **Conscious need or *Q2*:** In progression of the visceral need, the investigator transforms the actual need into a conscious mental description of the need as an ill-defined question. Taylor characterized the user's need at this stage as ambiguous in its descriptions and statements. The enquirer may discuss her ill-defined question with someone with the hope that the person will understand the ambiguities, and the dialog will help in clearing such ambiguities.

3. **Formalized need or *Q3*:** From *Q2*, the inquirer transforms her ill-defined mental description of information need into a rational statement in *Q3*. Unlike the representation in *Q2*, this statement is a rational and unambiguous description of the inquirer's needs. This transformation happens only through a dialog with a partner.

4. **Compromised need or *Q4*:** Finally, the compromised need represents the question as presented to the information system. From *Q3* to *Q4*, the inquirer recasts the question in anticipation of what is expected to get out of the system. Thus, the inquirer's expectation and prior experience with the system may affect the transformation to *Q4*.

Taylor's view on information need was seminally presented by (Belkin, 1980) in "ASK hypothesis" of the users' "anomalous state of knowledge" which they seek to repair. Referring Taylor, Ingwersen (1982) explained what he called "the label effect": the users "compromise" their need in the form of a label which consisted of one or several concepts out of the context which formed their real, formalized need. Due to the unspecifiable nature of information need (*Q1*), literature in information science devoted to studying adjacent or surrogate concepts

e.g., problematic situation and task, information transfer. For example, Belkin (1984) studied information transfer as the dynamic interaction among the user; a knowledge resource; and the intermediary mechanism between these two components. In this process, user initiates the interaction with the system because of some problem, goals, gap in knowledge (Belkin, 1980) whose management or realization may be furthered by information obtained from the appropriate knowledge resource. Thus, the intermediary's job is mediating between the user's desires, requirements, knowledge, and the knowledge resource's contents, and representation, so that if texts appropriate to the user's situation are in the knowledge resource, they can be brought to the user's attention. In other words, the view of information transfer envisages information retrieval system as a knowledge formulation or acquisition system. Further studies on this view suggests that information need is a secondary need and must therefore be contextualized in the user's situation and task in order to be meaningful (Wilson, 1981).

Studying the user's information need in the context of a task (Schamber, Eisenberg, & Nilan, 1990) suggests that the user's information need evolves over the course of the task in the hand, with the user's knowledge structure constantly evolving from the beginning to the end of the task as the user gathers more knowledge about the topic area and the task itself. According to this view, the user narrows in on the message of the task; the information need becomes more specific more focused (Kuhlthau, 2004). Informed by these previous work, this dissertation on conversational IR system also focuses on users' information need in the context of a task.

Further research on information need has shown a great deal of interest to explicate the nature of Taylor's information need levels (Qs) and its relation with the task. In this direction, Cole (2011) suggested that the information need manifested itself to the user in different ways over the course of performing a task. This work on information need differentiated between the phases to the performance of user's information-based task and the underlying information need. As per Cole, there is only one information need instantiation to its deepest $Q1$ level in the focusing stage, which stays the same (unless the user abandons the task or radically changes direction for some reason). Thus, all transformation in information need happens before the instantiation in the pre focus stage. In the post focus stage of performing the task, therefore, the information need stays the same as the user conducts command-type information searches. On the other hand, Ruthven worked on classifying information need levels from user

generated questions. Studying the postings in major internet discussion groups, Ruthven's experiments suggested significant linguistic differences between the expressions of *Q2* and *Q3* levels of need. This work further suggested expressions of conscious needs to be more emotional in tone involving more sensory perception with different temporal dimensions than descriptions of formalized needs. Taylor's information need model and the follow-up research are of importance here on at least two accounts. First, based on the findings, it is evident that the intermediary has a crucial role to play towards the realization of a users' actual need (*Q2* or *Q3*) beyond just retrieving results as per the user's request. Secondly, Taylor's framework described the interaction as having a reciprocal influence between the inquirer and the system, which is of interest here. Such reciprocity in influence suggests at least to some extent, the interactions between the user and the system as mixed-initiative, which makes it compatible for analyzing conversations in the context of conversational IR. While this dissertation does not explore the user's information need level in the conversation, however, it is assumed that the information need is not static, rather it evolves and becomes more specific through conversation with the dialog agent.

To be noted here, there exists a host of empirical studies in the information science literature on information needs and uses, which was reviewed by, for instance, Paisley (1968), Dervin and Nilan (1986) and Wilson (1994). However, the focus of these reviews was theorizing human information behavior, rather than advancing the design of conversational systems. Broader information seeking models such as information behavior (Wilson, 1999), ISP (Marchionini, 1997) that used the concept of information need to model human information behavior or information seeking strategies are useful in discussion of information need; however, they will not be covered in this discussion.

2.2 Negotiation of Information Need

There has been limited research on under standing interactions between a user and an intermediary that ventures beyond the topic or domain of the information need and the associated task that triggered the dialog. In this direction, Taylor (1967) analyzed the conversation between an expert intermediary (information specialist) and the user (inquirer) as a process, abstracting the

topic and domain of the user's information need. Taylor's characterization of the intermediary's negotiation actions towards eliciting the inquirer's information need as a form of directed and structured process is an important step and at the core to this dissertation work as such characterization helps us to study the clarification questions to be asked by the intermediary as part of the structured negotiation process and independent of the task domain.

Taylor decomposed the negotiation process into five filters through which the user's inquiry was processed by the intermediary to generate appropriate responses throughout the conversation. The term "filters" was used to denote the five components, rather than calling them simply "codes" or "labels" to highlight their roles in the negotiation process, i.e., to refine the user's earlier questions or expression(s) of information need with the aim to retrieve appropriate information to aid the user in her task. As per Taylor (1967), these five filters are:

- **Determination of the subject (Topic):** This filter determines the limits and provides some delineation of the information space. For example, after applying this filer on the inquiry posed by the user, the system may ask follow-up questions 'Is this what you mean' or 'Is this in the ballpark' (Taylor, 1967) as a response. Therefore, the job of this filter is to classify users' topics of interest.

- **Motivation and objective of the Inquirer (Motivation):** The second filter or category of information negotiation is related to: why the inquirer wants this information; what is objective the objective; what motivates the user to look for this information etc. It may further distill the subject or may even alter the meaning of the entire inquiry. In this stage of negotiation, the system tries to ascertain the cause of difficulties faced by the inquirers to express his need.

- **Personal background of the inquirer (Preference):** The third filter or category of information that affects the negotiation process has to do with the personal background of the inquirer. This includes, but may not be limited to, the following questions: What is his background? Has he used the system before? What's the relationship between his current inquiry and what he already knows etc. Answers to these types of questions help the system to determine the urgency, the negotiation strategy, level or depth of any dialog, and the critical acceptance of search results etc. Thus, the filter is associated to deduce

user's preference.

- **Relationship of inquiry description to file organization (Search strategy):** Through this filter, the intermediary or the information specialist interpret and restructure the user's inquiry that best fits for effective retrieval purpose. In the symbolism discussed earlier, in complex task scenarios, here the system constructs or suggests a Q4, or a set of Q4s, so that the total content of the system can be searched efficiently.

- **What kind of answer will the inquirer accept (Anticipation):** When an inquirer approaches the information system, he has some picture in mind as to what he expects the information to look like, e.g., it's specificity, format, modality etc., which in some way shapes his information need to Q4 from Q3. The system, through negotiation, tries to alter the inquirer's a priori picture of what it is he expects, therefore making the user more embracing to information that can fulfill his Q3 and even Q2.

In the context of information need labels discussed earlier, the five filters (Taylor, 1967) represents the broad functions the expert intermediary performs to negotiate with the user to support her compromised need (*Q4*); to work with the inquirer back to the formalized need (*Q3*), possibly even to the conscious need (*Q2*); and then to translate the need into a useful search strategy. Thus, with respect to automated dialog-based IR system, these five filters represent the clarification strategies the system should adopt to navigate the negotiations with the user on her information need. This dissertation work on supporting the user's information need is based on Taylor's conceptual model of information needs and filters, but looks at the negotiation process in complex tasks that an average user may aim to accomplish with an "ideal" dialog based IR system.

2.3 Conversational or Dialog Models of IR

As per Cambridge English dictionary, the word conversation means "an informal, usually private, talk in which two or more people exchange thoughts, feelings, or ideas, or in which news or information is given or discussed." From this definition of a conversation, a few important characteristics of the interaction emerge which should be considered when discussing

conversations or dialogs in IR. In particular, the idea of "exchange" is of interest here, which suggests that the initiative may belong to both parties at different points in the conversation. Hence, a conversation is a mixed-initiative interaction; with reference to IR which means it is not enough to just answer the searcher's questions, but to have a meaningful conversation, the intermediary i.e., the system must take the initiative at appropriate points during the course of interaction. Following this definition, Radlinski and Craswell (2017) defined the conversational search system as *a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user*. The research in this dissertation embraces the definition of conversational IR by Radlinski and Craswell (2017), and discusses only those models of discourse analysis of human-intermediary dialogs that were specifically developed to drive a human-machine conversation in this chapter. Discourse between the user and the intermediary pair as they interact while negotiating the user's information need, serves the function of user modeling (Saracevic et al., 1997). Previous experiments suggested various ways to design IR system through dialog, e.g., (Oddy, 1977), (Sitter & Stein, 1992), and (Belkin, Cool, Stein, & Thiel, 1995). Here the last two models are discussed in detail as they are based on similar assumptions and can be used as templates for generating response for a conversational IR system. Both models were developed for two-party conversations, where the user (information seeker) and the system (information provider) were treated as dialog partners cooperating with each other to fulfill the user's information need. Each dialog was considered as a well-formed sequence of dialog steps. However, despite their similarities, the two models had important differences, noticeably in identification of dialog steps.

Sitter and Stein (1992) developed the COnversational Roles (COR) model based on the dialog steps or dialog acts as a general model for information-seeking dialog. As shown in Figure 2.2, each node represents the dialog state, and an edge between two nodes represents a dialog act or dialog contribution. This arrangement is analogous to a speech act suggested by the theory of speech acts (Searle, 1985). By executing a dialog step or dialog act, the actor (A) undertakes a social role (e.g., request), and in turn, assigns the complementary role (accept) to the dialog partner (B). Consequently, these dialog steps are identifiable from the expectation

the actor sets on his partner's future behavioral responses. Thus, the COR model identifies the structure of any dialog session in terms of a sequence of underlying dialog acts. A basic schema is identified by analyzing enough dialog sessions. This schema is stored and modeled in the form of a transition network (see Figure 2.2) in the system. The transition network, in turn, can be used to generate responses for the system in future sessions to move forward the conversation.

The bold line in the network represents the optimal path to fulfill the goal of current dialog in COR model. Thus, Sitter and Stein's model identifies the current state in the dialog network based on the prior dialog acts in the conversations to predict the next dialog act that is acceptable as a response. The COR model uses the dialog acts to create a structure for the conversation, and, abstracts away from the task in hand as it does not rely on the domain-dependent knowledge for response generation. Consequently, the model lacks support in identifying global information seeking strategies, in accordance with the task. Nonetheless, the COR model is still useful to analyze when the user takes a conversational role and in turn, expects the system to take initiative in a dialog act during a conversation.

Sitter and Stein (1992)'s model was further enhanced into broader information seeking structures in scripts from Belkin et al. (1995). The latter argued that the user might employ different information seeking strategies (ISS) to fulfill different types of information need. Consequently, a dialog-based system should provide support multiple types of interaction as needed for each ISS. Further, the system should be able to suggest the user which interaction type might be appropriate to fulfill the information need for the task. Belkin et al. (1995) proposed a specific type of dialog or sequence of dialog acts as a script. During a dialog session, multiple scripts could be combined following a dialog plan to guide the user through interaction. This dialog plan could, in turn, be derived from previous cases of successful retrieval sessions. Thus, Belkin et al. (1995)'s model used previously stored instances of interaction patterns or sequence of interaction from successful retrieval sessions to suggest the current interaction type.

Both COR model and the scripts support a form of prediction for the next dialog act in a conversation that is appropriate and necessary following on from a previous move. Both models suggests that to provide effective interaction support, a conversational IR system should

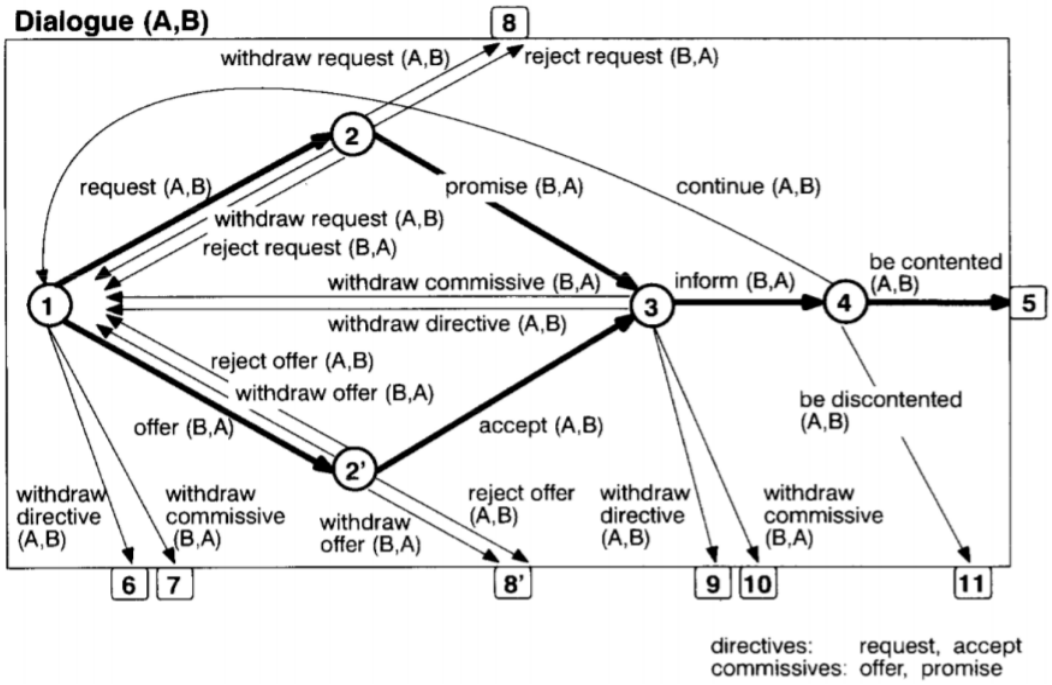


Figure 2.2: Conversational Role (COR) model (Sitter and Stein,1992)

analyze and store a comprehensive list of previous dialog sessions and the associated interaction patterns, index these patterns and use this information along with the the dialog act from the current user utterances to provide an appropriate response to the ongoing user's act. Thus, irrespective of their differences, both models suggested identification of current dialog step in conjunction with a global plan as prerequisites of prediction of which kind of interaction would be necessary for the next step in both models. This research uses the COR model to analyze the dialog acts and to identify turn-taking of initiatives from previously recorded conversations. The global plan for the dialog is derived from the structure of the negotiation process.

2.4 Goal or Non-goal Driven Dialog Agents

It is worth mentioning here that not all the dialog-based systems are conversational IR systems. Conversation or discourse analysis has received attention from multiple disciplines, from philosophy to cognitive-psychology, from computational linguistics to artificial intelligence (AI). Here, the last two disciplines are of interest, research on discourse analysis from these two

fields shares some common goal with conversational IR, i.e., to drive human-computer communication. The result of these efforts is two prongs. On one hand, a host of natural language interfaces, e.g., ELIZA (Weizenbaum, 1966), CSIEC (Jia, 2009) have flourished lately, which sometimes are referred as chatbots. On the other hand, the research has delivered a broad range of task-based communication programs such as airline or restaurant booking (Bordes et al., 2017).

Investigation in the fields of IR and natural language processing (NLP) has studied various aspects of creating automated dialog agents. Early approaches on automated dialog systems focused on developing rule-based conversational systems, e.g., DARPA Communicator (Walker, Passonneau, & Boland, 2001) and Williams et al. (2014), while the other direction in this line of research studied natural language understanding approaches (Hemphill, Godfrey, & Doddington, 1990; He & Young, 2005). Recent studies on developing automated agents to perform some actions in goal-driven systems have suggested that modeling human-computer dialog as plan-based (Bruce, 1975; P. R. Cohen & Perrault, 1979; Allen & Perrault, 1980) collaborative behavior is useful, where the dialog agent's job should be to recognize its partner's plans to achieve the joint goal and then to perform the actions to facilitate them.

In contrast to goal-based dialog systems, chatbots are designed to "chit-chat" about arbitrary topics, as they provide methods for imitating a dialog but not for participating in one (P. R. Cohen, 2018). Recent studies in natural language interfaces found end-to-end neural networks trained on large corpora to be useful (Sordoni et al., 2015) for designing systems that can mimic human responses in open domain dialogs. Neural approaches also found applications in conversational recommendation space (Dodge et al., 2016). Sun and Zhang (2018) utilized a semi-structured user query and a combination of facet-value pairs to represent a conversation history and proposed a deep reinforcement learning framework based personalized conversational recommender system. In the movie recommendation domain, earlier works focused on human-human movie recommendation (Johansson, 2004). Recently, Dalton, Ajayi, and Main (2018) demonstrated a conversational movie recommender system based on Google's DialogFlow[1].

[1] https://dialogflow.com/

Although earlier dialog systems were designed to converse with users primarily via text inputs, the voice based communication has gained more prominence lately due to recent developments of accurate automated speech recognition (ASR) systems. ASR and task-based communication systems coupled with text-to-speech synthesis together, in turn, have been further developed into voice-based intelligent personal assistant systems such as Alexa, Ok Google, Cortana, and a host of associated appliances, e.g., Echo, Google Home, etc. Indeed, the recent surge of interest in Conversational IR appears to be somewhat influenced by the easy availability of such voice-based intelligent assistants and appliances. However, in reality, these assistants work more as task-based dialog systems (*Alexa, Open magic 98.3!*) and are very limited as information retrieval systems (*Alexa! What is the capital of Colombia?*). However, neither the chatbots nor the task-based communication devices (or voice-based personal assistants) represent true conversational IR systems at present. The chatbots are not necessarily concerned with the user's intention, goal or problem, while the task-based communication systems are designed to handle very limited and well-defined tasks in a closed domain. Such constraints are unrealistic for any IR system. Hence, despite their shared part-common goal with the conversation IR, both systems have some major underlying differences with the latter. Nonetheless, the research on dialog systems have implications for the conversational IR research. For example, a spoken conversational IR system also depends on accurate ASR and natural language understanding models for interpreting the user's spoken questions or queries.

In goal-driven systems that aim to assist users with specific tasks, the challenge is to understand the user's request and query a database of the task (e.g., flight schedule information) accordingly. Intelligent dialog agents in this context found applications in the domain of flight (Hemphill et al., 1990) and train information (Aust, Oerder, Seide, & Steinbiss, 1995). Intention detection and slot filling have been the prevalent approach in goal-driven system development. In this approach, the core problem is represented as to fill out required and optional attribute-values (termed "slots") in an action schema or "frame". For example, the list of attributes in a slot may include the date, time, and the number of people for a restaurant reservation task. If an attribute or argument is missing, the system prompts the user to supply it. More recent research in task-based dialog agents relies on an architecture typically consisting of a natural language understanding module, state tracking, a dialog policy and a

response generation module (Y.-N. Chen, Celikyilmaz, & Hakkani-Tur, 2018). Each of these modules is often implemented using neural network based architecture and optimized individually (Young, Gašić, Thomson, & Williams, 2013). Based on these studies on both goal-driven and non-goal driven system designs, neural network-based architecture seems to produce best prediction models in dialog generation/selection tasks.

2.5 Conversational Search System

Most of IR and Web search research focuses on the analysis of keyword and boolean queries (Spink & Saracevic, 1997). However, conversational approach to IR enables the user to express the information need in more natural forms such as questions. Therefore, studying the characteristics of users queries in other formats, such as question is an important and growing field for the development of more effective search system (Spink & Ozmultu, 2002). Recent advances in the development of conversational agents attracted research in various aspects of conversational information access (Aliannejadi, Zamani, Crestani, & Croft, 2018; Yan, Song, & Wu, 2016). Due to lack of true conversational search systems, most of the previous studies used either role-playing (Trippas, Spina, Cavedon, Joho, & Sanderson, 2018) or Wizard-of-Oz settings (Trippas, Spina, Sanderson, & Cavedon, 2015; X. J. Yuan & Sa, 2017; Vtyurina, Savenkov, Agichtein, & Clarke, 2017; Jung et al., 2019; Avula & Arguello, 2020). In Wizard-of-Oz experiments, a wizard (i.e., a human and often an expert in the search system) simulates the activities of an automated conversational agent. Users (i.e., information seekers) unaware of this arrangement, communicates their information need to the wizard in order to complete their information seeking tasks. The human "wizard" simulating the critical role of the "ideal" conversational search system is selected carefully for the experiment and needs to have some experience in dealing users' information need (e.g., call center operator, customer care representative, librarian). The discourse between the user and the agent is analyzed by the researchers to elicit the nature or the topic of interest related to a "true" conversational search system. For example, Thomas, McDuff, Czerwinski, and Craswell (2017) studied task completion in conversational search system using Wizard-of-Oz experiments. Radlinski, Balog, Byrne, and Krishnamoorthi (2019) focused on conversational preference elicitation. Ghosh (2019) proposed to use similar set up to

study the presentation of retrieved results in a conversational search system. Avula and Arguello (2020) used Wizard-of-Oz experiments to study how a conversational search system should take the initiative when engaging with users during collaborative search. Conversation modality's effect on the search system was also studied employing Wizard-of-Oz experiments. (Avula, Chadwick, Arguello, & Capra, 2018) used Slack, a popular messaging app to simulate collaborative information-seeking tasks with the help of a conversational agent in written (text) mode. In comparison, in role-playing method, both parties in the conversation are aware that they are communicating with a human being. Thus, Trippas et al. (2018) studied conversations of pairs of participants, one of whom was assigned the role of "user"; the other the role of "agent" to identify commonly used interactions for spoken conversational search.

In study of conversational search, Radlinski and Craswell (2017) proposed a theoretical framework highlighting the importance of multi-turn interactions between users and the conversational agent for narrowing down the users' specific information needs. Another line of research in the context of conversational IR systems analyzed data to understand how users would interact with voice-only systems (Spina, Trippas, Cavedon, & Sanderson, 2017). X. Yuan, Belkin, Jordan, and Dumas (2011) found that task type had a significant effect on users' query behavior in a spoken language interface. Similarly, studies that compared written versus spoken queries found that the user might express higher satisfaction in terms of the naturalness of the system for spoken queries (Crestani & Du, 2006). This finding was also confirmed by (Yan et al., 2016), where users initiated significantly fewer but longer queries for both interpretive and exploratory tasks in the spoken language interface than in the textual interface. (Kiesel, Bahrami, Stein, Anand, & Hagen, 2018) studied the impact of voice query clarification on user satisfaction and found that users like to be prompted for clarification in the spoken conversational system. The next section discussed previous research on conversational systems that were directly related to clarification questions to be asked by an intelligent assistant.

2.6 Asking Clarifying Questions

Research on clarifying questions attracted significant attention in the past from both natural language processing (Stoyanchev, Liu, & Hirschberg, 2014; Rao & Daumé III, 2018, 2019) and

information retrieval community (Aliannejadi et al., 2019). Earlier natural language processing research on clarifying questions leveraged community question answering (CQA) sites (Braslavski, Savenkov, Agichtein, & Dubatovka, 2017), e.g., Yahoo! Answers, Stack Exchange, Quora etc., in pursuit of machine reading of comprehension literature (Duan, Tang, Chen, & Zhou, 2017; Heilman & Smith, 2010). In CQA sites, a user (information seeker) posts the information need typically in the form of a question, to which other users (information provider) are expected to respond with appropriate answers or requests for further clarifications when the question has ambiguity. Thus, dialogs between the users on a CQA website can be a rich repository of information-seeking interactions. Trienes and Balog (2019) analyzed CQA posts to classify unclear posts that require further clarification. Braslavski et al. (2017) suggested that studying the type of clarification questions asked by CQA users in association with their overall interaction behavior could have implications for search as a dialog paradigm. Further analysis on the intent of each utterance (Shah, Oh, & Oh, 2009; Choi, Kitzie, & Shah, 2012), including clarifying question types in human-generated dialogs suggested highly recurring patterns in user intent (Qu et al., 2018) during an information-seeking process on CQA websites. The other research direction on clarifying questions in CQA websites looked at generating (Duan et al., 2017; Zhou et al., 2017) and ranking (Heilman & Smith, 2010) questions whose answers appeared in a given passage. In this direction past research suggested neural models e.g., convolutional network (Duan et al., 2017) to be useful in generating questions.

Apart from CQA, asking clarifying questions was also studied in other contexts of language processing, such as open-domain question answering systems (De Boni & Manandhar, 2003, 2005), dialog systems (Lurcock, Vlugter, & Knott, 2004; Quintano & Rodrigues, 2008) and automatic speech recognition systems (Stoyanchev et al., 2014). However, these contexts are fundamentally different from asking a question to clarify the user's information need due to difference in objectivity of the systems, and therefore, their findings are not discusses here. Asking clarifying questions to point out missing information in a passage has also received significant attention from natural language processing research, which is a more relevant line of work for this dissertation. In this direction, Rao and Daumé III (2019) proposed an attention-based sequence model for generating clarifying questions for eliciting missing information in a closed-domain. Reinforcement learning approaches to optimize a utility function based on the

value added by the potential response to the clarifying question was also found to be useful (Rao & Daumé III, 2018).

The other method of simulating clarification questions by the intermediary in users' information-seeking conversations that gained prominence in more recent experiments used experimental systems to simulate the conversational search system. For example, Aliannejadi et al. (2019) designed a system that could ask for clarification to the user if the system determined that it needed more information to resolve the ambiguity in its response. The system's questions were aimed to clarify ambiguous, faceted or incomplete queries (Vtyurina et al., 2017). In this system, two high dimensional neural network based modules were used, one designed to retrieve a set of questions (Question Generation module) with the focus on maximizing the recall for a given query, and the other was to select a set of questions (Question Selection module) retrieved by the previous module with the aim of maximizing the precision at the top of the ranked list of questions. For a given query, Aliannejadi et al. (2019)'s system decided its confidence in the retrieved documents against the user's query in order to decide between whether to present the result, or ask a clarifying question. In cases where the system determined not have enough confidence about the quality of the result, it passed the query and the context (i.e., earlier utterances exchanged by the user and the system) to the Question Generation module to generate a set of clarifying questions, from which a Question Selection module determined the best question to be presented to the user. The questions dataset was generated from a Human Intelligence Task (HIT) employing crowd-sourced workers that was designed to collect a reasonable set of questions covering multiple facets of every topic used in the experiment. A workflow diagram of asking clarification question process in Aliannejadi et al. (2019) is given in Figure 2.3. The result of the experiments by Aliannejadi et al. (2019) suggested that asking only one good question could lead to over 170% retrieval performance improvement in terms of Precision at 1 (P@1). Despite the positive outcome of this heuristic based approach adopted here, Aliannejadi et al.'s experiment only covered elicitation of the user's information need when some facets of the query was missing, and lacked explanation of the relationship between the situations when user struggling to explain her information need, and the system struggling to understand the user's "intent", and therefore, chose to generate clarification questions.

In a similar approach, Zamani, Dumais, et al. (2020) proposed another experimental set up

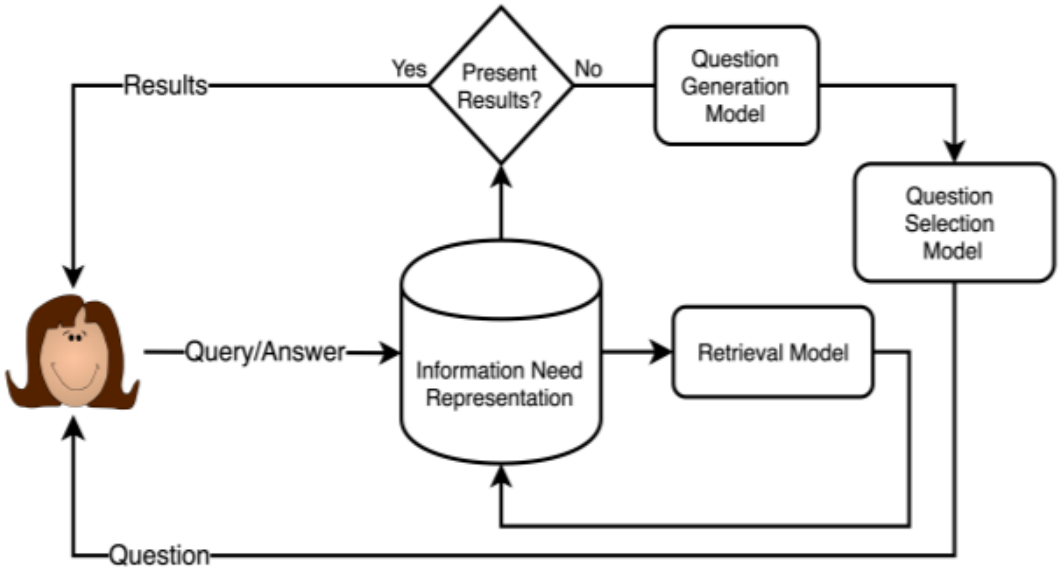


Figure 2.3: Workflow diagram for asking clarifying questions in Aliannejadi et al. (2019).

for studying users' interactions with the clarification questions in more traditional query based search. They used a modified commercial search engine interface, Bing, with a clarification pane that included representations of both clarifying questions and candidate answers. The users' interactions with the clarification pane were recorded. The interface of the experimental system in (Zamani, Dumais, et al., 2020) is given in Figure 2.4. The researchers also conducted user studies interviews to collect qualitative data on user experience with the clarification pane. Their analysis suggested that even with non-relevant clarification questions, the participants did not felt that the search experience was degraded. This experiments towards generating questions suggested that a slot-filling approach could be useful to generate clarifying questions on user preference, and topic of interest.

The inventory of clarification questions used in the clarification pane were based on the taxonomy of clarification types adopted by analyzing a large scale Bing query log and query reformulations. Thus, the types of clarification questions were identified in this experiment were based on the reformulation behavior in traditional Web search engine and hence, put a restriction on the type of clarification questions allowed in the system. Additionally, Zamani, Dumais, et al. (2020)'s experiment simulated user's interaction with the clarification questions only in the text or written mode of conversation. Hence, despite their implementation of the

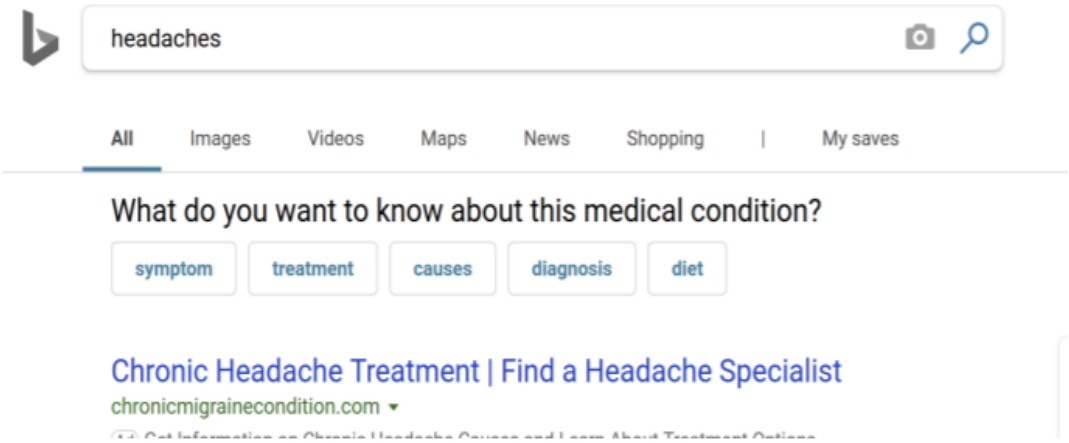


Figure 2.4: Interface of the experimental search system with clarification pane in Zamani, Dumais, et al. (2020)

near-real systems, both systems had limitations in generating "ideal" dialogs between humans and conversational search systems that could be used to generate a taxonomy of clarification questions. Nonetheless, the experiments mentioned above highlighted the importance of asking clarifying questions in conversational search systems.

In the realm of IR, Allan (2004) organized HARD Track at TREC 2004, which allowed the participants to submit clarifications. More specifically, the participants could submit a form containing human generated clarifying questions in addition to their submission run. Radlinski and Craswell (2017) raised the importance of asking for clarification in their theoretical framework for conversational search. (L. Yang, Zamani, Zhang, Guo, & Croft, 2017) proposed a neural matching model on Quora data and Ubuntu chat logs for retrieving the next question in conversation. Kiesel et al. (2018) analyzed the impact of voice query clarification on user satisfaction and suggested that users preferred to be prompted for clarification. In more recent studies, Zamani, Lueck, et al. (2020) analyzed users' click behavior on the queries and interactions with clarification pane collected from the search logs of a major commercial web search engine, Bing. Their analysis suggested that the existing click models that were primarily designed for web search did not perform as per expectation for search clarification. Hashemi, Zamani, and Croft (2020) explored the utilization of user responses to clarifying questions in a Transformer network (Vaswani et al., 2017). Zamani, Dumais, et al. (2020) studied the task of

generating clarifying questions for open-domain information retrieval and suggested a taxonomy of clarification questions from analysis of Bing search logs with clarification pane. While this dissertation work was largely influenced by the work of Zamani, Dumais, et al. (2020), this dissertation focused on creating a robust set of clarification question types that could be deployed in future conversational search systems in multiple modalities.

2.7 Research on Query Suggestion, Reformulation, Disambiguation

Previous research on query suggestion, query reformulation and disambiguation also have implications for clarifying questions in conversational search system. Coden, Gruhl, Lewis, and Mendes (2015) studied the task of asking clarifying questions for entity disambiguation mostly in the form of "did you mean A or B?". While their approach was useful for entity disambiguation, however had limitations when applied to a broad spectrum of queries including faceted queries. In contrast, query suggestion in IR tasks is envisioned as recommending a set of possible queries likely to be searched by the user. Research on query suggestion is of interest as query suggestion and reformulation could also be posed as clarification questions in a conversational system, e.g., "did you mean A?", "did you mean A instead of B?". A commonly used query suggestion method proposed in the literature is to find similar queries in search logs and use those queries as suggestions for each other (Wen, Nie, & Zhang, 2001; Baeza-Yates, Hurtado, & Mendoza, 2004). In line with this direction, the previous experiments suggested mining pairs of adjacent or co-occurred queries in the same query sessions to be useful (Huang, Chien, & Oyang, 2003; Jones, Rey, Madani, & Greiner, 2006). When one of the queries from the pair is encountered, the system uses the other as a suggestion. Apart from query suggestion and reformulation, query expansion techniques is the other most common form of support found in most of the current search systems. When the query issued by the user is too restrictive for a direct match or any retrieval, the system usually supplements the original query with additional terms to increase the recall (Efthimiadis, 1996). Previous studies on query expansion techniques suggested that expansion by adding terms that were most similar to the concept of the query (Fonseca, Golgher, Pôssas, Ribeiro-Neto, & Ziviani, 2005), rather than the query terms itself to be more useful (Qiu & Frei, 1993). Experiments on query expansion techniques using

lexical-semantic relations found little difference in retrieval effectiveness if the original queries were relatively complete representations of the information being sought; however for less-well developed queries retrieval effectiveness significantly improved when the expansion terms were chosen from hand-picked concepts (Voorhees, 1994). Previous work on users' query behavior (Hassan, Shi, Craswell, & Ramsey, 2013; Liu, Gwizdka, Liu, Xu, & Belkin, 2010) suggested that an overlapping query syntax between two consecutive queries in a search session could be an indication of low satisfaction with the first query, and users were likely to describe their intents more clearly in the second query. Result diversification and personalizing emerged as the key components for query suggestion (Jiang, Leung, Yang, & Ng, 2015), especially in small-screen devices. Query reformulation behavior was also extensively used to study various IR tasks, such as query suggestion and query auto-completion (Cai & De Rijke, 2016; Mitra, 2015; H. Yang, Guan, & Zhang, 2015). In this direction, Boldi et al. (2008) used query reformulation data to construct a query flow graph. (Mitsui, Liu, Belkin, & Shah, 2017) used query reformulation types (Rha, Mitsui, Belkin, & Shah, 2016) along with users' search behavior to predict users' intentions in information seeking sessions. Diaz (2016) studied query reformulation as a discrete optimization problem by constructing an unweighted graph of queries. Szpektor, Gionis, and Maarek (2011) used entity type information in association with query reformulation behavior to improve the query suggestion quality in tail queries. The usefulness of query suggestion and reformulation in IR tasks from previous studies suggested that query suggestion as clarification questions could be worth exploring in conversational search system.

Much work on asking questions was done on conversational recommendation space. By asking questions about various attributes of the target item the system can provide more accurate recommendation (Sun & Zhang, 2018; Zhang, Chen, Ai, Yang, & Croft, 2018) to the user. For example, Christakopoulou, Radlinski, and Hofmann (2016) designed a restaurant recommender system that could interact with users to collect information about their preferences before recommending the venue. The unique challenges and techniques used for identifying different facets of search queries and generating clarifying questions in response to user queries are fundamentally different from those reviewed in this section. However, inspired by the findings from previous research, this dissertation analyzed the clarification questions and query suggestions asked by an "ideal" conversational assistant in eliciting the user's information need.

2.8 Summary

This chapter reviewed the previous work from related disciplines that influenced this research and explained in respect to previous research where the proposed work is situated. The review primarily covered literature on information needs from information seeking models, dialog models of information retrieval from interactive information retrieval research, and research on the development of goal and non-goal driven dialog agents from computational linguistics. Finally, some recent work on conversational IR system that explored various aspects of how an ideal conversational IR system should act was reviewed. Among information need models, Taylor (1962)'s information needs model is influential as it suggests the gradual transformation of user's information needs throughout dialog with the intermediary. Besides, the model highlights the importance of the intermediary's role in the negotiation process. Taylor (1967)'s proposition for the intermediary's role systemizes the negotiation on user's information need as a structured process, consisting of smaller identifiable functions (filters), while abstracting the task or domain of information need being discussed in the conversation. Previous studies on conversational search that were discussed earlier in this chapter highlighted the usefulness of Wizard-of-Oz experiments to simulate an "ideal" system. While experimental systems (Zamani, Dumais, et al., 2020) have come up in most recent studies, these systems have limitations to be considered as a real-time "ideal" conversational search system.

Among dialog models of IR, the COR model (Sitter & Stein, 1992) is highlighted in this chapter as it provides a domain-independent annotation scheme to help understand the nature of the conversational turn-taking between a human-intermediary pair. And secondly, analyzing the conversational roles taken in utterances can help us to identify how both parties take initiatives systematically in a information seeking conversations. Moreover, past experiments based on the scheme achieved some success in developing a dialog plan to generate the agent's response to user's request. Finally, a review of recent experiments on task-based dialog agents and chatbots suggested the potential usefulness of neural network based architectures in developing a dialog agent's response generation models in human-machine conversation. The influence of each of these findings in this research design is elaborated in the methodology section in Chapter 3.

Chapter 3

Methodology

The introduction chapter described the problem statement that this dissertation attempts to address. The research problem is about understanding how a conversational search system can clarify a user's information need when the expressed form of the need is inadequate. Complex task scenarios are of interest, because an ordinary user is likely to face difficulties in expressing the need in such scenarios. This chapter consolidates the research problem into three discrete research questions (RQs) and describes the methodology adopted in this dissertation work to answer the RQs. More specifically, the chapter introduces the data collection process, study procedure, and data analysis methods used for the experiments in this dissertation. The data collection process involves qualitative coding of conversations between a user and an intelligent assistant pair on the former's information need, collected in previous user studies. The data analysis process is consisted of multiple quantitative analysis techniques, including descriptive statistics, frequency analysis, state transition models, and other probabilistic models for predicting the agent's response type including clarification question types in an ongoing conversation. Each of the components from data analysis is discussed in detail in the subsequent chapters.

3.1 Research Questions (RQs)

3.1.1 RQ1

The purpose of this dissertation is to explore the ways a user's information need can be supported by a conversational search system. Previous experiments on search systems highlighted the importance of multiple types of support needed from the system to assist the users in their

information-seeking activities. For example, the system should be able to suggest or recommend queries when the query provided by the user is too vague or ambiguous to retrieve relevant information. More recent work on supporting user's information need in conversational search system explored the follow-up questions an intelligent conversational system should ask to clarify the user's information need (Aliannejadi et al., 2019; Zamani, Dumais, et al., 2020), which is also the focus of the research. Recent advancements in automatic speech recognition systems (ASR), natural language understanding (NLUs), and development of models that can mimic human conversation provide us the opportunity to design voice or text-based search systems that can respond to a user's request for information with follow up questions when further clarifications is needed. A conversational IR system with the capability to ask clarifying questions can take more initiative in the conversation and thus, have more responsibility towards understanding the user's "actual" information need and, which, in turn, can reduce the user's cognitive load during the conversation than a traditional search system. However, before we can develop such a system, the first step is to understand what type(s) of clarification to be asked by a conversational system that can help the users clarify their information needs for open-domain information-seeking dialogs. To this end, the research in this dissertation explored directions to develop a taxonomy of clarification questions for conversational search systems in open-domain information-seeking dialogs. More specifically, the first research question this dissertation tries to address:

- **RQ1:** What type of clarification questions does an "ideal" dialog agent ask in a conversational search system?

Method to Address RQ1

To address the RQ1, recordings of prior conversations between a user and an "ideal" conversational system on the user's information requirement, needs to be analyzed. However, obtaining such dialog data is a challenge as, according to the best of our knowledge, no such "ideal" conversational search system exists in the public domain from where conversational data can be collected at the time of this work. Since the ultimate goal of such a system is to be able to handle complex human information behaviors, it would seem that learning from

human-human conversational data is the better choice for conversational search system development (Budzianowski et al., 2018). However, learning from purely human-human dialogs presents challenges of its own. In particular, human conversation has a different distribution of understanding errors and exhibits turn-taking idiosyncrasies which may not be useful for interaction with a conversational search system (Byrne et al., 2019). In the absence of an "ideal" system, previous research found simulated human-machine information-seeking dialogs useful for conversational search system (Trippas et al., 2015; X. Yuan et al., 2011). The WOz framework, introduced by (Kelley, 1984) as a methodology for iterative design of natural language interfaces, presents a more appropriate approach to human-human dialog collection while simulating human-machine conversation. In this setup, users are led to believe that they are interacting with an automated assistant but in fact it is a human, the "wizard" behind the scenes who controls the system responses. Given the human-level natural language understanding, users can express their intent comfortably and naturally without modifying behaviors as is normally the case with a fully automated assistant. The appropriateness of these experiments relies on how appropriately the "wizard" can simulate the role of the "ideal" search assistant, and thus, selection of the "wizard" for such experiments is of utmost importance and often are based on the prior experience in handling user's information need problems. Previous studies that used Wizard-of-Oz settings to simulate dialogs in conversational search systems differed in study goals. For example, Thomas et al. (2017) studied task completion; Radlinski et al. (2019) focused on conversational preference elicitation or how the modality affected the conversation, i.e., through written text (Avula et al., 2018), spoken dialogs (Trippas, Spina, Cavedon, & Sanderson, 2017).

The alternate method of obtaining information-seeking conversation data between a human and an intelligent system pair that gained momentum in more recent studies suggested using experimental search interfaces, either based on suitable modifications of existing applications (Zamani, Dumais, et al., 2020; Avula & Arguello, 2020) or thorugh developing new search interfaces (Aliannejadi et al., 2019). However, such experimental systems have limited functionalities, typically tailored for the aspect of conversational search the developers wanted to simulate, and require significant amount of effort to adopt the system for any other conversational search experiments. Thus, the analysis done on the dialogs collected from such

system would be biased by the existing system and likely mimic its limitations (Williams & Young, 2007). Due to the limitations with experimental systems, in this dissertation, we analyzed information-seeking conversations which were collected from Wizard-of-Oz settings and through role-playing. Several dialog dataset from previous experiments for conversational search are made available as open-source by the researchers, including MISC (Thomas et al., 2017), Frames (El Asri et al., 2017), MultiWOZ (Budzianowski et al., 2018) dataset with follow-up versions (Zang et al., 2020), and TASKMASTER-1 (Byrne et al., 2019). Serban et al. (2018) discussed the major advantages and differences among the existing offerings in a detailed survey of available corpora for data driven learning of dialog systems. We used the Taskmaster-1 dataset for the experiments in this dissertation work. The complete dataset consisted of 13,215 task-based dialogs in English. Taskmaster-1 was chosen over the other open-source dialog dataset because the data collection for Taskmaster-1 involved a mixed procedure, comprising of dialogs collected in Wizard-of-Oz setting and "self-dialogs" collected through role-playing. For Wizard-of-Oz dialogs, a trained call center operator was used to act as an "intelligent voice assistant" and crowdsourced workers were recruited to play the role of "users". For a given task scenario, the "user" was asked to complete the task with the help of the "assistant". In role-playing, crowdsourced workers were asked to write down the full conversation themselves following the task outline provided by the experimenters. Thus, the same participant played the roles of both user and the assistant in a role-playing conversation. Further details on the instructions provided by the experimenters in both data collection processes are provided in Chapter 4. The tasks simulated in both settings were all complex tasks, chosen from multiple domains, and required faceted elicitation. Due to the task complexity, the dialogs were expected to represent complex negotiations between both parties on the user's information need involving back and forth exchange of questions and clarifications. Thus, inspecting the dialogs should provide us a rich inventory of clarification questions that were used to drive the negotiations. The goal here is to identify a common pattern in the clarification questions that appeared in the dialogs irrespective of the task domain. To identify the pattern of clarification questions, first we need to identify the role of each utterance in the negotiation process. To achieve this, each utterance in the conversation was labeled based on the functions it served in the negotiations and also towards completing the user's task. The annotation labels were

adopted from previous research on negotiations in the dialogs between a user and a human expert intermediary pair (e.g., a reference librarian) on the user's information need (Taylor, 1967). The negotiation (filter) labels were broadly defined as topic, motivation, preference, search strategy, anticipation, general conversation and task management. For the experiments conducted in this dissertation, only a small subset of dialogs was used for labeling. Details on data collection in Taskmaster-1 and the subset used in the experiments conducted in this dissertation were provided in the follow-up Chapters 4 and 5.

Each utterance in the conversational dataset used in this research was further annotated as per the speaker's conversational role in the current utterance from the COR model (Stein & Maier, 1995). The conversational role labels were used to identify and separate the utterances that were explicit questions posed by the agent from the rest of the responses in the dialogs. Since the conversations were user driven and focused on the users' information-seeking tasks in hand, all questions posed by the agent were assumed to be towards clarifying the user's information need. Secondly, the conversational role labels were helpful to characterize the nature of the utterances and dialogs by comparing the roles taken by the two parties engaged in the conversation. And last, the sequence of role labels can be useful in predicting the clarification question to be asked by a system in a previously unseen conversation, which we intend to explore in future work. Since the conversational role labels could be easily identified by an ordinary person and would not require an information specialist, getting conversational roles labeled by human annotators is cheaper than obtaining the same dataset annotated on the negotiation filters. As the conversational role label annotations often do not require in-depth semantic knowledge of the context and only need to consider the current utterance, a supervised approach to generate conversational labels for the utterances in future work can be worth exploring.

3.1.2 RQ2

As discussed in Chapter 2, previous research on conversational search systems has identified the differences in user's search behavior with the change of conversation modalities. For example, Crestani and Du (2006) found that using speech to formulate one's information need provided the user a way to express it more naturally and encouraged the formulation of longer

queries than in case of written ones. As a result, the user can be expected to express higher satisfaction when interacting with a spoken conversation system. The difference between written and spoken queries was also confirmed by Yan et al. (2016)'s experiment, where users initiated significantly fewer but longer queries for both interpretive and exploratory tasks in the spoken language interface than in the textual one. However, how the change in user's query behavior between the two modalities affects the system's clarification questions has yet to be explored in research. To the best of our knowledge, the only work that explored clarification questions on voice queries was by Kiesel et al. (2018). Their study on the impact of voice query clarification on user satisfaction found that users liked to be prompted for clarification in the spoken conversational system. However, no comparative analysis of text query clarifications and voice query clarifications was reported.

To address the knowledge gap on conversation modality's effect on clarification questions, this dissertation compares the clarification question types the agent used between written and spoken conversations. To this end, the second research question this dissertation attempts to address is the following:

- **RQ2:** How does the modality of conversation affect the clarification types in a conversational search system?

Method to Address RQ2

To study the RQ2, this dissertation used the same Taskmaster-1 dataset (Byrne et al., 2019) used for RQ-1 analysis. In Taskmaster-1 dataset, the self-dialogs simulated the users' written information seeking conversations with the dialog agent. Out of 13,215 dialogs in the complete dataset, 7,708 (58.33%) were collected in this fashion, as written dialogs. On the other hand, the rest 5,507 conversations (41.67%) were collected in Wizard-of-Oz settings and represented the spoken conversation between the user and the "voice assistant". Provided the types of clarification questions identified from the agent's utterances in analysis for RQ1, RQ2 needs comparisons of clarification types' appearances to identify any difference in their distributions between the two modalities. The dialog agent's clarification questions depend on the user's initial expression of the information need, and users with the same information need may not

start the conversation with the same initial expression. Thus, not all the dialogs may involve the same number or types of clarification questions even in case of the same information need, let alone tasks from different domains. Therefore, the length of a conversation (total number of utterances), including the number of clarification questions, may vary significantly depending on the task, domain, and the user's initial questions or queries, which can also affect our distribution of the clarification question types. We considered both total distribution and relative distribution of clarification types per dialog to offset the effect of longer or shorter conversation when comparing distributions between written and spoken dialogs for RQ2 analysis. The, use of existing user study data for the experiments conducted here did not affect the novelty of this work, as the previous experiments conduced on the Taskmaster-1 dataset differed in the study goal (Mosig, Vlasov, & Nichol, 2020). Also, before doing any analysis, the Taskmaster-1 dataset was tailored with utterance labels as per our study requirement.

With respect to RQ1 and RQ2, it is worth mentioning here some similar scenarios where the need to clarify a user's information need may arise. An inquirer may deliberately be vague some times while expressing the information need and therefore, can choose to issue ambiguous or multi-faceted queries to the system. This dissertation does not differentiate between such cases and when the user's expression of information need does not entirely represent the formalized need due to the user's inability to express the latter from the complexity of the task. Therefore, if the simulated users in Taksmaster-1 data collection intentionally chose to be vague or ambiguous while expressing the information need, such cases should be considered as false positives with respect to the dialogs analyzed here.

3.1.3 RQ3

Given a taxonomy of domain-independent clarification question types, an intelligent conversational search system should be able to decide not only what clarification questions to ask but also at what point(s) in an information-seeking conversation the clarification is required from the user. Previous experiments on the location of clarification questions in a dialog suggested heuristic-based approaches. For example, Aliannejadi et al. (2019) used a confidence measure to decide if the information requirement provided by the user could be used to present the retrieved result. If the system had low confidence in the retrieved documents, the system

chose not to present the result, and instead prompted the user with clarification questions. In contrast, Zamani, Dumais, et al. (2020)'s study was concerned with the usefulness of the clarification types in query-based web search. The clarification types were identified by mining query reformulation logs and using a slot-filling approach to generate clarifications for new conversations. Therefore, the participants from the treatment group in Zamani, Dumais, et al. (2020)'s experiment had access to the clarification pane all the time during their search tasks. However, in a conversational search system, not all search tasks can be expected to require faceted elicitation, and for the same reason, not all information-seeking conversations can be expected to have facet-related clarification questions. Additionally, asking clarifying questions that are not helpful can be more detrimental to the search performance for a voice-based search than a text-based system due to the linear nature of communication in the first medium. An "ideal" conversational search system should present a clarification question to the user only when the potential answer to the question can help with the negotiations on the user's information need. In order to create responses for such a system, we have to study the relationship between user's utterances and the clarification responses it follows in "ideal" information-seeking dialogs. In this aspect, the third research questions this dissertation attempted to address was the following:

- **RQ3:** What is the relationship between the characteristics of user's utterances and the clarification questions by an intelligent agent in a conversational information-seeking dialogs?

Method to Address RQ3

The overarching goal of the RQ3 is to understand how the agent's clarifying questions constitutes the negotiation process. To be able to address this question, this dissertation analyzed the order in which clarification types (as identified from RQ1 analysis) appeared in the user's prior information-seeking conversations with the agent. It is worth mentioning here that the objective of this experiment was not to derive an universal order of clarification types for all information-seeking tasks, which deemed unattainable. However, on the flip side, such exploration was necessary at least to establish if there was a pattern in which the agent's clarification types appeared in the "ideal" information seeking conversations. It is a necessary first step towards

building prediction models for an intelligent conversational system's response for open-domain information-seeking dialogs. To this end, following the previous work (Hendahewa & Shah, 2013) on segmenting information-seeking episodes, each annotated dialog was segmented into four sequences with an equal number of utterances in each sequence. The frequencies of negotiation labels in each of the four sequences were analyzed. Following the hypothesis of RQ2 that the agent's use of clarification questions in the negotiation process was expected to differ between written and spoken conversations, the segmentation and frequency analysis of labels in RQ3 was done separately for written and spoken conversations. We also explored state transition models towards investigating the sequence of negotiation labels in the information-seeking conversations. Previous experiments on state transition models found applications in analyzing sequence of dialog acts by focusing on the temporary assignment, acceptance or refusal of conversational roles during the conversations (Stein & Maier, 1995). From the experiments in RQ3, a transition network for the negotiation process with each state representing a clarification emerged which, when considered with other context of the dialog, can be used for generating responses for an automated agent in future information-seeking conversations.

The location of the agent's clarification questions in a dialog can be affected by additional factors, e.g., the user's background knowledge on the topic, motivation behind the task, user's initial expression of the need and realization of change in information need during a conversation. Since in this dissertation work the focus is on the conversation itself, we considered only user's utterances and examined their relationship with the subsequent clarification questions from the agent. Characteristics of prior user's utterances that prompted a clarification question by the agent in a dialog were analyzed to examine the relation between the user's utterances and the agent's clarification type. The utterance characteristics used to examine the relationship were based on a mix of lexical and semantic attributes. Sequential models that considered the context of the clarification questions were used to examine this relationship. Further details on the attributes and the models used in these experiments, are discussed in Chapter 5.

3.2 Summary

The earlier sections in this chapter introduced the research questions (RQs) concerning the research problem discussed in Chapter 1, which the experiments in this dissertation attempted to address. Specifically, the three RQs were:

- **RQ1:** What type of clarification questions does an intelligent agent ask in a conversational search system?
- **RQ2:** How does the modality of conversation affect the clarification types in a conversational search system?
- **RQ3:** What is the relationship between the characteristics of user's utterances and the clarification questions by an intelligent agent in a conversational information-seeking dialogs?

The RQ1 was about establishing a taxonomy of clarifications for open-domain information-seeking dialogs. Creating a taxonomy can help us to design responses for a conversational search agent that can ask the user clarifying questions in the future conversations. Prior experiments on conversation modality pointed to the difference in user's search behavior between written and spoken search systems. Based on this observation, we hypothesized that modality should also affect the clarification questions the system can ask the user, which was tested in RQ2 analysis. Due to the recent popularity of commercial voice-based assistants, voice-based search gained prominence in the research, and yet most of the current models of clarification questions were generated for traditional web search. The RQ2 aimed to address this gap. The last question, RQ3 explored the location of clarification types in the negotiation process. To this end, the order of clarification types in the dialogs was investigated. The goal was to create a transition network for the agent's responses in the negotiation process with each state representing a clarification question type. The transition network can serve as a baseline template for generating the agent's response in future information seeking conversations. Additionally, following the prior work on segmenting information-seeking sessions, each dialog was segmented into multiple sequences, and the frequencies of clarification types in each sequences were compared. We further probed the relation between the user's utterances and the agent's subsequent

clarification question types to examine how the user's expression of the need affected the order of clarification questions. The utterances characteristics used for this experiments were lexical and semantic attributes.

The methodology adopted in this research involved analyzing the user's prior information-seeking dialogs with a simulated "true" conversational assistant in complex task scenarios. The dataset used for the experiments in this dissertation was from Taskmaster-1 (Byrne et al., 2019), which consisted of written and spoken conversations. The spoken dialogs were collected in Wizard-of-Oz setting, where a call center operator was used to play the role of the "true" assistant, whereas, for the written dialogs same crowdsourced worker simulated the roles of both the user and the assistant following the script of the task scenarios outlined by the experimenters. To analyze the negotiation process between the user and the system required the dialogs to be further labeled in a coding scheme that encapsulated the utterances' roles with respect to the global goal of the user's information need. In this dissertation, we used two coding schemes to label the utterances, (1) based on Taylor's filters (Taylor, 1967) to analyze the type of questions asked by the assistant to the user, in conjunction with (2) labels from the COR model (Sitter & Stein, 1992) elucidating the exchange of conversational roles that happened during the negotiation process. The conversational roles are identifiable expectations concerning future behavioral responses from the conversational partner. The characteristics of the utterances were used to predict the type of clarifying questions to be asked by an intelligent conversational IR assistant model in previously unseen conversations. The following chapter covers more details on the data collection process and annotations.

Chapter 4

Data Collection

4.1 Conversational Dataset

The experiments in this research used the Taskmaster-1 dataset, which is a dialog corpora consisting of written and spoken dialogs. Two procedures were used to create the data collection, each with it's own advantages and disadvantages (discussed in the following sections). Both procedures involved recruiting participants, e.g., 'users' from a crowd-sourcing platform, Amazon Mechanical Turk (MTurk). Each user was given a set of tasks to complete by the experimenters (Byrne et al., 2019). The tasks were set up from six domains, e.g., ordering pizza, creating auto repair appointments, setting up ride service, ordering movie tickets, ordering coffee drinks, and making restaurant reservations.

4.2 Spoken Dialogs

In spoken dialogs, Wizard-of-Oz methodology was used to collect the data where users were instructed to complete the given tasks through spoken conversations with a simulated conversational assistant. The role of the 'assistant' was played by trained call center operators, who were hired from a pool of dialog analysts and were trained on the setup and interface for two hours and on how to handle anticipated challenges such as technical glitches and unreasonable users. Technical challenges included dropped sessions (e.g. connection failure) or cases in which the user could not hear the agent's voice or vice-versa. Uncooperative users' behavior typically involved ignoring agent's input or rushing through the conversation with short answers.

Instructions provided by the experimenters to both parties for one of the six tasks chosen for the Taskmaster-1 data collection are given in Table 4.1.

Table 4.1: Instructions provided to 'user' and 'agent' by the experimenters in Taskmaster-1 dataset (Byrne et al., 2019)

Instructions for the 'user'	**Instructions for the 'agent'**
In this conversation you're going to pretend you need to take your car to the mechanic, so you need to get an appointment scheduled.	In these conversations, users will call their assistant to set up an auto repair appointment with a repair shop called "Intelligent Auto Imports".
MAIN TASK: Use your voice-powered, personal digital assistant to make an appointment at an auto repair shop called "Intelligent Auto Imports".	Your job as the Assistant is to set up the appointment on their behalf. By the end of the conversation, you'll need to gather the following information. Do not feel compelled to do things in this exact order though. AND of course you'll need to take several turns to gather these bits.
Your car: car make, model, year	• name.customer • "Gina Jones" (not their real name)
In addition to the car, you need to give a name and phone number. DO NOT USE YOUR REAL NAME AND PHONE NUMBER UNDER ANY CIRCUMSTANCES!!!	• phone.customer • 10-digit telephone number (not their real number) CONFIRM the name and phone number HERE—do not wait until the end of the call
Describe the following reason for your appointment in your own words: {reason}	• reason.appt • Users will describe their problem, e.g. "tune up", "there's a funny noise when I turn", "It keeps stalling", "pulls to the left", "leaking oil", etc.

Continued on next page

Table 4.1 – continued from previous page

Instructions for the 'user'	Instructions for the 'agent'
As far as the day and time are concerned, date-time The assistant should end the conversation by confirming BOTH the details you gave, the appointment time, as well as the fee for inspection.	• Date.appt • You should start by asking them which day/date they need the appointment for. • Can be actual dates like "April 29th" as well as days of the week "this Friday", etc. • time.appt • You should let them know that the usual procedure is for them to drop off their car before 8:30am on the day of the appointment. • However, some users will give you a particular preference like: "I'd like someone to take a look this afternoon", "at 2pm", "ASAP", etc. IMPORTANT: If they do ask for a non-standard time such as that described above, tell EVERY OTHER customer of this type that right now the shop only has availability tomorrow and that they should drop their car off before 8:30. (For the other half you can accommodate their more immediate time request.)

In this set up, users were led to believe that they were interacting with an automated system to complete the task while it was, in fact another human, allowing them to express their information need in natural turns but in the context of an automated interface. Thus, dialogs collected from Wizard-of-Oz set up are better suited for modeling human-computer conversation than other human-human dialogs. The user's audio-only portion of the dialog was transcribed and then merged with the assistant's typed input to create a full transcription of the complete dialog. Finally, these text versions of the dialogs were checked for transcription errors and typos.

While the two-person setup for data collection in WOz methodology creates a realistic scenario for robust, simulated human-machine spoken dialog corpora, this technique is time consuming, complex and required considerable technical implementation including administrative procedures to train and manage agents and workers.

4.3 Written Dialogs

The written conversations were collected by engaging crowdsourced participants to write the full conversation themselves (i.e., self-dialogs) based on scenarios outlined by the experimenters for each task. In this setup, users were asked to imagine that they had a personal assistant to help them take care of various tasks in real time. Given a task scenario, they were told to imagine that they were interacting to their assistant while the assistant accessed the services for the given tasks. The users then wrote down the entire conversation as they envisaged the conversation to look like. Thus, in written dialogs, the same participant played roles of both the user and assistant in every conversation. Instructions given to the recruits for writing the dialog for one of the task, pizza ordering is given in Table 4.2.

Table 4.2: Instructions provided to 'user' for self-dialogs in Taskmaster-1 dataset (Byrne et al., 2019)

Instructions for the 'user'
To begin, think of one of your favorite pizza places and think about the types of pizzas you like to order. Make sure you study their menu to confirm the details of the choices offered like: toppings, sizes, prices, specialty pizzas, sides. IF YOU'VE DONE THIS BEFORE, CREATE A NEW STORY/VERSION.
Continued on next page

Table 4.2 – continued from previous page

Instructions for the 'user'
The pizza should have at least two toppings. If it's a specialty pizza, ask if you can exchange one of the toppings for something else.(You must specify it, like "Instead of the x can I get y on that instead?") MAIN TASK: Pretend you call your personal assistant on the phone to have them order ONE pizza for you from this place. Write the conversation that would happen between you and your assistant in order to buy the pizza online. (Try to make the order realistic–like for a meal/event with you and your family or friends.) MAKE SURE the assistant asks about all relevant details. Don't order other items-just the pizza. You can assume you already have an account with this business which your assistant knows, so no credit card information is necessary. The assistant should confirm all of the details of the order. To end the conversation the assistant will tell you that your pizza order is complete and the pizzas will be ready for pickup in about 25 minutes. Payment: Your assistant can tell you that your receipt will be sent to your mobile device via text message. DO NOT GIVE ANY PERSONAL INFORMATION: no phone numbers or addresses, names, etc. If you want to include this type of info, make it up NOTE: IT doesn't hurt to include a turn or two where what you want isn't available or where your assistant has to correct you as to what things are called, sizes available, etc. This makes it more realistic!! :-)
Continued on next page

Table 4.2 – continued from previous page

Instructions for the 'user'
YOUR TASK: Write the conversation that results between you and your assistant. It must be at least 10 turns long (for both you and the assistant). Below we have provided 15 turns in case you need more. KEEP IT NEW AND FRESH! DON'T REPEAT DIALOGUES FROM THE PAST!

Since the workers were not restricted to detailed scripts or to a small knowledge base and hence, the self-dialog technique rendered quality data without some of the challenges seen with the two person approach. Since the same person is playing both sides of the conversation, the data collection did not reflect any communication issues, misunderstanding or frustration as it was sometimes experienced between interlocutors in the two-person WOz approach. Also, self-dialog approach is far more simple, without the need of any transcription or trained agents and hence, more efficient and cost-effective approach to create large scale dialog corpora. Despite the advantages, self-dialog approach is not without its limitations. The written conversations in self-dialog technique cannot recreate the disfluencies and more complex error patterns that are typically common in the two-person spoken dialogs.

The Taskmaster-1 dataset collection used tasks from six domains. In total, the dataset had 13,215 task-based dialogs in English, including 5,507 spoken and 7,708 written dialogs.

4.4 Data Labeling

The experiments in this research involved analysis of the negotiation process of the user's information need irrespective of the task domains, as happened in conversation with a conversational assistant. Thus, the negotiation process had to be dissected into individual functions each conversational turns served towards the global goal of fulfilling the user's information need. These functions were drawn from Taylor (1967)'s work (i.e., filters) on human-intermediary negotiations on the user's information need and these labels were topic of information need or

simply *topic*, *motivation* behind the need, user's background or *preference*, alternate *search strategy*, and *anticipation* of the nature of the information by the agent. However, there were important differences between the human-intermediary dialogs considered in Taylor's work and the type of "simulated" human-machine dialogs considered here. In our case, the information-seeking dialogs were clearly about an overarching task that the user wanted to accomplish with the help of the agent. The agent were expected not only to provide the relevant information, but also be able to access the relevant services to complete the task goal as per user's direction. Taking cognizance of the difference, an additional category of *task management* was introduced in our coding scheme to label those instances where an utterance served towards completing the global task without advancing the discussion on information need. Additionally, it was observed in the dataset that both speakers took regular turns just to maintain and manage the communication process as natural in human language behavior, e.g., opening greetings, closing remarks etc. The category of *general conversation* was added to our coding scheme to label such utterances. Detailed descriptions of each of the categories from the codebook is provided below.

- Determination of the subject (Topic, or 'T'): This label described the topic of user's information need and provided some delineation of the information space. For example, user's initial utterances typically are expected to provide from partial to complete description of the topic. The agent may ask follow-up questions such as "Is this what you mean" to further clarify the topic. Following is a snippet of dialog that where both parties use this filter to move the negotiation forward.

- Motivation and Objective of the Inquirer (Motivation, or 'MO'): The utterance marked with this label represented why the inquirer wanted this information. For example, the agent may the ask user what is the motivation behind this inquiry. The user may respond to such requests with information such as "I want this because .." In this case, both speaker's utterances represented were to be marked with 'motivation' label.

- Personal Background of the Inquirer (Preference, or 'PR'): The utterances that were to elucidate the user's background, familiarity with the system, and the aspect of the information need already known to the user, were to be marked with this label. Answers

to these types of questions help the system to determine the urgency, the negotiation strategy, level or depth of any dialog, and the critical acceptance of search results etc.

- Relationship of Inquiry Description to File Organization (Search strategy, or 'SS'): Through this utterance types, the agent restructured the user's inquiry that seemed best fit for effective retrieval. For example, the agent may ask follow-up questions such as "Have you tried this". Utterances of this type can be thought of conversational equivalent of query suggestion or query recommendations from traditional query-based search system.

- What Kind of Answer Will the Inquirer Accept (Anticipation, or 'AN'): When an inquirer approaches the information system, he has some picture in mind as to what he expects the information to look like, e.g., it's specificity, format, modality etc. The agent's utterances that were to resolve what the user was expecting the information to look like, were to be labeled in this category.

- Managing some aspects of the task (Task management or 'TM'): Utterances in this category represented the speaker performing some action related to managing the task, e.g., asking the status of a process "Are we done?", "Should I book the table?", or giving some action directives, such as "please hold", "click the button" etc.

- Maintaining communication process(General conversation, or 'GC'): These utterances were not associated with negotiation of the task or task management but rather serve the function of communication management as per the language behavior and social norms, e.g., Greetings ("Hello", "Good bye") or acknowledgement ("OK", "Sure").

- Other: Where none of the above seven labels apply, the annotators were instructed to label the utterances as "Other" category.

In addition to the negotiation (i.e., filter) labels, each utterances from the dialogs were labeled with the dialog act or conversational role label, from the COR (Sitter & Stein, 1992) model. The dialog acts represent the social role a speaker takes on in the current utterance while assigning the complementary role to the hearer. The set of social roles that were permitted as labels in this coding scheme is provided in Table 4.3.

Table 4.3: A sample conversation labeled with conversational roles as per the COR model (Stein & Maier, 1995)

dialog Act	**Utterance (Speakers: A, B), <Role >**
request	A: When does the next WSDM conference takes place? <request(A, B) >
offer	B: In March 2021, <offer (B, A)>
reject offer	A: But when in March? <reject offer(A, B) >
assert	B: I don't know. <inform (A, B) >
promise	B: OK, I'll have a look <promise (B, A) >
accept	A: OK. <accept (A, B) >
be contended	A: Thanks <be contended (A, B) >
withdraw request	A: Never mind. <withdraw request (A, B) >
withdraw offer	B: Sorry I can't find the schedule in the invitation <withdraw offer (B, A) >
be discontented	A: Can I have at least the dates? <be discontented (A, B) >
reject request	B: I don't have the dates either <reject request (B, A) >

A sample dialog labeled in both coding schemes is provided in Table 4.4. For annotation, two persons who specialized in handling users' information needs on a routine basis, i.e., librarians, were recruited through emails from university library and staff directory. A sample recruitment email is provided in Appendix B. Prior recruitment, the the Institutional Review Board (IRB) approval was obtained [Pro2020000991]. No recruit specific identifier including name, age, gender, and e-mail was collected or stored for the purpose of this study. No in-person meeting was required to participate in the annotation. Each participants had to read and sign the consent form electronically before participation. The recruits for annotation had to meet the following eligibility criteria:

1. The annotator must be a native English speaker.

2. The annotator must be a librarian with at least some (average 2 years) of experience in handling library users' information problems.

3. Each annotator should be familiar with the qualitative coding process and had done such coding in the past.

Upon completion of study participation (approximately between 3-4 hours), each participant was compensated by $75 for their time and effort. Both recruits were provided sufficient

training on the coding schemes for this task, before coding any real data to be used for analysis to address the research questions in this dissertation. The training process involved virtual meeting with the annotators, explaining the coding schemes, and assigning 20 randomly chosen dialogs for practice labeling. A codebook was developed that contained descriptions of each labels from both annotation schemes along with an example of a completely coded dialog. The codebook is provided in Appendix A. Each annotator was explicitly instructed to go through the codebook first to get familiar with the labels, before starting any labeling work. Additionally, they could refer to the codebook any time during the coding process.

Once the annotators were familiar and comfortable with the coding process, they were assigned a new set of 20 dialogs for labeling. These 20 dialogs were randomly sampled from the Taskmaster-1 dataset and contained 10 spoken and 10 written dialogs. Both written and spoken set had at least one dialog from all six task domains. These 20 dialogs were common between the two annotators and had no overlap with the dialogs that were used for practice labeling. Written dialogs and spoken dialogs were assigned for labeling in separate files, however, the coders were not aware of the dialog collection process.

The 10 spoken dialogs included 101 user utterances (46.98%) and 114 assistant utterances (53.02%), containing a total of 215 utterances. In comparison, the same number of written dialogs had 107 (50.47%) user utterances and 105 (49.52%) assistant utterances, with a total of 212 utterances. The utterance labels, as coded the annotators in this round, were compared. The agreements and disagreements on labels between both annotators were analyzed to test the reliability of the coding scheme, and to evaluate if further training was required. The next section provides more details on coding reliability.

4.5 Coding Reliability

This section discusses the agreement on negotiation labels (i.e., Taylor's filters) between the two annotators first, followed by details on the agreement for the conversational role labels. At the end of this section, implications of the agreement and disagreement is discussed.

Table 4.4: A sample dialog labeled with both data annotation schemes

Utterance ID	**Speaker: Utterance (U: user A: agent)**	**Filter**	**Filter reason**	**Conversational role**
1	U: Hi, I'm looking to book a table for Korean food.	topic	The user describes the topic of this conversation as booking a table for Korean food.	request (U, A)
2	A: Ok, what area are you thinking about?	topic	Agent elicits for more information on the topic	request (A, U)
3	U: Somewhere in Southern NYC, maybe the East Village?	topic	User clarifies the initial topic description with more information.	assert (U, A)
4	A: Ok, great. There's Thursday Kitchen, it has great reviews.	GC, motivation	Agent's acknowledgement followed by a suggestion with reasoning with hope the reasoning will match user's motivation.	offer (A, U)
5	U: That's great. So I need a table for tonight at 7 pm for 8 people. We don't want to sit at the bar, but anywhere else is fine.	motivation, preference	Further clarification on motivation is provided by the user followed by details of preferences.	accept (U, A), request (U, A)
6	A: They don't have any availability for 7 pm.	TM	Agents response in negative saying it cannot proceed with the task.	assert (A, U)
7	U: What times are available?	search strategy	User suggests change in search strategy from looking for restaurants to look for times/slots when tables are available.	request (U, A)
8	A: 5 or 8.	TM	Agents responds to user's request	offer (A, U)
9	U: Yikes, we can't do those times.	TM	User responds that agent's retrieved information is not helping in completing the task.	reject offer (U, A)

4.5.1 Coding Reliability on Filter labels

The coding labels on utterances, as labeled by the two annotators, were analyzed and compared for agreement for both written and spoken dialogs together and also separately. For written dialogs, the first annotator (henceforth referred to as coder-1) labeled the utterances with the topic filter (T) on 72 occasions, which accounted for 32% of all the labels and approximately 33.96% of all the utterances. General conversation (GC) was used to label the utterances on 43 occasions (20%) from the same dialogs, which also was the most frequent label after topic in coder-1's data. Preference (PR) and anticipation (AN) labels from the coding scheme were the next most frequent ones in coder-1's labels appearing on 37 (17% utterances) and 33 (16% utterances) occasions, respectively. The other two labels, search strategy (SS) and motivation (MO) from Taylor's filters, were the least observed in coder-1's data, with the last appearing on only 5 (2%) occasions. Apart from the above six labels, a significant number of utterances (25 occasions and in 15.57% of utterances) in written dialogs were labeled as about managing the task (TM) by coder-1. Moreover, coder-1 labeled 13 utterances (6%) from written dialogs with multiple labels, with two labels on each occasion. T and GC were the most frequent labels (3 utterances) that co-occurred in coder-1's labels.

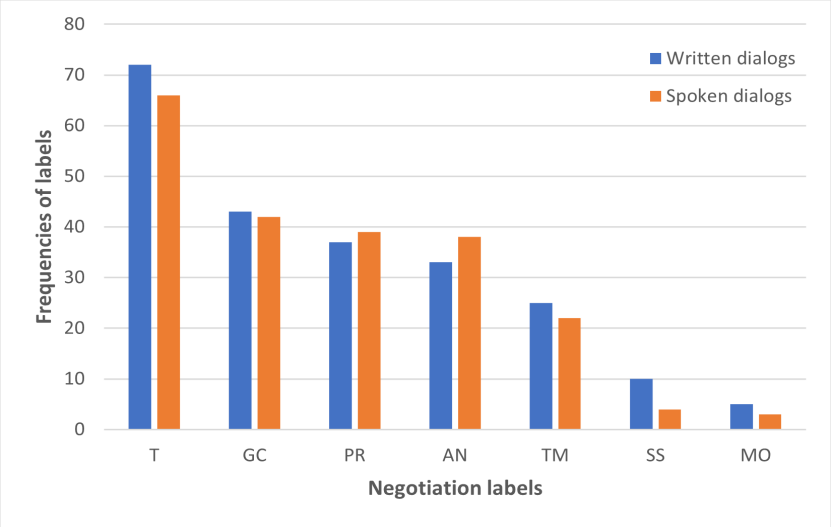


Figure 4.1: Distribution of negotiation labels (Taylor, 1967) in written and spoken dialogs as labeled by the first annotator (T = Topic. GC = General conversation, MO = Motivation, TM = Task management, PR = Preference, AN = Anticipation, SS = Search strategy)

The distribution of labels in the 10 spoken dialogs was near identical to the written ones

in coder-1's data, with a slightly smaller percentage accounting for all the labels except for anticipation (AN), and preference (PR). Out of 215 utterances, 39 utterances (18.14%) were labeled as PR label, and 38 were (17.68%) labeled as AN. Thus combined, these two labels accounted for 36% of all the labels assigned to utterances in spoken dialogs by coder-1. However, compared to written dialogs, only on 3 occasions(1.40%), the same utterances were coded with multiple labels by coder-1, all three involving task management (TM) and anticipation (AN) labels. Moreover, unlike written dialogs, on 4 occasions (1.86%), none of the existing seven labels were deemed appropriate and hence labeled as other (O) label by the same annotator in spoken dialogs. The distributions of negotiation labels in coder-1's data were compared between written and spoken dialogs in Figure 4.1.

Table 4.5: Use of negotiation labels by the two annotators for written dialogs

Label(s)	**Used by the first annotator**	**Used by the second annotator**
Topic (T)	72	52
General conversation (GC)	43	56
Preference (PR)	37	6
Anticipation (AN)	33	16
Task Management (TM)	25	69
Search Strategy (SS)	10	6
Motivation (MO)	5	9
Other (O)	0	0
Total	225	214

The distribution of labels in coder-2's data for written and spoken dialogs is provided in Figure 4.2. Compared to coder-1, the second annotator (henceforth referred to as coder-2) labeled a significantly higher number of utterances with task management (TM) for both written and spoken dialogs. Out of a combined total of 427 utterances in written and spoken dialogs, 138 utterances (32.32%) were tagged with TM labels by coder-2, compared to only 47 TM labels (11.01%) used on the same dataset by coder-1. Moreover, the distribution of labels between written and spoken dialogs for coder-2 showed more significant variations than coder-1, especially for less frequent labels, such as anticipation (AN), motivation (MO), and preference (PR). In spoken dialogs, the number of times AN label occurred (5 times, 2.33%) was less than

half than the same occurred in written dialogs (16 times, 7.55%). At the same time, the MO and PR labels were assigned more than twice for written dialogs by coder-2. The distribution of labels for written dialogs is compared between the two annotators in Table 4.5.

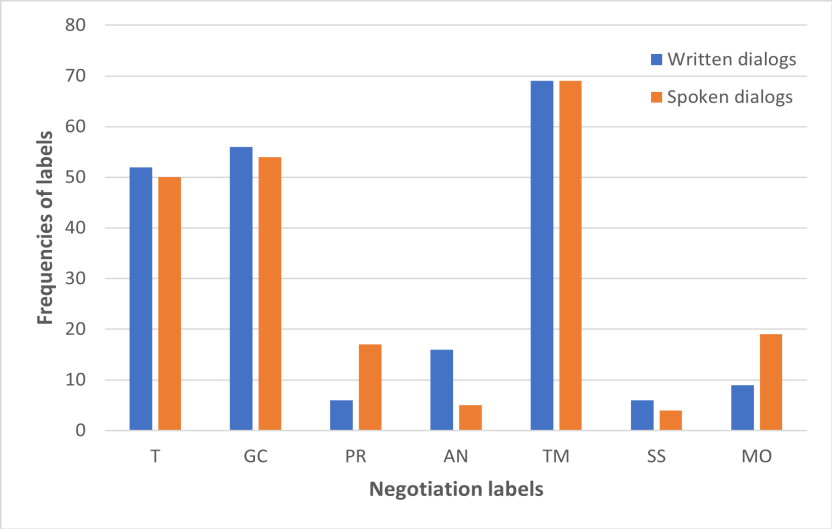


Figure 4.2: Distribution of negotiation filters (Taylor, 1967) in written and spoken dialogs as labeled by the second annotator

Similar to coder-1, the frequencies of labels for the other three criteria, topic (T), general conversation (GC), and task management (TM), as used by coder-2, were almost similar between written and spoken dialogs. For written and spoken dialogs, GC label was observed 56 (26.42%) and 54 times (25.12%) respectively. The topic label (T) was used to tag 52 (24.52%) utterances in written and 50 utterances (23.26%) from spoken dialogs by coder-2. Compared to coder-1, coder-2 used multiple labels on the same utterances only on 3 occasions in combined written and spoken dialogs. On all three occasions, AN and TM labels were used. Unlike coder-1, coder-2 never used the other label (O) for any utterances in either written or spoken dialogs. The negotiation labels' distribution for spoken dialogs is compared between the two annotators in Table 4.6.

The agreement on coding labels between the two annotators was measured in Cohen's Kappa (J. Cohen, 1960). The Kappa score calculated based on the 20 labeled conversations suggested moderate agreement (Landis & Koch, 1977) between the two annotators on negotiation labels for written conversations (0.42) and spoken conversations individually (0.48) and

Table 4.6: Use of negotiation labels by the two annotators for spoken dialogs

Label(s)	**Used by the first annotator**	**Used by the second annotator**
Topic (T)	66	50
General conversation (GC)	42	54
Preference (PR)	39	17
Anticipation (AN)	38	5
Task Management (TM)	22	69
Search Strategy (SS)	4	4
Motivation (MO)	3	19
Other (O)	4	0
Total	218	218

also as combined (0.45). Among all disagreements on negotiation filters, the majority of utterances involved TM (39%) or PR (32%) labels. In comparison, both annotators mostly agreed on T and GC labels. Out of a total of 15 utterances that were assigned multiple labels by at least one of the annotators, none of the cases both annotators agreed on both labels. The Kappa scores improved to 0.44 for written and 0.51 for spoken dialogs when these 15 utterances involving multiple labels were ignored from coding agreement measurements. Since the labels from negotiation filters were going to be used to predict the system's response type, the agreement between the two annotators on the assistant's utterances alone was also measured for spoken and written dialogs separately. In both cases, a moderate agreement with minor increases in Kappa scores was observed among both annotators, 0.43 for written and 0.51 for spoken dialogs. At the end of this stage, we met with both annotators together to discuss the disagreements and to collect their feedback on the coding process. It was evident from the coded dialogs that the coder-2 was heavily biased towards the task management label, and assigned the same to majority of the utterances. The disagreements were resolved upon explaining each labels providing more examples of coded dialogs. The disagreements on TM, PR and AN labels were resolved by explaining each label with more examples of coded dialogs. At the end of this session, both annotators felt confident on negotiation labels coding. Both annotators suggested an excellent coverage of the coding scheme, which was also reflected through the occasional use of the other (0) label. Both coders also noted the longer time needed to label the

spoken dialogs compared to written ones due to larger variation in dialogs from the spoken set.

4.5.2 Coding Reliability of the Conversational Role Labels

Similar to negotiation filters, Cohen's Kappa was used to measure the agreement on conversational role labels between the two annotators. Request and assert labels were observed most frequently among all conversational roles in both written and spoken dialog utterances by both annotators. Coder-1 assigned the request label to 69 (32.55%) and 64 (32.56%) utterances from written and spoken dialogs, respectively. Whereas, in coder-2's data the same label was used 58 (27.36% of utterances) and 64 (29.76% of utterances) times for written and spoken dialogs, respectively. Among the rest of the role labels, be discontented, reject offer, and withdraw offer were overall the least frequent labels on the dataset for any annotator. Together these three labels were used to label only 41 utterances (4.80%) by both annotators combined. Overall, both coders noted a similar distribution of conversational roles between written and spoken dialogs. Only coder-2's use of offer and promise labels showed significant difference between written and spoken sets. Comparisons of role labels' distribution between written and spoken dialogs were provided in Table 4.7 and Table 4.8 for coder-1 and coder-2, respectively.

Table 4.7: Comparison of conversational role labels' distribution as used by the first annotator

Label	**Count in written dialogs**	**Count in spoken dialogs**
Request	69	70
Assert	78	69
Accept	34	19
Be contended	12	9
Offer	18	21
Promise	17	19
Be discontented	7	9
Reject request	8	12
Reject offer	2	5
Withdraw offer	3	1

In comparison, a significant difference was observed between the two annotators in the number of utterances coded with multiple labels. Coder-1 used multiple labels on 38 spoken (17.67%) and 23 written (10.85%) dialogs. In comparison, coder-2 had co-occurrence of

multiple labels only on 3 (1.40%) and 7 (3.30%) occasions for spoken and written dialogs, respectively. Among all the utterances tagged with multiple labels by coder-1, request and assert co-occurred most frequently for spoken dialogs (12 times, 5.58% of utterances), accept, and asset labels for written dialogs (17 utterances, 8.02% of utterances).

The Kappa scores suggests substantial agreement between the two annotators on conversational role labels for written conversations (0.61) and spoken conversations individually (0.64) and also combined. Among all disagreements on conversational roles, the majority of utterances involved assert and request labels. Removing the utterances where at least one of the annotators used multiple labels from agreement score measurement, the Kappa score further improved to 0.68 for written and 0.73 for spoken dialogs. Nonetheless, overall the annotators had a good agreement on conversational roles for all the utterances (0.63) in written, and spoken dialogs combined.

Table 4.8: Comparison of conversational role labels' distribution as used by the second annotator

Label	**Count in written dialogs**	**Count in spoken dialogs**
Request	58	64
Assert	77	66
Accept	21	24
Be contended	15	6
Offer	15	32
Promise	12	3
Be discontented	3	0
Reject request	3	2
Reject offer	1	3
Withdraw offer	1	0

4.6 Summary

This chapter described the data collection process for the experiments involved in this dissertation. The data used for analysis was the Taskmaster-1 dataset, comprising self-dialogs (written) and WOz dialogs (spoken), simulating the conversation between a user and a conversational system on the user's information need. We used a subset of this dataset and further labeled the

subset's dialogs with our annotation scheme. The annotation scheme was designed to label each utterance in the dialogs with respect to its function in the negotiation process and exchange of conversational roles during the negotiation. The annotation scheme was based on Taylor's work on the negotiation of user's information need by the intermediary and the COR model (Sitter & Stein, 1992). For labeling the dialogs, two coders who had prior experience in handling user's information-seeking questions in the library were recruited. The study process involving the recruitment, obtaining consent from the participants, and their compensation was approved by the IRB [Pro2020000991] (see Appendix B).

Before labeling the primary dataset reserved for annotation and analysis, both annotators were provided sufficient training in the coding process. As part of the training, the annotators were provided with a codebook explaining each label in the coding process and use case scenarios for each label (Appendix B). Next, a set of 20 dialogs randomly sampled from Taskmaster-1 (excluding the subset earmarked for our experiments) was assigned for practice labeling. Once the annotators were well-versed with the coding process from practice labeling, another set of 20 dialogs (similarly sampled) with equal numbers of written and spoken ones were assigned to both coders for annotation. The labels from these 20 dialogs were used to test the agreement between the two annotators on coding labels. In general, analysis of utterances' labels suggests moderate agreement between the annotators on negotiation labels and substantial agreement on conversational role labels. However, for none of the utterances with multiple labels (by at least one annotator), both annotators agreed on all the labels, suggesting challenges in coding the same utterance with multiple labels in these annotation schemes. At the end of this stage, the experimenter virtually met with the annotators to discuss agreements and disagreements in their labels and collect feedback on the coding process. Once the disagreements were resolved, the subset earmarked for the experiments was divided equally and assigned to the two annotators for labeling. The following chapter discusses the primary data, the distribution of task domains, the analysis of the labeled data obtained from the two annotators, and the analysis findings.

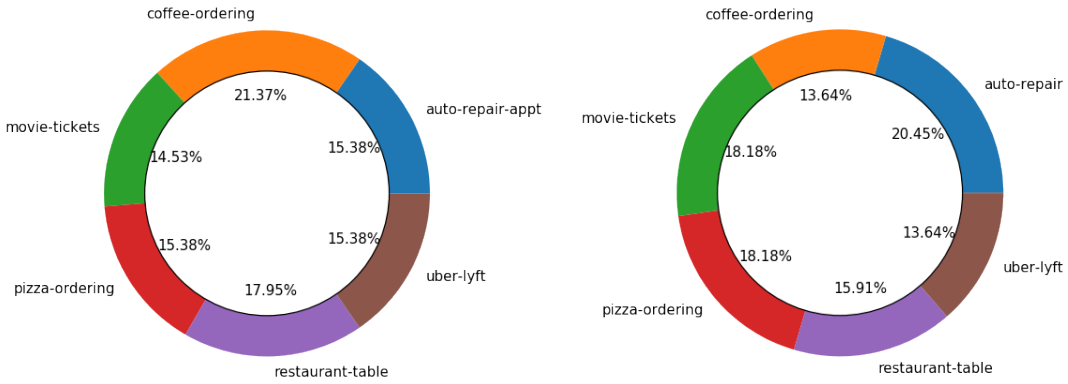
Chapter 5

Analysis and Results

5.1 Data Description

To answer the three RQs outlined in the methodology chapter, we used 161 dialogs from the Taskmaster-1 dataset that were further labeled by the two annotators in the coding scheme outlined in Chapter 4. Out of these 161 dialogs, 117 were written dialogs, and the rest 44 were from the spoken set. These 161 dialogs were sampled from the Taskmaster-1 dataset in a manner that the distribution of all six task domains was consistent between the sample and Taskmaster-1; however, to avoid any bias from the order of the dialogs' appearances in the sample, their order was randomized. The distributions of the task domains between written and spoken dialogs in the set of 161 dialogs are provided in Figure 5.1. The figure suggests a close to even distribution of the task domains in the selected sample. Dialogs from coffee-ordering task were the most frequent ones (25 dialogs, 21.37%) among written dialogs, whereas dialogs from auto-repair tasks had the maximum presence (9 dialogs, 20.45%) in the spoken set. From the 161 dialogs selected, the first annotator (i.e., coder-1) labeled 81 dialogs (59 written and 22 spoken), and the second person (coder-2) did the rest 80 (58 written and 22 spoken). The distribution of task domain was even between the two annotators and was consistent with the whole sample.

The rest of this section discusses the length of a conversation in terms of the number of utterances per dialog. The 161 dialogs contained a total of 3,598 utterances that were labeled by the two coders. The dialogs were relatively long (mean number of utterances per dialog 22.35), which suggests that the task scenarios assigned to simulate the dialogs were complex and needed long conversations to complete the task. In general, spoken dialogs (mean number of utterances: 24.50) were longer than written ones (mean number of utterances: 21.54). Additionally, the range of utterance counts per dialog was larger for the spoken dialogs than in case



(a) Written dialogs (b) Spoken dialogs

Figure 5.1: Distribution of task domains in the annotated dialogs

of the written ones despite a much smaller sample size in the spoken set. The larger variation could be due to different negotiation strategies or higher error rate in understanding the dialog partner for the spoken dialogs. Nonetheless, both annotators noted the higher variation and consequently, longer time for labeling the spoken dialogs in the practice round, which is the reason a smaller number of spoken dialogs were chosen.

The maximum and minimum number of utterances observed in the spoken conversations were 6 and 46. In comparison, the longest written dialog had only 30 utterances, and the shortest one contained 19. The difference in distributions of utterance counts per dialog between written and spoken dialogs can be observed in Figure 5.2. The distributions of utterance counts were found to be normal for spoken dialogs (p-value = 0.18) but not normal for written dialogs (p-value $<0.001^{***}$) from the Shapiro-Wilk normality test. Using the Mann-Whitney U test, the utterance counts' distribution for written dialogs was found to be significantly different (p-value $<0.001^{***}$) from spoken ones. Further discussion on similarities and differences between written and spoken dialogs is covered in the following section. To clarify the contribution and implication of the data analyses, the results are organized according to the proposed RQs in the following sections.

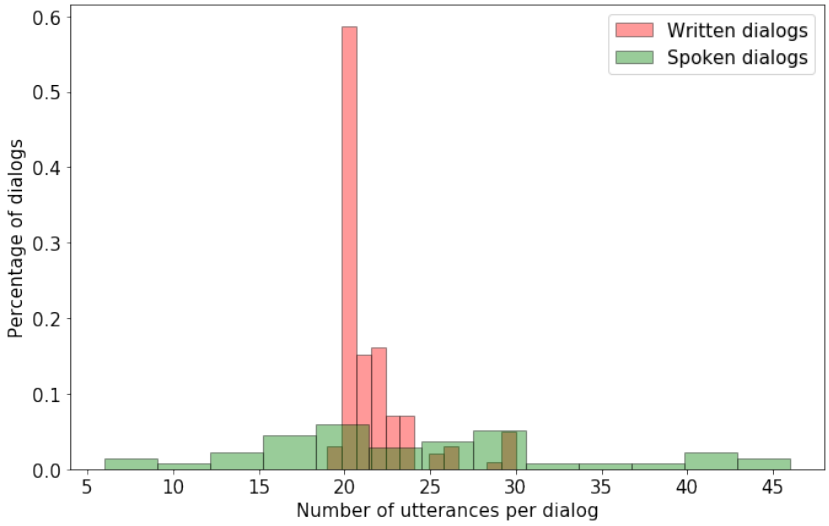


Figure 5.2: Distribution of count of utterances between written and spoken dialogs in the labeled data

5.2 Experiments for RQ1: The Type of Clarification Questions An Intelligent Agent Asks in Conversational Search

To suggest the type of clarifications a conversational system should ask the user in future dialogs, we need to analyze the negotiation process and identify the clarifications that were asked by the simulated system in our labeled information-seeking conversations. Each negotiation label (i.e., filter label) in the coded dialogs represents the labeled utterance's function in the negotiation process. Therefore, the analysis looked at the distribution of negotiation labels on the coded dialogs in this section. Additionally, by analyzing the conversational role labels, we can identify the utterances where the agent took the initiative and asked a question to the user. Thus, the analysis in this section also investigated the distributions of conversation roles between the two parties' utterances. The results of these analyses are reported in three parts. The first part discusses the distribution of all labels in the complete dataset, while the last two parts discuss the distribution of labels between two speakers separately. The second part is dedicated to the discussion of the negotiation labels' distribution in the information seeker's utterances, while the last part discusses the same for the agent. Analysis of both filter labels and conversational role labels are provided in each section. Additionally, the distribution of filter labels on the agents' utterances that are question types is also identified, and their frequencies are reported.

5.2.1 Analysis of the Complete Dataset

Distribution of filter labels

The annotators used a total of 3,791 filter labels to annotate the 3,598 utterances from 161 dialogs. None of the utterances were tagged with more than 2 filter labels simultaneously, which means that only 193 utterances (5.36%) were tagged with more than one label. The relatively low number of co-occurrences of filter labels could be due to two probable causes. First, most of the utterances were simple sentences, and therefore, either speaker could complete only one function of negotiation (as denoted by the filter labels) per utterance. Second, the low number co-occurrence suggests that the filter labels were mutually independent and did not have any significant overlap. Out of 3,791 labels, 3 (0.08%) labels had typos due to malfunctioning of the annotation tool which could not be resolved to the labels the annotators meant to use, and hence, not included in the rest of the analysis.

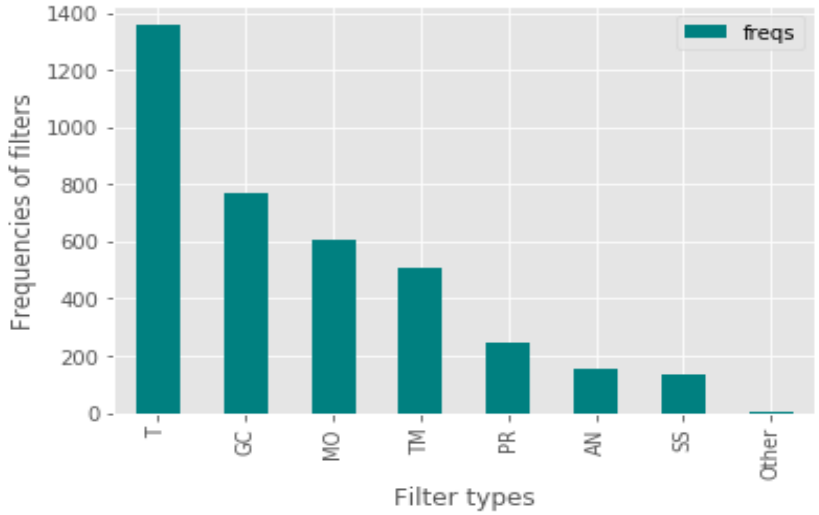


Figure 5.3: Distribution of filter labels in the labeled data (T = Topic. GC = General conversation, MO = Motivation, TM = Task management, PR = Preference, AN = Anticipation, SS = Search strategy)

The distribution of the remaining 3,788 labels among the eight types is provided in Figure 5.3. As shown in the figure, negotiations on the topic of information need (denoted by T) was the most frequently (1,355, 35.77%) observed label in the annotated data, which is in line with our expectations that the majority of the utterances in the information-seeking dialogs

contributed to the negotiations on topic for delineation of the information space. The annotators used the general conversations (GC) label 773 times (20.40%) to tag the utterances that did not contribute to the negotiation but were needed to maintain the communication as per social norm (e.g., exchanging greetings, closing rituals), suggesting the idiosyncrasies of human interactions even for focused conversations as the ones considered here. Among Taylor's five filters on question-negotiation process, motivation (MO) was the second most frequent label (607 times, 16.02%) after T observed in our data. Despite the "scripted" dialogs used in our experiments, the MO label's frequent occurrence suggests the importance of understanding the user's motivation behind the information-seeking activity in fulfilling the information need. User's preference (PR) was discussed on 248 utterances (6.55%), which were most likely used to personalize the search result.

In comparison, anticipating the type of information need (AN: 157 utterances, 4.14%) and discussion on alternate search strategies (SS: 135 utterances, 3.61%), from Taylor work, had the least presence in our labeled conversation, which suggests the relatively below par use of conversational query suggestion or query recommendation techniques in the simulated dialogs. As per the labels, the dialogs contained a considerable number of utterances (508 times, 13.41%) about performing some actions related to task management (TM), which points to the overhead of using a conversational search system for completing a complex task. Only on five occasions the coders felt none of the existing labels were appropriate to describe the utterance's function, and hence, used the other label (O). The low number of O labels (1.40%) suggests an excellent coverage of the filter labels in describing the utterances' functions. The following section discusses the distribution of conversational role labels on the annotated dataset.

Distribution of conversational role labels

As discussed in Chapter 4, the taxonomy of conversational role labels used in labeling was derived from the dialog acts in the COR (Sitter & Stein, 1992) model. These 10 labels are 'request', 'offer', 'reject offer', 'assert', 'promise', 'accept', 'be contented', 'withdraw request', 'withdraw offer', 'be discontented', and 'reject request'. Unlike the COR model, this study did not put any restrictions on the conversational roles for the speakers, i.e., annotators were

allowed to assign any roles from the labeling scheme to either speaker's utterances. The frequencies of role labels observed in the annotated dataset is provided in Figure 5.4.

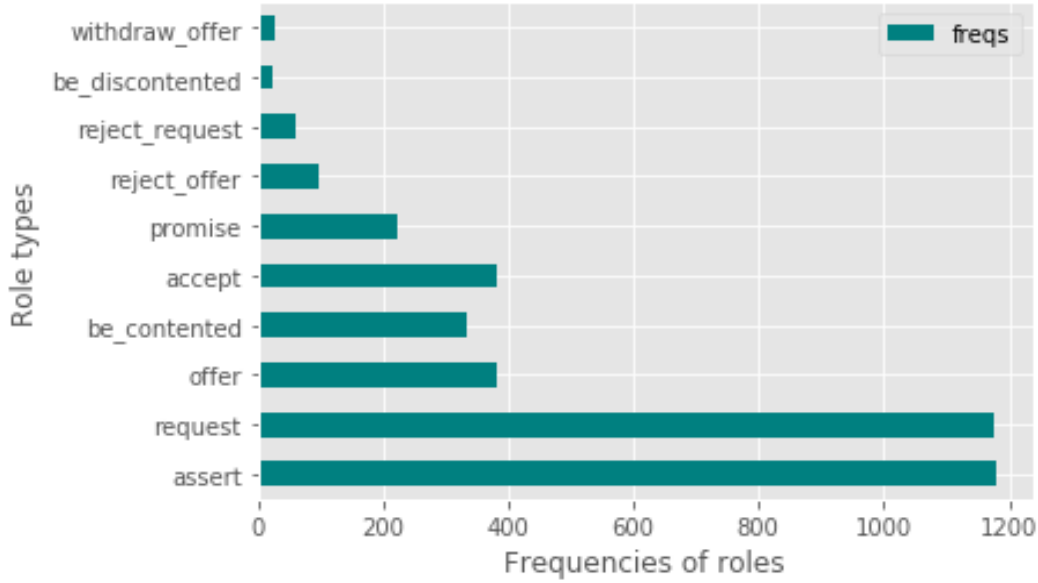


Figure 5.4: Distribution of conversational role labels for all utterances in the annotated corpus

In total, 3,898 role labels were assigned to the 3,598 utterances annotated from the dataset. No utterance was annotated by more than two role labels. Hence, on 300 occasions the same utterance was labeled with multiple conversational roles. Out of the 3,898 labels, 10 had typos which could not be resolved (0.26%) decisively. The rest of the analysis reported here was done after excluding these 10 labels. The most common role labels were 'assert' (1179 times) and 'request' (1176 times) types. Together these two types constituted 60.57% of all the role labels. The next most frequent labels in the utterances were of 'offer' (382 times) and 'accept' types (381 times), each appearing 9.83% in the labels. Noticeably, in dialog acts these two roles were complementary to each other, i.e., by taking the role of 'offer', the speaker could assign the role of 'accept' or 'reject offer' to the other party involved in the conversation. The role label 'be contented' was used to label 336 utterances (8.64%). In comparison, the 'promise' label appeared 223 times (5.72%). The role types 'reject offer', 'reject request', 'be discontented' and 'withdraw offer' were much less frequent in the utterances, with only 99 (2.55%), 60 (1.54%), 22 (0.57%) and 27 (0.69%) occurrences. The low occurrences of these labels can be due to mostly implicit use of these roles by the speakers in our annotated dialogs. A detailed

analysis of distributions of conversational role labels between the two parties is presented and compared in the following sections.

5.2.2 Analysis of Information Seeker's Utterances

Distribution of negotiation labels

A total of 1,779 user's utterances were recorded in the annotated dataset, and 1,876 negotiation labels (filters) were used to code these utterances. The most common filter label that appeared in this data was topic (T), which accounted for 37.37% (701 times) of all labels used on the user's utterances. The T labels' frequency was slightly higher than the same in the complete dataset. The general conversation (GC) label appeared 440 times (23.45%) and was the second most frequent label in the data. The distribution of all the filter labels in user utterances was very similar to the distribution of filter labels in the complete dataset (user and agent), except for the task management (TM) label, which was observed only 105 times (5.60%) in the user utterances. The low number of TM labels in the user's utterances suggests that the agent did most of the discussion and possibly, taking actions for task management while in conversation. The motivation label (MO) was used on 341 occasions (18.18%) to annotate the user utterances. Whereas the preference label (PR), alternate search strategy (SS), and anticipation (AN) were used 168 (8.96%), 72 (3.84%), and 42 times (2.24%), respectively. The distribution of the labels on user utterances was provided in Figure 5.5. Barring the seven labels discussed above, the 'other' label appeared on a single occasion and was not included in the figure.

Distribution of the conversational role labels

A total of 1,918 role labels were used to annotate all the user's utterances. All the 10 typos in labels detected in the complete dataset were from user utterances. Hence, the rest of the analysis reported here considered the remaining 1,908 role labels. Similar to the distribution in complete dataset, the most common role label appeared in the data was 'assert' type, which was used 701 times (36.74%). The 'request' type label was the second most freqent label (443 times). However, the frequency of the 'request' label was much less in the user utterance (23.21%), when compared to the frequency of the same in the complete dataset (30.27%), which suggests

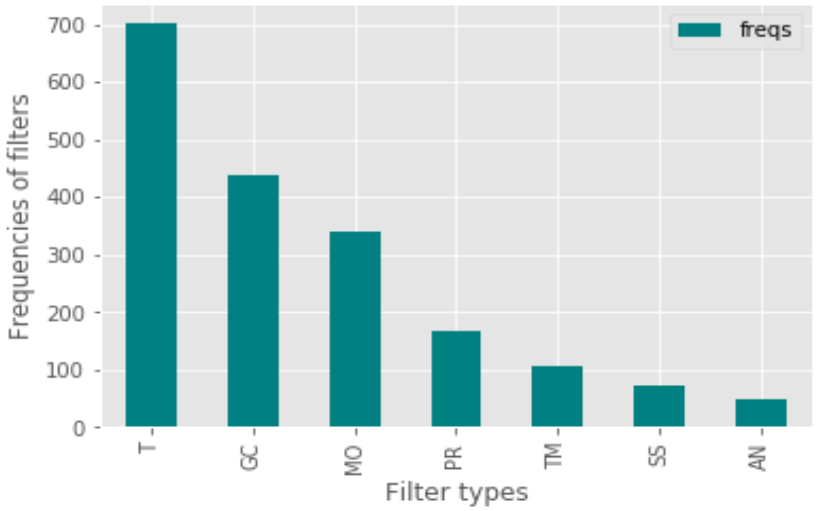


Figure 5.5: Distribution of filter labels in the information seeker's utterances

that the users explicitly requested the agent for information on fewer occasions and more times relied on responding to the agent's clarification requests to express their information needs.

The distribution of the labels on user utterances is provided in Figure 5.5. As shown in the figure, the 'reject request', 'offer' and 'be discontented' roles were rarely present in the user utterances, and were used on 15, 11 and 18 occasions respectively. The rarity of 'offer' role in user utterances was expected as this role was mostly taken by the intermediary in any information seeking dialogs. The low occurrence of 'reject request' suggests majority of the occasions when the intermediary had asked for information the 'user' responded favorably in the annotated dialogs. Whereas, 'be discontented' role was less frequent in user utterances due to two probable reasons. The user was satisfied in majority of the dialogs used in this analysis, or in case they were unsatisfied with the information provided by the agent, the dissatisfaction was not expressed explicitly. Unlike the compete annotated dataset, no occurrence of a 'withdraw offer' or a 'promise' label was observed in any user utterance. The absence of these roles in user utterances suggests that they were agent specific conversation roles, which also corroborates the use of these roles in the COR model (Sitter & Stein, 1992).

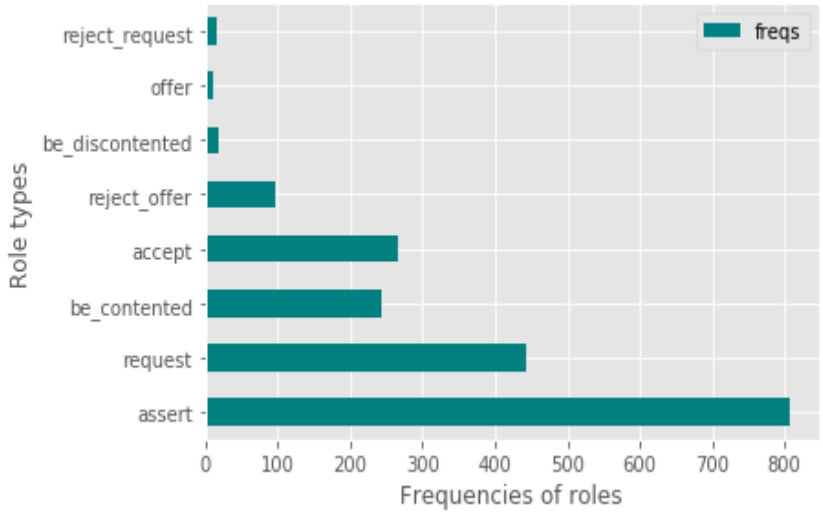


Figure 5.6: Distribution of conversational role labels in the information seeker's utterances

5.2.3 Analysis of the Agent's Utterances

Distribution of the negotiation labels

Out of 3,598 utterances in 161 dialogs, 1,819 (50.56%) were recorded as assistant's utterances. A total of 1,915 labels were used to tag these utterances. Out of the 96 utterances where more than one label co-occurred, on 68 occasions (69.89%), one of the co-occurring labels was either task management (TM) or general conversation (GC). Only on 28 occasions (30.11%), the annotators used more than one label from (Taylor, 1967)'s filters per utterance, which suggests that the dialog agent only occasionally completed more than one function of negotiation in a single utterance. Among all 1,912 labels used to annotate the assistant's utterances, GC appeared 333 times (17.42%), while TM was used on 403 occasions (21.08%). The distributions of Taylor's five filters in the agent's utterance labels were consistent with the distributions in both speakers' utterances. As shown in Figure 5.7, the topic of the user's information need (T) was the most frequent label observed with 654 counts (34.21%). The motivation label (MO) was present 265 times (13.86%), which suggests that the agent spent a significant amount of conversational turns discussing the user's motivation for the information need. In comparison, the presence of preference label (PR), anticipation (AN), and search strategy (SS) were less frequent in the assistant's utterances. The search strategy (SS) was observed 63 times (3.29%), whereas the anticipation label (AN) and preference (PR) label were used 109 times (8.31%) and

80 times (4.18%), respectively. On 4 occasions the annotators noted that none of the existing labels were appropriate (other label, or 'O') to code the function of the agent's utterances in the annotated dialogs. These 4 other tags were from 4 different dialogs.

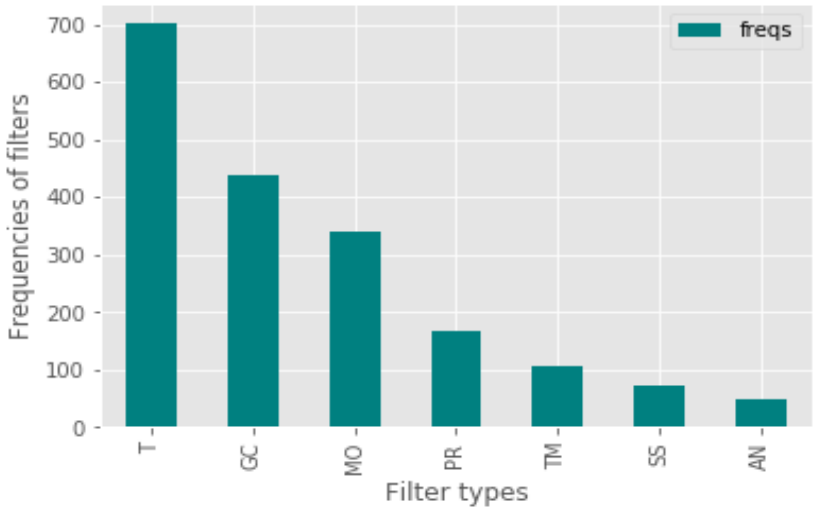


Figure 5.7: Distribution of filter labels in the dialog agent's utterances

Distribution of the conversational role labels

A total of 1,980 role labels were assigned to 1,819 assistant's utterances, out of which the most frequent one was 'request' type with 777 occurrences (39.24%). In fact, the agent utterances had more 'request' role labels than user utterances, which further corroborates the importance of clarifying questions by the agent in fulfilling user's information need. The frequencies of 'offer' and 'assert' roles were almost identical with 371 and 372 occurrences respectively and were the most frequent role labels after 'request'. Unlike user's utterances, 'promise' role label was more frequent on the agent's utterances with 218 appearances (11.01%). As per the annotators, on 45 occasions the agent rejected user's request ('reject request') and 27 times withdrew an earlier offer made to the user ('withdraw offer'). The 'reject offer' and 'be discontented' roles were almost non-existent in the agent' utterances with just 4 and 3 occurrences. The low frequencies of these two labels suggests that these roles are user specific. The distribution of the conversational role labels on the agent's utterances is shown in Figure 5.8.

The above two sections described the distribution of in the agent's utterances regardless of the nature of utterances. More specifically, the distribution considered all the agent's utterances.

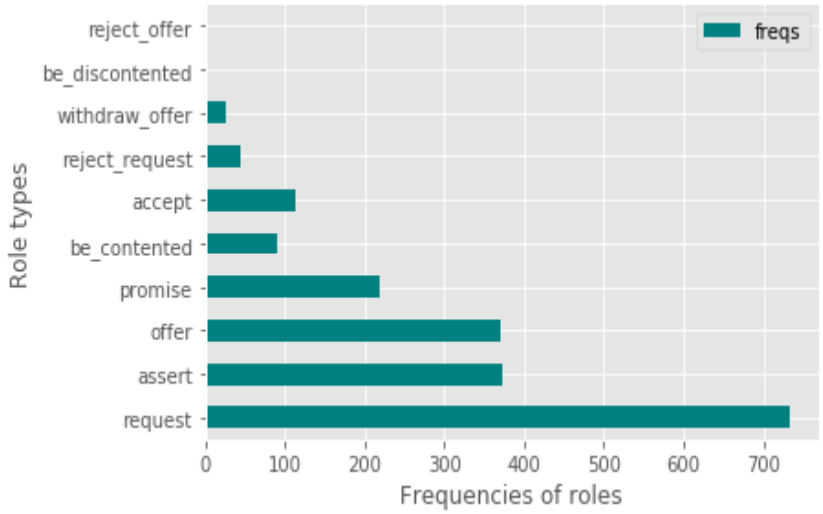


Figure 5.8: Distribution of conversational role labels in the agent's utterances

However, to identify a taxonomy of clarification questions, the utterances where the agent specifically asked a follow up question to the user, needs to be identified and the distribution of the associated negotiation labels needs to be evaluated. The question type utterances were identified from the conversational role labels, where the system assumed the role of request and in turn assigned the complementary conversation role to the user. The next section discusses the distribution of negotiation labels on the agent's clarifying questions.

5.2.4 Distribution of negotiation labels in agent's follow-up questions

In this section, only those utterances are considered where the agent took the conversational role of request as per the labels assigned by the annotators. A total of 777 clarifying questions were identified in this process and their labels were analyzed. The distribution of filter labels on these utterances is provided in Figure 5.9. As the figure suggests, the distribution of labels on the clarifying questions were in line with the distribution of labels on the complete set of agent's utterances. Overall, majority of the agent's questions were on user's topic, with 339 (43.63%) instances identified in the dataset. The next most frequent labels on the agent's clarification questions was motivation (MO). A total of 178 clarification questions (22.91%) were of this type. The frequencies of anticipation and preference labels were almost similar observed in agent's clarifying questions, with 62 (7.98%) and 54 instances (6.95%) respectively. Only on 18 occasions (2.32%) the agent offered an alternate search strategy (SS) to the user. As the

SS labels are expected to represent the conversational approach to query expansion and query suggestion behavior, it seems the agent in simulated dialogs used this approach rarely for the tasks. Moreover, it is evident that the frequencies of general conversation labels (GC) and task management labels (TM) in agent's clarifying questions were much lower than the frequencies of these labels in the complete dataset of the agent's utterances. Therefore, it could be safely said that most of the TM and GC labels in the agent's utterances were due to responding to the directives of the users towards their task completion.

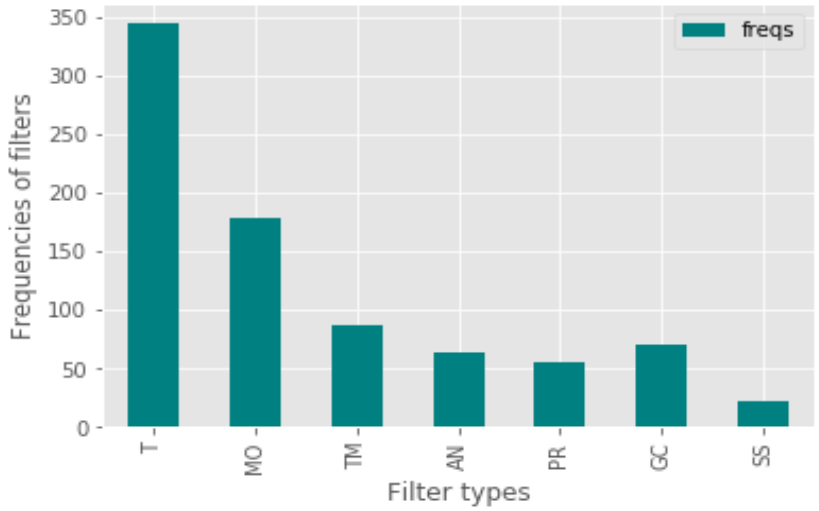


Figure 5.9: Distribution of filter labels in the agent's clarifying questions

The above statistics in the analysis of RQ1 suggests that both speakers, the user, and the agent spent a significant amount of conversational turns discussing the topic of user's information need and what motivated the user to seek the information. On infrequent occasions, the assistant took turns negotiating the type of information being anticipated by the user and discussing the user's preference. Only occasionally the assistant proactively suggested new search strategies for the task to the user. The above statistics used labels from the assistant's utterances in both written and spoken conversations. However, it is not clear how the modality of conversation may have impacted the negotiation process. The following section compares the filter labels' distribution on the assistant's utterances between written and spoken dialogs.

5.3 Experiments for RQ2: The Affect of Conversation Modality in the Clarification Types in a Conversational Search System

This section characterizes the affect of conversation modality on the clarification types that were identified in RQ1 by comparing their distribution between written and spoken subsets. This is done by analyzing the difference in filter labels' distribution on the assistant's utterances between written and spoken conversation. Since the 'Other' label (O) was rarely present in the data, this comparison does not consider the 'Other' label's distribution. Detailed statistics about the frequency distributions of filter labels per dialog are provided in Table 5.1. Consistent with the rest of the data, the topic label (T) appeared more frequently than any other filter from Taylor (1967)'s model. Out of 117 written dialogs, 107 dialogs (91.45%) had at least one utterance of agent labeled with T. Whereas, among 44 spoken conversations, 27 (61.36%) had at least one utterance of agent coded as T by the annotators. Among the 107 written dialogs, the T label appeared on the agent's utterances with an average (mean) of 5.21 times per dialog, whereas the mean number of T labels per spoken dialog was 3.56. The topic labels distribution in the data indicates that the agent needed to negotiate the topic of the user's information needs more frequently in written conversations, and while negotiating the topic, the assistant consistently had to use more conversational turns in written dialogs than spoken ones, which is one on the important findings from this research.

The distributions of the preference label (PR) and search strategy (SS) were similar between the written and spoken dialogs. These two labels appeared in 35.90% and 29.91% of written dialogs, respectively. Among all spoken dialogs, the PR label appeared in 29.95%, and the SS appeared in 20.45%. The mean number of labels per dialog for PR and SS labels were also similar between written and spoken conversations. In comparison, the frequencies of the motivation filter (MO) and the anticipation filter (AN) on the agent's utterances varied significantly between written and spoken dialogs. The MO filter appeared in 53 (45.30%) written and 32 (72.72%) spoken dialogs. Moreover, the mean number of MO filters per dialog was far greater for spoken dialog (4.34) than written ones (2.38). Thus, MO filters' distribution suggests that the agent consistently invested more effort into negotiating the user's motivation behind the information-seeking activity and did so more frequently in spoken conversations

Table 5.1: The negotiation labels' distributions per dialogs on agent's utterances in written and spoken conversations

Modality	**Filter label**	**No. of dialogs (%)**	**Maximum no. of labels per dialog (Min = 1)**	**Mean no. of labels**
Written	T	107 (91.45%)	11	5.21
Spoken	T	27 (61.36%)	10	3.56
Written	MO	53 (45.30%)	23	2.38
Spoken	MO	32 (72.72%)	16	4.34
Written	AN	20 (17.09%)	4	1.60
Spoken	AN	22 (50.00%)	9	3.50
Written	PR	42 (35.90%)	4	1.50
Spoken	PR	13 (29.55%)	2	1.31
Written	SS	35 (29.91%)	3	1.45
Spoken	SS	9 (20.45%)	3	1.33
Written	TM	107 (91.45%)	7	2.43
Spoken	TM	41 (93.18%)	8	3.49
Written	GC	95 (81.20%)	7	2.11
Spoken	GC	41 (93.18%)	9	3.22

than in written ones.

Out of all five Taylor's filters, only the anticipation label (AN) appeared in more spoken dialogs (20 counts, 17% of dialogs) than in written ones (22 counts, 50% of dialogs) despite the larger sample size (approximately 2.65 times) of the latter. For those dialogs where at least one AN label was used to code the agent's utterances, the average number of AN labels was more than double for written dialogs (3.5) than the spoken counterparts (1.6). The distributions of task management (TM) label and general conversation (GC) label in the agent's utterances were consistent and similar between written and spoken conversations. Both labels appeared in more than 80% of dialogs regardless of the dialog modality. For each of those dialogs, these two labels on average (mean) appeared between 2 and 4 times per dialog on the agent's utterances.

5.3.1 Differences in distribution of negotiation labels between written and spoken dialogs

The significance of the difference in filters labels' distributions on agents' utterances between the two modalities is also evaluated and reported in Table 5.2. As the table suggests, the difference in distribution between the two modalities was significant for all filter labels except for preference (PR) and search strategy (SS). Also, the significance level of the difference in distribution was high for the topic label, T (p-value = 5.13e-05***), motivation label, MO (p = 0.0009***) and the anticipation label, AN (p-value = 0.0014**). The differences in the distributions of filter labels were consistent with the reporting of Table 5.1. However, contrary to the indications from Table 5.1, the distributions of task management (TM) and general conversations (GC) labels were found to be significantly different between written and spoken dialogs.

Table 5.2: Difference in filter labels' distribution between written and spoken dialogs on the agent's utterances (p <0.05* , p <0.01**, p <0.001***)

Filter label	**Sample size**	**Difference in frequencies**		**Difference in relative frequencies**	
		Test statistic	p-value	Test statistic	p-value
T	134	751.0	p = 5.13e-05***	563.5	p = 4.96e-07***
MO	85	514.0	p = 0.0009***	665.0	p = 0.0486*
PR	55	249.5	p = 0.29	140.0	p = 0.0042**
AN	42	107.0	p = 0.0014**	142.5	p = 0.0258*
SS	44	131.5	p = 0.192	108.0	p = 0.0757
TM	148	1551.0	p = 0.0024**	1955.5	p = 0.1538
GC	136	1360.0	p = 0.0019**	1596.5	p = 0.0478*

The above findings of statistical difference in filter label's distributions on agent's utterances were based on total counts of filter labels irrespective of the number of utterances in dialogs. However, longer conversations with more than the average number of utterances per dialog might result in higher frequencies of filter labels' distributions than the normal, which may bias the significance test's output. As reported in section 5.1, the variance in utterance counts' distribution was significantly different between written and spoken dialogs. To address the effect of irregularities in utterance counts' distributions on the distribution of filter labels per dialog, relative frequencies of each filter label per dialog were used for significance testing. The relative frequencies of filter labels (rf) per dialog (d) were derived from the frequencies of a

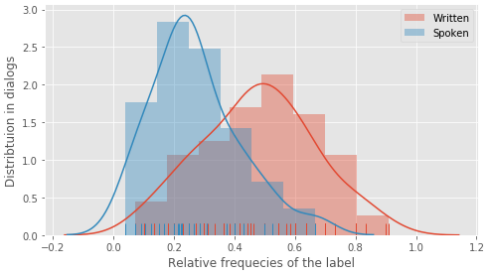
single filter label over frequency of all filter labels in a dialog. For example, relative frequency of i^{th} filter label ($rf_{i,x}$) was calculated using Eq (5.1) as:

$$rf_{i,x} = \frac{f(i,x)}{\sum_{j=1}^{n} f(i,x)} \tag{5.1}$$

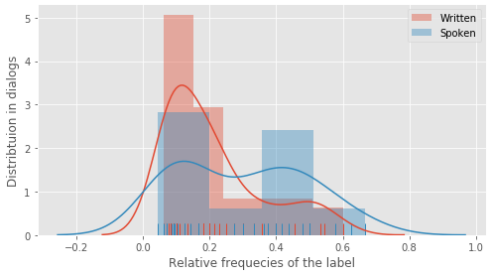
where $f(i,x)$ is the frequeny of i^{th} filter label in x^{th} dialog (d_x) and n is the total number of unique labels in d_x.

The difference in filter labels' relative frequencies per dialog between written and spoken conversations was evaluated using the same statistical test. The result for the same was provided in Table 5.2. Overall, the difference in the distribution of relative frequencies of filter labels between the two modalities was similar to the total frequencies' difference. For T, MO, and AN filters, the difference was significant in total counts as well as relative frequencies per dialog. For PR labeled utterances, unlike in total counts, the difference in the distribution of relative frequencies per dialog was found to be significant ($p = 0.0042^{**}$). For the SS label, the difference in the count of labels between written and spoken conversations was not significant either in total counts or relative frequencies per dialog ($p = 0.0757$). For the task management label (TM), contrary to the result from total counts per dialog, the difference in relative frequencies was found to be not significant ($p = 0.1538$) between written and spoken conversations. The distribution patterns of the labels between the two modalities of dialogs were shown in Figure 5.10.

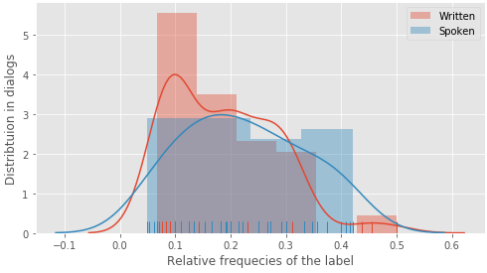
To summarize the result from the above section, it was observed that the agent invested a significant amount of conversational turns negotiating the topic of the user's information need regardless of the modality of conversation, which conformed to our expectation. Apart from the topic, the user's motivation behind the information-seeking episode was found to be the most frequent negotiation labels in the dataset regardless of the modality. In comparison, user's preference, anticipating the type of information acceptable to the user, and suggesting alternate search strategies that might fulfill the user's information need were not greatly invoked by the agent in the negotiation process. Despite similar negotiation themes in both modalities, a significant difference was observed in how much effort the agent had to invest in each negotiation themes between written and spoken dialogs. The analysis result suggests that the agent took



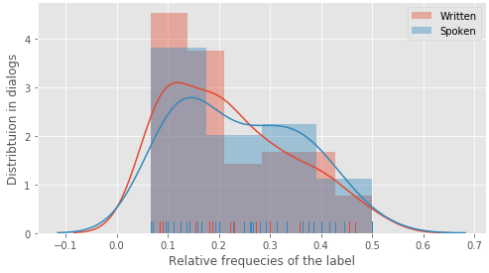
(a) Distribution of ‘T’ label



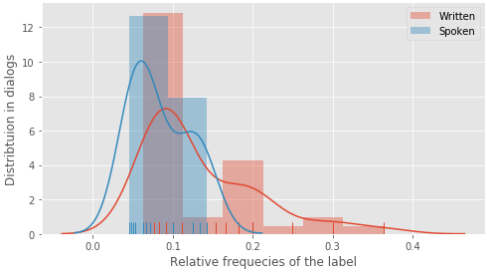
(b) Distribution of ‘MO’ label



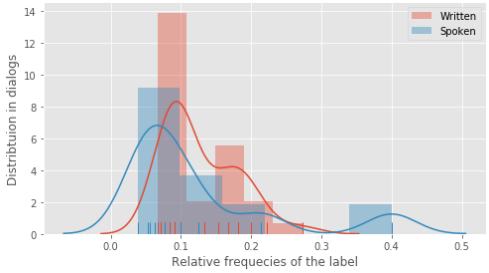
(c) Distribution of ‘PR’ label



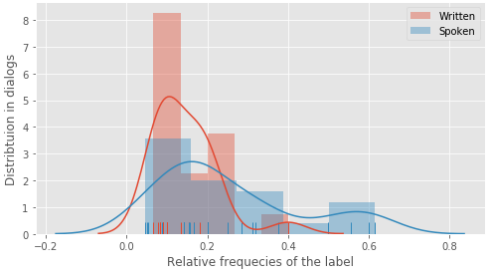
(d) Distribution of ‘AN’ label



(e) Distribution of ‘SS’ label



(f) Distribution of ‘TM’ label



(g) Distribution of ‘GC’ label

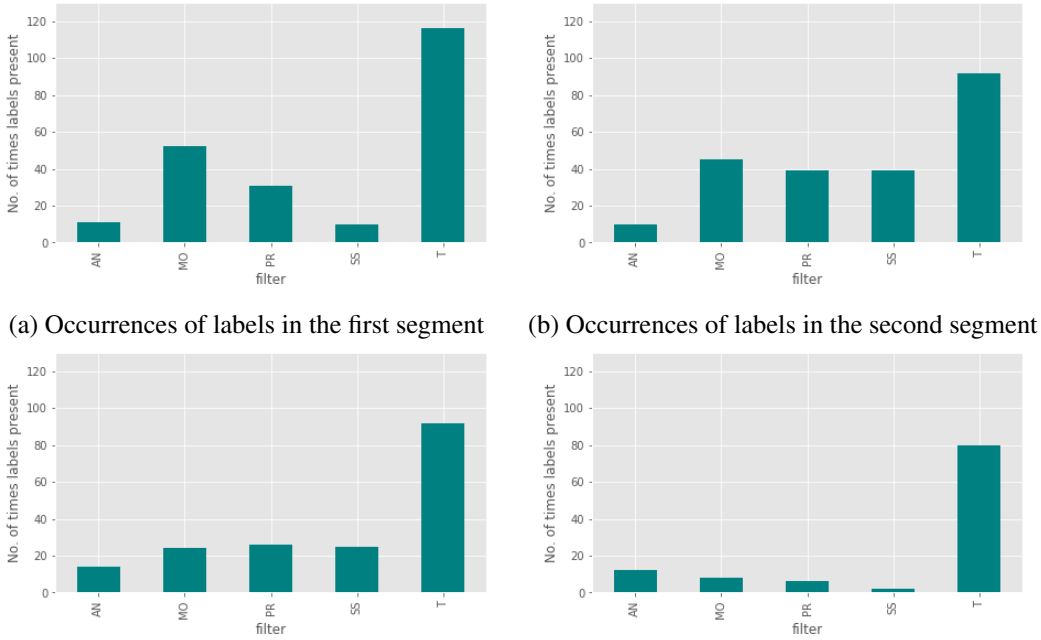
Figure 5.10: Relative frequencies of filter labels in the agent’s utterances compared between written and spoken dialogs

more conversational turns discussing the user's motivation and did so more frequently in spoken interactions than written ones. Moreover, unlike in written dialogs, the discussion on the user's motivation took precedence over the topic of information need in the agent's utterances in spoken conversations. Similarly, the agent negotiated the type of information the inquirer would accept more frequently and took more conversational turns to discuss this in spoken dialogs than in written ones. However, while offering alternate search strategies, no significant difference was observed in the agent's utterances between the two dialog modalities. The implications of these findings are discussed in detail in the next chapter.

The analysis from the above two sections highlighted what type of clarifications from the conversational agent might help the user to complete the task in hand. It also indicated how the modality of conversation might affect the type of clarification the agent could offer. However, it is still unknown when the conversational agent should ask for clarification from the user during a conversation about the latter's information need. The following section addresses this question.

5.4 Experiments for RQ3: The Relationship Between the Characteristics of User's Utterances and the Clarification Questions by an Intelligent Agent in a Conversational Information-seeking Dialogs

The previous sections' analysis presented a typology of clarifications in a conversational assistant's utterances that could be used to model the assistant's responses in future open-domain information-seeking conversations with the user. However, predicting an agent's appropriate response in future conversations with the user requires evaluation of the underlying patterns in use of clarifications types by the agent in prior dialogs. As a stepping stone in that direction, this section analyzes the order of clarification types in the assistant's utterances. Following the analysis from the previous section, the frequencies of clarification types in the agent's responses depended on the conversation modality; this section analyzes the order of the agent's clarification types for written dialog and spoken dialog separately.



(a) Occurrences of labels in the first segment

(b) Occurrences of labels in the second segment

(c) Occurrences of labels in the third segment

(d) Occurrences of labels in the fourth segment

Figure 5.11: Presence of the labels in the four segments of the written dialogs

5.4.1 Order of clarification types in dialog segments

Due to the high cost of obtaining human-labeled data, the number of data points with filter labels available at the dialog level (161 dialogs in written and spoken dialogs combined) was too small for any predictive modeling. Following Hendahewa and Shah (2013)'s work on segmenting information-seeking episodes, each dialog was segmented into four sequences with an equal number of utterances in each sequence to workaround the data scarcity. Thus, a total of 644 segments of dialogs were constructed from 161 written and spoken dialogs. In the next step, each label's presence was counted per segment for all dialogs collectively. Unlike the previous section's analysis, a label's distribution at the segment level was counted based on the presence or absence of the label instead of its frequencies (total or relative). The sections below describes the distribution of labels per segment in written and spoken dialogs. Since the analysis here was concerned with understanding the order of negotiation themes in agent's responses, labels that did not contribute to the negotiation process, e.g., general conversation (GC) label and task management (TM) labels, were ignored in this analysis.

5.4.2 Written dialogs

The presence of labels in the dialog segment level for the agent's utterances in all written dialogs is provided in Figure 5.11. The figure suggests that the clarifications on the user's topic of interest and anticipating the user's expectations were present almost equally in all four segments in the agent's responses. However, the distributions of the other three labels were different in all four segments. As shown in the figure, the motivation labels (MO) were predominantly present in the earlier segments of the conversations with the highest number of presence in the first (52) segments, before gradually decreasing in second (45) and third segments (24) to barely present in the last segments (8). Clarifications on the user's preference (PR) were also mostly present in the first and second segments (68.63%) before diminishing in the subsequent segments with barely present in the last segments (6 times, 5.89%). However, unlike the MO label, clarifications on PR labels appeared in higher numbers in the second segments (39 times) than the first segments (31 times). Clarifications on search strategies (SS) had a low presence in the first and second segments, with 10 (13.15%) and only 2 (2.63%) appearances, respectively. Most SS labels appeared in the second (39 times, 51.32%) and the third segments (25 times, 32.89%).

The above analysis result suggests a pattern in the order the clarification types appeared in the agent's utterances. Aiming to go beyond predefined dialog segments and explore the dynamic aspect of the complex negotiation process, the state transition patterns for utterance labels were examined in agents' utterances. From a conversation perspective, the difference in state transition pattern represents the divergence in the way the intelligent assistant negotiated the uncertainty in the user's information need and thus might help to disambiguate different paths of the complex negotiation process. Modeling state transition patterns can enhance understanding of how predefined negotiation functions and the task natures manifested dynamically in search sessions. Figure 5.12 illustrates the state transition patterns (including the frequency distributions of start states and end states) of all seven labels (including task management, and general conversation) assigned to the agent's utterances in written dialogs. Since the agent's utterance labels produced seven separate states, the edges with a transition probability of lower than 15% were omitted from this diagram to improve the clarity and readability. The complete

state transition table with transition probabilities is provided in Table 5.3.

Table 5.3: A state transition table for written dialogs with rows as source vertices and columns as target vertices.

	T	**MO**	**PR**	**AN**	**SS**	**TM**	**GC**
T	0.485	0.132	0.058	0.044	0.042	0.086	0.154
MO	0.264	0.337	0.066	0.048	0.048	0.105	0.132
PR	0.257	0.100	0.280	0.034	0.102	0.150	0.077
AN	0.361	0.034	0.076	0.178	0.048	0.119	0.183
SS	0.357	0.092	0.053	0.048	0.191	0.189	0.071
TM	0.197	0.048	0.058	0.032	0.098	0.094	0.474
GC	0.280	0.053	0.047	0.026	0.026	0.202	0.366

Overall, the results demonstrates the negotiation process of the user's information needs in complex search tasks in written conversations to be linear, with transition loops only within states (i.e., remaining in the same state). Loops involving transitions between states were observed involving only topic (T), general conversation (GC), and task management (TM) labels, which suggests that the discussion on the topic of interest went through several iterations. Furthermore, at the end of each such iteration, the topic possibly went through some modifications by completing some activities related to task management. In general, the probability of transitioning within states was observed to be higher than transitioning between states. However, transitioning probabilities within states were lower for search strategy (SS) and anticipation (AN) labels. Almost no direct transition was observed between search strategy (SS), anticipation (AN), and preference (PR) labels. Overall, the transition probabilities suggests that the agent used the other four negotiation functions except for topic (T) linearly and also exclusively per dialog, with almost no cross-transition involving the other four labels.

5.4.3 Spoken dialogs

Similar to the written dialogs, spoken dialogs were also segmented into four sequences. Thus, the analysis in this section is based on 176 sequences generated from 44 spoken dialogs. The presence of labels in each sequence of the dialog in all spoken dialogs is illustrated in Figure 5.13.

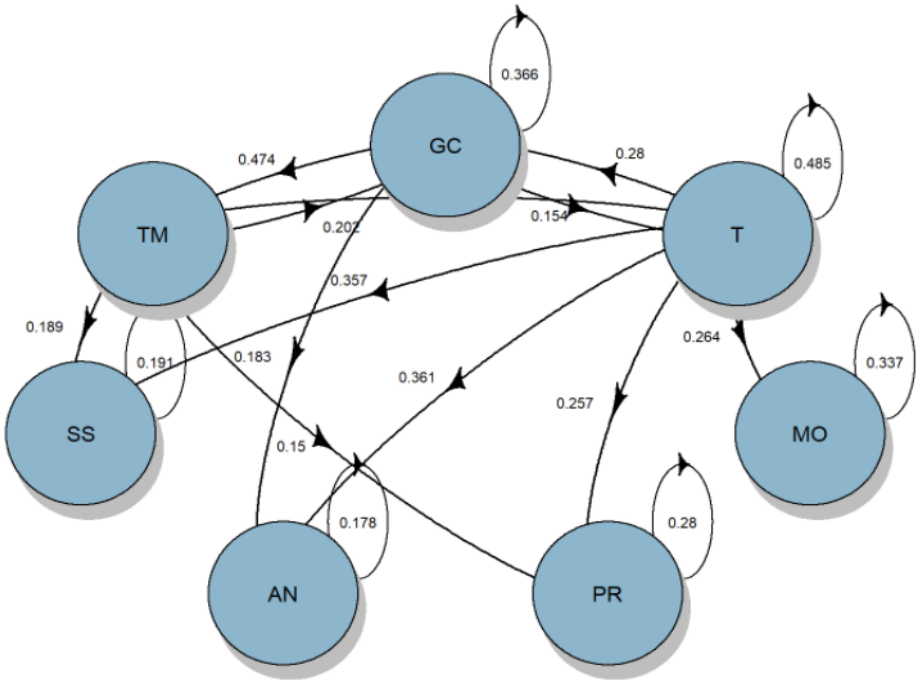
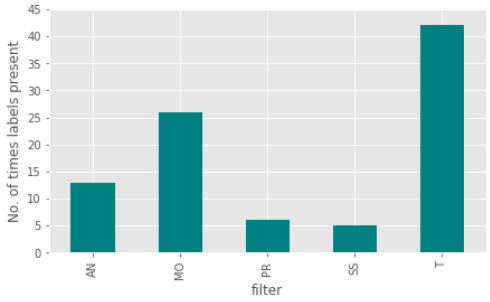
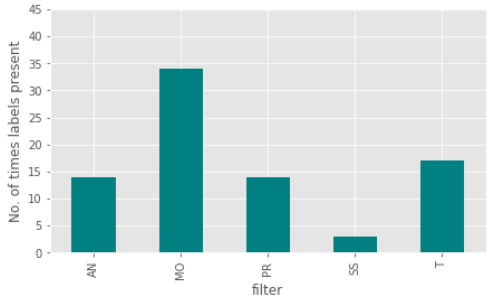


Figure 5.12: State transition of labels in agent's utterances in written dialogs

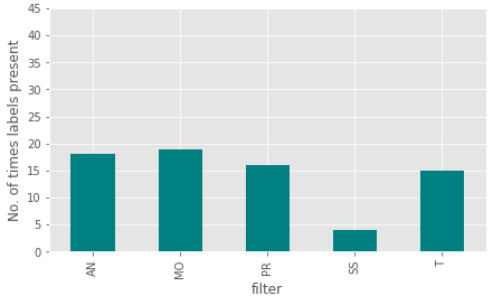
Overall, a more significant variation in the presence of labels in four sequences was observed in spoken conversations than in written dialogs. The first noticeable difference was for the topic label (T). Unlike in written conversations, the topic label (T) was present in the agent's utterances in a significantly smaller percentage in the last three sequences than in the first sequence. Out of 88 presence of topic labels in 176 sequences, 42 (47.73%) occurred in the first sequences. The preference label (PR) was detected only 6 times (14.29%) in both the first and the last sequences. Whereas, in the second and the third sequences, PR labels were more frequently present with 14 (33.33%) and 16 times (38.10%) successively. To this matter, the most noticeable difference between second and third sequences was observed in the MO labels' presence, with the number of second sequences with MO labels present was 1.79 times of third sequences, despite having the same number of utterances. In comparison, the anticipation labels (AN) were more evenly present in the first three sequences in 13 (25%), 14 (26.92%), and 18 times (42.86%), but less frequent in the final sequences (16.67%). A total of 92 presence of labels were detected in the first sequences, compared to only 34 presence



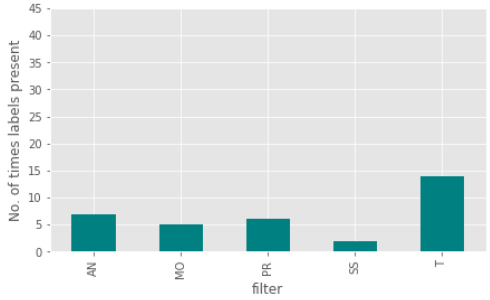
(a) Occurrences of labels in the first segment



(b) Occurrences of labels in the second segment



(c) Occurrences of labels in the third segment



(d) Occurrences of labels in the fourth segment

Figure 5.13: Presence of the labels in the four segments of the spoken dialogs

in the final sequences despite having the same number of utterances. The disparity is labels' distribution between the segments can be due to two probable causes. First, the first sequences might have more co-occurrences of labels in the agent's utterances; and second, the last sequences have significantly more labels from task management (TM) and general conversation (GC) category, which were excluded in this analysis. The first probable reasons could mean more cross-interactions between the labels than what was observed for written dialogs. This was further probed through analyzing the state transition diagram of spoken dialogs in the next paragraph.

The state transition patterns for spoken dialogs are illustrated in Figure 5.14. Similar to written dialogs, edges with transition probabilities less than 15% were omitted from this diagram. The complete transition table is presented in Table 5.4 Overall, the diagram suggests more non-linear transitions between the states in the negotiation process in spoken conversations than what was observed for written dialogs. For filter labels, transition loops involving the five filter labels were observed both within states and between states. The transition loops for

within states were observed for T, MO, and PR filter labels. Among these three labels, the transition within the same state had the highest probability for MO, which suggests that the agent in spoken dialogs spent multiple consecutive utterances understanding the motivation behind the user's information-seeking activity. The transition loops between states were observed involving AN and PR labels, which suggests that negotiation on the user's preference and anticipating the user's expectations followed each other in the negotiation process. Overall, comparing state transition patterns from written and spoken dialogs, it was evident that the agent used more non-linear and complex negotiation strategies in the spoken dialogs.

Table 5.4: A state transition table for spoken dialogs with rows as source vertices and columns as target vertices.

	T	**MO**	**PR**	**AN**	**SS**	**TM**	**GC**
T	0.219	0.326	0.089	0.121	0.042	0.065	0.138
MO	0.0.48	0.536	0.063	0.102	0.038	0.146	0.068
PR	0.156	0.099	0.174	0.161	0.062	0.299	0.049
AN	0.445	0.063	0.209	0.075	0.063	0.081	0.063
SS	0.163	0.059	0.068	0.130	0.083	0.225	0.273
TM	0.075	0.089	0.083	0.063	0.020	0.285	0.386
GC	0.171	0.130	0.043	0.086	0.034	0.117	0.419

5.4.4 Towards Prediction Models for Clarification Questions

The analysis presented in the above sections for RQ3 suggests that the clarification questions by the agent in information seeking dialogs appeared in a general order in the conversations and hence, any prediction model for clarification question responses by an automated agent should consider the utterance depth in the conversation as an important feature. However, the dialogs considered here and as a matter of fact, any information seeking dialog between a user and an intermediary should be user mediated conversation. Therefore, any prediction model for generating an automated agent's responses in future information seeking conversations should also consider the characteristics of the user utterances to predict the agent's response type. The analysis in the following section evaluates the relationship between the characteristics of the user's utterances and the intermediary's responses in our simulated conversations.

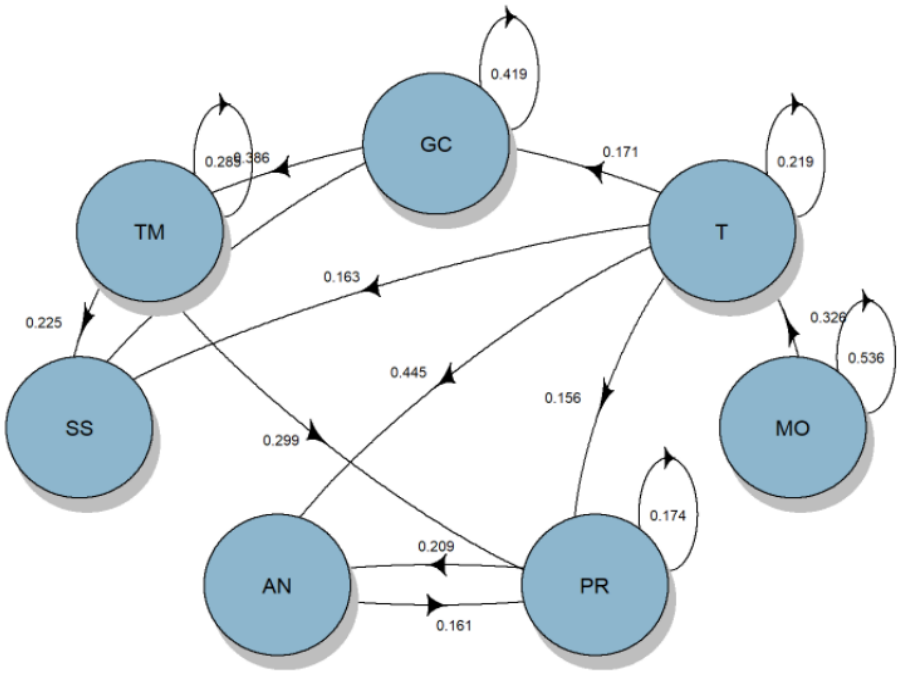


Figure 5.14: State transition of labels in agent's utterances in spoken dialogs

Characteristics of the user utterances

The characteristics of user utterances used in the experiments conducted in this section were a combination of lexical and semantic features of the utterances and described below.

- Word count: The number of words excluding the stop words present in the user utterances were used to characterize the user utterance. The word count feature represented the nature of user utterances (e.g., simple vs complex sentences) in the conversation. Word count feature is expected to represent important characteristics of conversations, including directness, informative, and the user's willingness to participate in the conversation.

- Presence of wh-words: The presence of 5W1H words (who, what, where, when, why and how) was computed and used as a boolean feature. The presence of such words in the user's prior utterance represents the scenarios where the agent temporarily refused to accept the user's request and opted to ask for more information.

- Similarity between consecutive utterances: To compute the similarity between two utterances, at the first step, each utterance was represented as a vector. A skip-gram model based on pre-trained embeddings trained on Wikipedia data (Wikipedia2Vec[1]) was used to obtain the embeddings. The dimension of each utterance vector was set at 100. Next, the word vectors were converted to utterance vectors by taking the normalized summation of all the word vectors in an utterance as shown in Equation 5.2. Distance between the current utterance ($\overrightarrow{u_i}$), and the last utterance ($\overrightarrow{u_{i-1}}$) was computed using the cosine similarity metric. The similarity measure was represented by the Equation 5.3. The lexical similarity between the two consecutive utterances is expected to reflect if there was any sudden change in the negotiation process that prompted the agent to ask a clarifying question.

$$\overrightarrow{t_i} = \frac{1}{N_i}\sum_{j=1}^{N_i}\overrightarrow{w_{ij}},\ and\ \overrightarrow{w_{ij}} = \overrightarrow{0}\ if\ w_{ij} \notin\ Wikipedia2Vec \tag{5.2}$$

$$Similarity(\overrightarrow{u_i}, \overrightarrow{u_{i-1}}) = cos\theta = \frac{\overrightarrow{u_i}.\overrightarrow{u_{i-1}}}{|\overrightarrow{u_i}||\overrightarrow{u_{i-1}}|} \tag{5.3}$$

Using the above set of features, experiments were conducted to predict when an automated agent should ask a clarification question and what type of clarification question it should ask at any point of time during a conversation. To simplify the first classification problem, the experiments conducted have only two classes: 'request' and 'other'. The 'request' class represented a clarification question, whereas other class denoted any 'other' conversational role taken by the agent. Due to the lack of sufficient data in spoken dialogs, the prediction models were created for the entire dataset (spoken and written). However, the mode of the conversation (spoken vs written) was used as a feature while creating the model.

To build the model, the entire dataset was divided into training and test cases. The training data consisted of 1278 assistant utterances, from 110 dialogs (65%), where a total of 447 clarification questions were present. The test data had 476 assistant utterances, from 57 dialogs (35%) out of which 149 were with 'request' role label and hence represented the clarification questions. 4 dialogs had typo in conversational role labels which could not be resolved and

[1] https://wikipedia2vec.github.io/wikipedia2vec/

hence, were excluded from the experiments. A conditional random field (CRF) model that considered the sequential dependencies between the features from neighbouring utterances was generated using the training data. The accuracy (f-measure) of the same model on the test data was found at 73.83%. Further details of the accuracy of the model was presented in Table 5.5.

Table 5.5: Accuracy of the CRF model in predicting the agent utterance role (TP = True Positive, FP = False Positive, TN = True Negative, TN = False Negative)

		Predicted response type	
		Request	**Other**
Actual response type	**Request**	237 (TP)	61 (FN)
	Other	56 (FP)	93 (TN)

An ablation experiment was conducted to investigate the importance of the features used for classification; the result is reported in Table 5.6. As the table suggests, the lexical features of user's utterances such as word count, and semantic features on similarities between the consecutive utterances were found to be most useful for classification when considered along with the utterance depth. In comparison, presence of wh-words in prior user utterances (i.e., user's questions) were found to less useful in this classification task.

Table 5.6: Ablation study of features used to classify the agent's response type (Request vs Other)

Features	**Accuracy (F-measure)**	**Loss**
ALL	0.7383	–
Word Counts + Utterance depth	0.6011	0.1372
Presence of Wh-words + Utterance depth	0.5807	0.1576
Word Counts + Presence of Wh-words + Utterance depth	0.7104	0.0279
Similarity between consecutive utterances + Utterance depth	0.6813	0.0570

In the next step, the same set of features, and training and test cases were used to identify the filter labels from the agent's utterances. Since this classification problem was a multi-class multi-label problem, multiple boolean classifiers were independently applied in this approach. For each class, a separate classifier was trained on the training cases and its prediction accuracy of filter labels was evaluated on the test cases. For example, the classifier to predict the topic filter 'T' labeled the test cases as 1 (for 'T') or 0 (for all other labels). The accuracies of

the CRF-based classifiers are presented in Table 5.7. As shown in the table, the classifier for predicting the topic filter (T) achieved lowest accuracy among all the classifiers with accuracy (measured in f-measures) of 0.60. Whereas, all the other classifiers achieved close to 90% accuracy in predicting the filter label for the agent's utterances. However, on closer inspection, such high accuracy was achieved due to the highly imbalanced nature of the data for these classes in both training and test sets. As a result, models generated for these classifiers were highly biased towards Class '0'.

Table 5.7: Accuracies of the CRF based classifiers in predicting the filter lables in the agent's utterances

Classifier	**Training data**		**Test data**		**Accuracy (f-measure)**
	Class '1'	**Class '0'**	**Class '1'**	**Class '0'**	
T	427 (34.21%)	821 (65.79%)	202 (45.39%)	243 (54.61%)	0.60
MO	183 (14.67%)	1065 (85.33%)	31 (6.97%)	414 (93.03%)	0.91
PR	59 (4.68%)	1203 (95.32%)	24(5.39%)	421 (94.61%)	0.97
SS	38 (3.04%)	1210 (96.96%)	20 (4.49%)	425 (95.51%)	0.95
AN	59 (4.73%)	1189 (95.27%)	37 (8.31%)	408 (91.69%)	0.89

The result from the last section suggests that the features used in the experiments, specifically characteristics of the user utterances worked well for predicting when a conversational agent should ask a clarification question instead of responding to user's request, however, did not perform so well for predicting the clarification question types. For the last classification problem, further experiments with classification algorithms that were better suited to handle multi-class multi-label classification problem are needed. Post-Hoc corrections on the CRF-based classifiers can also be worth exploring.

5.5 Summary

This chapter presented in detail the various analyses conducted on the task-based conversational dataset as labeled by the annotators. Through the analyses, a typology in the conversational agent's clarifications was identified from the dialog dataset (RQ1). The typology describing

the conversational agent's negotiation on the user's information need was domain and task-independent. How the modality of conversation could impact the clarification process was analyzed next (RQ2). Through the analyses, it was observed that the agent made a substantial effort in understanding the users' motivation behind the information-seeking activity in spoken conversations.

In contrast, the user's topic of interest garnered the most attention in written dialogs. Finally, the order of the clarifications on the agent's utterances was compared between the two modalities to find patterns in their appearances that can be used to model an agent's responses in future conversations with the user (RQ3). The analysis result suggests that compared to the written mode of conversation, the agent used a more non-linear and complex order of clarification types to negotiate the user's information need in spoken mode. Based on the characteristics of the user's utterances, CRF-based classification models were used to predict the position in a conversation where the agent should ask a clarification question and the clarification questions type. The classification model achieved high accuracy in predicting the positions of clarification questions in a dialog. However, similar models to predict the clarification question types were found to be highly biased due to the class-imbalance nature of the data.

Chapter 6

Discussion and Conclusion

In the experiments presented in Chapter 5, results obtained from analyzing utterance sequences of a user and an agent pair's dialogs were described. The dialogs were on the user's information need. Complex task scenarios were used to simulate situations where an average user is expected to face difficulties in expressing information needs accurately and thus likely to enter into a complex negotiation with the agent. In such scenarios, the user's initial utterances or queries may represent only partial information need. An intelligent system should be able to respond with appropriate clarifying questions that can help the users realize and express their information needs more accurately and completely. Analyzing the utterance sequences, a set of basic functionalities was identified that could be used to model clarification responses for an intelligent dialog agent for open-domain information-seeking conversations.

Most related work so far on this topic focused on the intention detection and slot-filling approaches for domain-specific task-based dialog systems. While such approaches were successful in multiple domains, e.g., ticket reservation systems (Hemphill et al., 1990) and train information (Aust et al., 1995), they typically required many domain-specific handcrafted rules, which hindered scaling up to new domains and, therefore, not suitable for designing open-domain dialog agents. More recent approaches explored developing memory network-based end-to-end goal-oriented dialog systems (Bordes et al., 2017) or were based on heuristics, e.g., ranking the best question from a pool of questions covering all available facets of the user's questions or queries (Aliannejadi et al., 2019).

In this work, we took a different approach towards solving the problem of what clarification questions to be asked by a conversational search agent. Analyzing the complex negotiation in dialogs of a user-agent pair in multiple task scenarios, we identified a taxonomy of clarification types that are task or domain-independent, and therefore scalable for developing open-domain

systems. We also analyzed the dialog modality's effect on the clarification response types. To this end, two modalities of conversation were considered, speech and text. The same clarification types were identified in conversations from both modalities, which suggests the taxonomy was conversation modality independent. Finally, we investigated the sequence of clarification types in agent's utterances from both modalities to identify if they appeared in specific orders in the negotiation process, which, in turn, can be used to predict a clarification an agent could ask in future conversations. To this end, prediction models that accounted for sequential dependencies of the utterances were used to investigate the relationship between the agent's utterances and occurrences of clarification questions in the conversations. While this study did not explicate the stages of user's information need throughout the task, however, the rich inventory of clarification requests by the agent at least supported the inability of users' initial queries to express their information needs accurately and further corroborated the necessity for information systems to generate clarification requests for the user. The findings of our experiments are discussed in detail in the following sections.

6.1 Overall Result and Implications

To answer the RQ1 or identifying clarification types in an intelligent agent's responses, we analyzed the conversations between the user and the intermediary pair in simulated task scenarios. The conversational data (Byrne et al., 2019) was collected in two methods, spoken dialogs were generated from Wizard-of-Oz experiments, where a human intermediary played the role of a wizard, i.e., an automated intelligent conversational system in this case. The written dialogs were created by simulated users (crowdsourced workers) through self-dialogs, where given a task scenario, the user envisioned and wrote down the complete conversation. A set of clarification types were identified from previous research on negotiation strategies a human intermediary (librarian) would use to negotiate the user information need (Taylor, 1967). Upon closer inspection of the utterances in data, the clarification taxonomy was further modified with addition of labels on negotiations on task management and maintaining conversation. We asked two human experts on handling users' information needs (e.g., librarians) to identify

similar clarification types in the wizard's utterances from simulated task scenarios. The analysis of the labels used by the two annotators followed. Our analysis suggests that the inventory of clarification types identified had excellent coverage of the agent's responses in our dataset. Out of 3,788 utterances, only on 5 occasions (0.13%), the annotators felt none of the clarification types provided in the taxonomy represented the utterances' goal in the negotiation process. Additionally, co-occurrences of labels on same utterances were less frequent when compared to the total number of utterances. Only on 193 occasions, utterances (5.36%) were tagged with more than one label, and none of the utterances had more than two labels. Such a low percentage of co-occurring labels could be due to a combination of two probable causes. First, the nature of human-computer interaction probably constrained both parties in their interactions, in particular for self-dialogs. Thus, the speaker could complete only one negotiation function per utterance. Secondly, the low co-occurrence of labels suggests that the clarifications types chosen as labels were mutually independent and did not have any significant overlap in their goal in the negotiation process.

Overall, the result suggests that the agents used most of their responses to clarify the topic of users' information need. Such clarifications were intended to identify a more precise information need often through questions on sub-topic information. Clarification of topics is expected to help the agent in delineation of the information space. A sample of the assistant's clarification questions on sub-topics are provided in the following conversation snippet (Table 6.1) where the user was looking for information to make a reservation in a restaurant. Most of the clarification questions on topic were task specific and therefore a slot-filling approach could be used to generate clarification questions of this type for an automated system in future conversations.

The agent's most frequent question to the users after topic was about their motivation, i.e., what motivated them to engage in the present information-seeking activities. A dialog snippet representing the assistant's clarification question on the user's motivation is provided in Table 6.2. The high frequency of clarifications on user's motivation despite the "scripted" dialogs used in the experiments is of interest, and highlights the importance on understanding the user's motivation in the success of conversational search approach beyond the topical match of user's need. Understanding the user's motivation and intention can lead to a deeper understanding of

Table 6.1: A snippet of restaurant reservation task dialog showing clarification questions on user's topic by the agent

USER:	Hey assistant, can you book a table for me for tomorrow's dinner?
ASSISTANT:	Sure, where would you like to go?
USER:	Founding farmers in Washington DC
ASSISTANT:	No problem, how many people and what time?

the task and thus not only helps in improving the search result but may also contribute to user satisfaction. The high frequency of users' utterances discussing their motivation in the dialogs suggests that they were open to explaining their motivation to the dialog agent. Unlike clarifications on the topic, the motivation questions typically did not require consideration of all facets of the task or user's information need representation. Therefore, a set of nonspecific motivation questions curated from previous dialogs could provide an excellent template to generate a dialog system's clarification questions of this type for future use.

Table 6.2: A snippet of auto repair task dialog showing clarification questions on user's motivation by the agent

USER:	I'd like to schedule an appointment with Intelligent Auto Solutions, please.
ASSISTANT:	Sure. Why are you wanting an appointment?

Compared to topic and motivation, the clarification questions on the user's preference in the context of the present information-seeking activity were less frequent in our data. Clarifying the user's preference can help the agent filter the retrieved results and further personalize the responses as per the user's need. Previous experiments (Das, De Francisci Morales, Gionis, & Weber, 2013) on automated questions to leverage user preferences for recommendation suggested a hierarchical structure of questions could be useful in eliciting the preference in closed domains. A dialog snippet involving the assistant's question on the user's preference is provided below, in Table 6.3:

Table 6.3: A snippet of ride booking task dialog showing clarification questions on user's preference by the agent

USER: Yeah, can I get a ride from Applebee's to AMC 20?

ASSISTANT: Sure. Did you want to use lyft or uber?

The other two types of clarification questions identified in the taxonomy were about the nature of the information anticipated by the user, and the suggestion of alternative search strategies. Both clarification types were rarely observed in our data (<5%) than the other three clarification types. The rare use of alternate search strategies by the conversational agent is of interest and requires further investigation, especially because previous experiments on query suggestion and query recommendation found these services to be useful in traditional query search system (Huang et al., 2003). For an intelligent system, generating questions to clarify the nature of the information a user anticipates could be particularly challenging. Designing such a system would demand a deeper understanding of the task user is trying to accomplish and the conversation semantics for a context-sensitive response generation.

The analysis of conversational role labels' distributions suggests that in complex, multi-faceted task scenarios as the one used in the experiments, the user relied heavily on the agent's clarification questions to express her information need accurately. In an intelligent intermediary system designed to interact with humans in natural language, the conversation is expected to be simple utterances, limiting how closely a user can express her information need in a single utterance. As evident in our dataset, the agent took significantly more utterances to ask requests (clarification) to the user than the user did to the agent. Only, a handful number of times, the user explicitly turned down the agent's clarification requests (15 times, 1.93%) or expressed discontent (18 times, 2.32%) with the information received from the system, which could be conscious or unconscious efforts on maintaining politeness the interactions (Brown, Levinson, & Levinson, 1987).

In RQ2, we examined if the conversation's modality affected the agent's clarification questions asked to a user on the latter's information need. Two conversation modalities were considered, written, and spoken conversations. The distribution of task domains among conversations

from both modalities was similar. Our dataset showed a greater range in distribution of utterance counts in spoken conversations than the written dialog set, despite having a much larger sample size in the latter (2.65 times). This observation suggests the underlying complexity and nuances of the voice-based search systems. The same taxonomy of clarifying questions identified in RQ1 was used to evaluate the difference in clarification questions the agent used between the two modalities. Our result showed the taxonomy had excellent coverage for both written and spoken clarifying questions. However, frequencies of clarification types in agent's utterances were significantly different between the two modalities. In spoken dialogs, the agent spent most of the effort understanding the user's motivation. As a result, the agent's utterances had more clarification questions on the user's motivation in both utterance level and dialog level for spoken dialogs than written ones. The difference in agents' effort towards understanding the users' motivation between the two medium is of interest, and requires further investigation, specially because our findings are in contrast to previous research on synchronous and asynchronous communication mode on electronic negotiations (Pesendorfer & Koeszegi, 2006) which suggested that synchronous communications were less friendly, and prone to more competitive negotiation behavior. Understanding the human motivation is arguable more friendly, affective behavior from the agent, and yet we observed less negotiations on users' motivation in written dialogs. The less emphasis on elicitation of user's motivation in written dialogs can also be due to limitations of the data collection procedure, i.e., due to self-dialogs. Since the self-dialogs were written by the same person for the both roles, it is highly likely that the lower frequency of MO label in written, as compared to spoken (i.e. with different people for each role) dialogs is because in written, the person playing the intermediary role already had knowledge of the motivation, whereas in the spoken, this was not the case. Comparing our results with written dialogs in which there are two different participants can be worth investigating.

In comparison, the topic of user's information need took precedence in the agent's clarification questions in written information need. Furthermore, between the two modalities, the agent enquired the user's personal preference more frequently in spoken conversations. These findings suggests that a voice-based dialog system for IR should take more initiative in understanding the user's motivation and user's preference than a text-based system. For alternate search strategies, no significant difference was observed in the agent's utterances between the

two dialog modalities. Clarifying questions of alternate search strategies type were rare for both written and spoken dialogs, and no significant difference was observed in their appearance in two modalities.

In RQ3, we investigated if there was a pattern in which the clarification questions appeared in agent's utterances. More specifically, we explored the sequence in which the agent used clarification questions to negotiate the user's information need. Finding a sequence in clarification questions along with the taxonomy of question types earlier established could lead to a dialog plan that could be useful to predict an automated system's clarification questions in future information-seeking conversation with a user. Overall, a pattern in transitions between the clarification question types was observed in the dataset. However, this transition pattern was found to be different between written and spoken conversations. In written mode, initial clarification questions were mostly on the topic of information need and the user's motivation. After initial negotiations, the agent clarified the user's preference and occasionally suggested alternate search strategies. Towards the end of conversations, the agent rarely asked any questions on the user's motivation behind the information-seeking activity or preference and mostly focused on negotiating the topic. In comparison, in spoken dialogs at the initial stage of negotiations, the agent asked clarifications mostly on the user's topic of interest. In the next stage, clarifications on user's motivation took precedence, followed by questions mostly on the nature of information the user was anticipating and user's preference. In spoken conversations, transition loops between user's anticipation and preference were observed. Back-and-forth questioning from both types of clarifications suggests that the user's response to either type of question influenced the other. Moreover, the presence of clarifying questions on topic throughout the conversations suggests that in complex task scenarios such as the ones used in the dataset, the user's topic of interest could go through several iterations and possibly through modifications throughout the negotiation process.

6.2 Limitations

As always, despite the lessons learned and broader implications from this dissertation study, the work presented here was not without limitations. Some of these limitations and their probable

impact on the outcome were discussed below.

6.2.1 Limitations of the data

As discussed in Chapter 4, the data used in this work were based on simulated conversations between a regular user and a conversational intermediary. The intelligent system's role was played by a human who had experience in handling user's information need through conversations (i.e., a trained call center operator). As a result, the nature of the data used in this dissertation work was dependent on the quality of the simulation. Similarly, the self-dialogs used to collect the written search dialogs had limitations as a data collection method. Since the same person envisioned the role of both the 'user' and the 'system', the dialogs collected in this process may not reflect the disfluencies, error-corrections and other idiosyncrasies of human-computer interactions. The tasks used to generate the dialog data were predefined and provided in advance by the experimenters (Byrne et al., 2019) and not the participants' own tasks, which might have reduced the uncertainty about information need and effected the participants' motivations playing the role of regular users. Additionally, the high number of task management related utterance in the dialogs suggests the selection of the tasks used in the experiments were more suitable for task-based dialog systems than for conversational information-seeking.

For labeling, only a small sample of spoken dialogs (44 conversations) was selected for annotation, which might have impacted the analysis for spoken dialogs. Furthermore, despite the training on labeling, disagreement on labels were noticed in the inter annotator agreement score which might have impacted the annotation quality. For analysis only utterance data was used along with labels, and no other behavioral data (e.g., time taken for an utterance, pauses) were not considered for the analysis. Using richer data accompanied by the user's behavioral measures during a conversation with the agent might help to predict an intelligent dialog system's response.

6.2.2 Limitations of the method

The method we used to produce the taxonomy of clarification questions had limitations. In the simulation, a human intermediary (a call center operator) playing the role of an intelligent dialog system asked all the clarification questions. Since we did not have access to the

recruits from Taskmaster-1 data collection playing intermediary's roles, a different set of participants with similar experience in dealing with users' information needs (librarians) was used to identify the clarification types. There might be a difference in how the intermediary used the clarification questions to negotiate with the user and how the second set of recruits interpreted it despite their similar background. An alternate approach to eliminate this difference could be to in the form of a user study, where the conversational data could be collected from Wizard-of-Oz settings, and at the end of a dialog, the wizard could be asked to identify what it was trying to accomplish each time it asked clarification from the user.

6.2.3 Limitations of the result

The results obtained in this study and their implications were not without limitations. First, the taxonomy of clarification questions was produced based on the data from one user study, Taskmaster-1. The same taxonomy needs to be applied in other conversational datasets collected in similar contexts to evaluate the taxonomy's reproducibility and generalizability across all information-seeking dialogs. Secondly, the taxonomy was not perfect, as some of the clarifying question types (e.g., alternate search strategy) identified were rarely present in the data, and therefore the viability of keeping them needed to be evaluated in future studies. Third, the tasks used in the simulations were all complex tasks, but their complexities were different in task facets (Y. Li & Belkin, 2008). This study did not address the relation between task facets and clarification question types. More importantly, a regular user could be expected to accomplish various information-seeking tasks with different complexities with a conversational IR system, and not all task situations should demand clarification questions from the system. Therefore, the task complexities needed to be predicted with reasonable accuracy to design a system that could respond with clarification questions if needed.

6.3 Future work

The fundamental assumption of this dissertation work is that current conversational systems do not take enough initiative in the conversation for use in open domain information retrieval purposes. As use cases for such a system include various tasks of different complexities, a true

conversational IR system should determine when it should take the initiative, ask follow-up questions to the user to clarify her information need, and determine what clarifying questions to ask. To that end, this dissertation presents a comprehensive analysis of the agent's utterance types to understand the nature of negotiation in complex search tasks, which covers identifying the clarification types (RQ1), examining the effect of mode of dialog on clarifications (RQ2), and presenting state transitions between the types of clarifications in complex task scenarios (RQ3). The next step in this research direction is to combine the clarification types identified in our study with the current research on natural language understanding to generate clarifying questions in complex task scenarios and evaluate their effect in future conversations. Based on the current literature on conversational systems, both rule-based (e.g., slot filling mechanism) and reinforcement learning approaches are found to be useful and can be promising choices in this direction.

Furthermore, future research should address some of the limitations of our study, as pointed out earlier in this chapter. As mentioned, our analysis data consisted of only 44 task dialogs in spoken conversations. To evaluate the stability of our inventory of clarification types, we plan to collect more annotated data on both written and spoken dialogs and from a mix of tasks of various complexities. Getting conversational data annotated by skilled humans (e.g., librarians) can be expensive. A weak supervision approach can be useful to generate more annotated data in this direction. The nature of the task facets and its relation to the clarification types need to be explored to make the result more generalizable across all task types. Moreover, we need to develop an appropriate evaluation strategy of the impact of the agent's clarifications on multiple aspects of the conversation, e.g., conversation length, user's motivation, satisfaction, and task outcome. Also, in this work, the study data consists of mostly single-turn interactions, where a separate turn is used for each clarification type. Future research should analyze a more balanced dataset involving multi-turn interactions, in which the users answer to multiple clarifications in the same utterance. The user behavior and preference may change in a multi-turn setting, which is worth exploring. Finally, the endpoints of a human-human conversation are often delineated. How it affects user behavior in human-machine conversations and its consequence on clarification questions need to be explored. Generating session-aware clarifications involving short and long-term context can be of importance in future research in this direction.

Research on conversational search systems needs to overcome a lot of potential challenges involving understanding the users' needs and algorithmic development. Some of the primary research directions could be to enhance answer coherency, to understand long- and short-term user preference, to present results in a multi-modal setting, and to develop better evaluation strategies appropriate for conversational search. While a futuristic search agent is expected to possess all the above-mentioned intelligent capabilities, this thesis explored only one of the intelligent functions. Overall, our work provides insights into the design and development of conversational search systems in a more user-centered way.

Appendix A

A Codebook for Labeling Dialogs between a User and a Conversational System

A.1 Description of the work

In this annotation work, you are going to label the utterances in task-based two-party conversations, between a user (U) and an agent (A: an expert intermediary system). Each conversation falls into one of the following six domains: ordering pizza, creating auto repair appointments, setting up ride service, ordering movie tickets, ordering coffee drinks and making restaurant reservations. The goal of the annotation work is to comprehend what is being discussed/negotiated between the two parties and the conversational roles of each while taking turns in conversation.

A.2 Coding Scheme for Classifying What is being Discussed or Accomplished

This coding scheme is based on Taylor's (1967) five filters that the expert intermediary uses to elicit the inquirer's information need as a form of negotiation, a directed and structured process. The description of five filters are given below:

- Determination of the subject (Topic): This filter determines the limits and provide some delineation of the information space. For example, after applying this filer on the inquiry posed by the user, the agent may ask follow-up questions 'Is this what you mean' or 'Is this in the ballpark' as response. Following is a snippet of dialog that where both parties use this filter to move the negotiation forward.
  - U: I'm looking for information on aftermarket car radios for my 2011 Hyundai Sonata.

  - A: Ok, what are you thinking about?

- Motivation and Objective of the Inquirer (Motivation): This filter is about: why does the inquirer want this information? What is the objective? What is the motivation behind this inquiry? It may further distill the subject or may even alter the meaning of the entire inquiry. Following are two example utterances the belong to this label.

  - A: What features are you looking for in the aftermarket radio?
  - U: I want something with Apple CarPlay and superior sound.

- Personal Background of the Inquirer (Preference): The third filter affects the personal background of the inquirer and may not be limited to, following type of questions: 'What is his background? Has he used the system before? What's the relationship between his current inquiry and what he already knows?' etc. Answers to these types of questions help the system to determine the urgency, the negotiation strategy, level or depth of any dialog, and the critical acceptance of search results etc. A snippet of a dialog that depicts use of this filter given below:

  - A: Have you looked at any specific after-market radio before?
  - A: If you have any preference about manufacturer, I can look it up for you.

- Relationship of Inquiry Description to File Organization (Search strategy): Through this filter, the intermediary or the information specialist interprets and restructures the user's inquiry that best fits for effective retrieval purposes. For example a user looking for 'places that serve French toast in the evening' may not yield any result from the system but the system may decide to restructure the query as 'places that serve breakfast all day' for better recall. Following are two example utterances that belong to this category:

  - A: Have you tried Crutchfield? They have excellent selection of car radios.
  - A: Have you looked at any manufacturer who makes good radios?

- What Kind of Answer Will the Inquirer Accept (Anticipation): When an inquirer approaches the information system, he has some picture in mind as to what he expects the

information to look like, e.g., it's specificity, format, modality etc. This filter elicits information about what the user is expecting the information to look like. Following is a snippet of conversation showing the use of anticipation filter.

  - A: Are you looking for instruction manual on how to install it?
  - U: Oh yes! Any video instructions will be great.

Apart from the above five filters, some utterances may represent intermediary's functions that are not associated with eliciting the user's information need but rather managing some aspects of the task in hand or performing some functions of communication management. Description of these two classes are given below.

- Task management (TM): Utterances in this category represent the speaker performing some action that deals task management (e.g., asking the status of a process "Are we done?", "Should I book the table?"), or giving some action directives ("please hold", "click the button") etc.

- General conversation (GC): Task based dialogs involving two parties may include utterances that are not associated with negotiation of the task or task management but rather serve the function of communication management, e.g., Greetings ("Hello", "Good bye") or acknowledgement ("OK", "Sure"). Such utterances are to be labelled as 'general conversation'.

- Other: Where none of the above seven labels apply to an expert's utterance the annotator may choose to label such cases as "Other" category.

Important Notes: All utterances must be labeled. Some of the utterances may accomplish more than one filter, and coders may choose any number of labels from the first six categories in such cases, but not the "other" category. The "other" label cannot coexist with any label in the same utterance. Each label must be accompanied with a brief explanation (1 or 2 sentences) of what the utterance represents. Below is an example of complete dialog annotated in this scheme.

Table A.1: A sample conversation labeled with the coding scheme based on Taylor's filters (1967)

Utterance ID	**Speaker: Utterance (U: user, A: agent)**	**Filter**	**Filter reasons**
1	U: Hi, I'm looking to book a table for Korean food.	topic	The user describes the topic of this conversation as booking a table for Korean food.
2	A: Ok, what area are you thinking about?	topic	Agent elicits for more information on the topic
3	U: Somewhere in Southern NYC, maybe the East Village?	topic	User clarifies the initial topic description with more information.
4	A: Ok, great. There's Thursday Kitchen, it has great reviews.	general conversation, motivation	Agent's acknowledgement followed by a suggestion with reasoning with hope the reasoning will match user's motivation.
5	U: That's great. So I need a table for tonight at 7 pm for 8 people. We don't want to sit at the bar, but anywhere else is fine.	motivation, preference	Further clarification on motivation is provided by the user followed by details of preferences.
6	A: They don't have any availability for 7 pm.	task management	Agents response in negative saying it cannot proceed with the task.
7	U: What times are available?	search strategy	User suggests change in search strategy from looking for restaurants to loo for times/slots when tables are available.
8	A: 5 or 8.	task management	Agents responds to user's request
9	U: Yikes, we can't do those times.	task management	User responds that agent's retrieved information is not helping in completing the task.

A.3 Coding Scheme for Conversation Roles

In this coding scheme, utterances of both parties are to be labeled as dialog acts, as per the Conversational Role, abbreviated as COR (Sitter & Stein, 1992) model. Each dialog act represents the social role a speaker takes on in the current utterance while assigning the complementary role to the hearer.

Speaker (Agent: A, User: U)	**Utterance**	**Dialog act**
U	"When does the next WSDM conference start?"	request (U, A)

For example, with the utterance shown above, "When does the next WSDM conference start?" the user is taking the role of request and assigning the complementary role to the agent, and therefore labeled as *request(U, A)*. Following are the set of social roles that are permitted as labels in this coding scheme with examples:

Table A.2: A sample conversation labeled with conversational roles as per the COR model (Stein & Maier, 1995)

dialog Act	**Utterance (Speakers: A, B), <Role >**
request	A: When does the next WSDM conference takes place? <request(A, B) >
offer	B: In March 2021, <offer (B, A)>
reject offer	A: But when in March? <reject offer(A, B) >
assert	B: I don't know. <inform (A, B) >
promise	B: OK, I'll have a look <promise (B, A) >
accept	A: OK. <accept (A, B) >
be contended	A: Thanks <be contended (A, B) >
withdraw request	A: Never mind. <withdraw request (A, B) >
withdraw offer	B: Sorry I can't find the schedule in the invitation <withdraw offer (B, A) >
be discontented	A: Can I have at least the dates? <be discontented (A, B) >
reject request	B: I don't have the dates either <reject request (B, A) >

Appendix B

Pre-study Documentations

B.1 Institutional Review Board approval

RUTGERS eIRB

Arts & Sciences IRB - New Brunswick
335 George Street
Suite 3100, 3rd Floor
New Brunswick, NJ 08901
Phone: 732-235-2866

Health Sciences IRB - New Brunswick/Piscataway
335 George Street
Suite 3100, 3rd Floor
New Brunswick, NJ 08901
Phone: 732-235-9806

Health Sciences IRB - Newark
65 Bergen Street
Suite 511, 5th Floor
Newark, NJ 07107
Phone: 973-972-3608

DHHS Federal Wide Assurance Identifier:
FWA00003913
IRB Chair Person: Beverly Tepper
IRB Director: Michelle Watkinson
Effective Date: 6/10/2020
Approval Date: 5/30/2020
Expiration Date: N/A

eIRB Notice of Approval for Initial Submission # Pro2020000991

STUDY PROFILE

Study ID: Pro2020000991
Title: Clarifying User's Information Need in Conversational Information Retrieval Systems

Principal Investigator:	Soumik Mandal	**Study Coordinator:**	Nicholas Belkin
Co-Investigator(s):	Nicholas Belkin	**Other Study Staff:**	There are no items to display
Sponsor:	There are no items to display	**Approval Cycle:**	Not Applicable
Risk Determination:	Minimal Risk	**Device Determination:**	Not Applicable

Review Type:	Exempt	**Expedited Category:**	N/A	**Exempt Category:**	2 4
Subjects:	Unlimited	**Specimens:**	N/A	**Records:**	N/A

CURRENT SUBMISSION STATUS

Submission Type:	Research Protocol/Study	**Submission Status:**	Approved
Approval Date:	5/30/2020	**Expiration Date:**	N/A

Pregnancy Code:	No Pregnant Women as Subjects	**Pediatric Code:**	No Children As Subjects	**Prisoner Code:**	No Prisoners As Subjects

Protocol:	Research Protocol	**Consent:**	Consent form.pdf	**Other Materials:**	Recruitment email

ALL APPROVED INVESTIGATOR(S) MUST COMPLY WITH THE FOLLOWING:

1. Conduct the research in accordance with the protocol, applicable laws and regulations, and the principles of research ethics as set forth in the Belmont Report.

2. **Continuing Review:** Approval is valid until the protocol expiration date shown above. To avoid lapses in approval, submit a continuation application at least eight weeks before the study expiration date.

3. **Expiration of IRB Approval:** If IRB approval expires, effective the date of expiration and until the continuing review approval is issued: **All research activities must stop unless the IRB finds that it is in the best interest of individual subjects to continue. (This determination shall be based on a separate written request from the PI to the IRB.) No new subjects may be enrolled and no samples/charts/surveys may be collected, reviewed, and/or analyzed.**

4. **Amendments/Modifications/Revisions**: If you wish to change any aspect of this study, including but not limited to, study procedures, consent form(s), investigators, advertisements, the protocol document, investigator drug brochure, or accrual goals, you are required to obtain IRB review and approval prior to implementation of these changes unless necessary to eliminate apparent immediate hazards to subjects.

5. **Unanticipated Problems**: Unanticipated problems involving risk to subjects or others must be reported to the IRB Office (45 CFR 46, 21 CFR 312, 812) as required, in the appropriate time as specified in the attachment online at: https://orra.rutgers.edu/hspp

6. **Protocol Deviations and Violations**: Deviations from/violations of the approved study protocol must be reported to the IRB Office (45 CFR 46, 21 CFR 312, 812) as required, in the appropriate time as specified in the attachment online at: https://orra.rutgers.edu/hspp

7. **Consent/Assent**: The IRB has reviewed and approved the consent and/or assent process, waiver and/or alteration described in this protocol as required by 45 CFR 46 and 21 CFR 50, 56, (if FDA regulated research). Only the versions of the documents included in the approved process may be used to document informed consent and/or assent of study subjects; each subject must receive a copy of the approved form(s); and a copy of each signed form must be filed in a secure place in the subject's medical/patient/research record.

8. **Completion of Study:** Notify the IRB when your study has been stopped for any reason. Neither study closure by the sponsor or the investigator removes the obligation for submission of timely continuing review application or final report.

9. The Investigator(s) did not participate in the review, discussion, or vote of this protocol.

10. Letter Comments: *There are no additional comments.*

B.2 A Sample Recruitment Letter

Hi X,

My name is Soumik Mandal, and I am a PhD candidate at the department of Library and Information Science in SC&I, Rutgers, New Brunswick campus. I'm reaching out to you to see if you can help with my dissertation study for which I'm looking to recruit three librarians.

The research study is on clarifying user's information need in conversational search system. Participation in this study consists of annotating a conversational dataset (approximately 100 dialogs) in two coding schemes.

An ideal candidate should meet all the following requirements:

- a native English speaker.
- a librarian with experience (2 years) in handling library users' information problems.
- should have some qualitative coding experience.

No recruit specific information, including name, age, gender, and e-mail will be collected for the purpose of this study. The complete annotation work can be done remotely, and no in-person meeting is required. Upon completion of annotation work (approximately between 3-4 hours), each participant will be compensated by $75 for participation.

The study has been approved by the Institutional Review Board (IRB) at Rutgers [Pro2020000991] and is supervised by Dr. Nicholas J. Belkin (copied here) at SC&I.

It would be immensely helpful if you can participate in the study. Please let me know if you have any questions. Also feel free to pass along this email among your colleagues who may fit the recruitment criteria.

Thanks very much.

Best wishes,

Soumik Mandal

References

Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2018). Target apps selection: Towards a unified search framework for mobile devices. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 215–224).

Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 475–484).

Allan, J. (2004). Hard track overview in trec 2004 (notebook), high accuracy retrieval from documents. In *The thirteenth text retrieval conference (trec 2004) notebook* (pp. 226–235).

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, *15*(3), 143–178.

Al-Rfou, R., Pickett, M., Snaider, J., hsuan Sung, Y., & Strope, B. (2016). *Conversational contextual cues: The case of personalization and history for response ranking* (Tech. Rep.). Retrieved from `https://arxiv.org/abs/1606.00372`

Aust, H., Oerder, M., Seide, F., & Steinbiss, V. (1995). The philips automatic train timetable information system. *Speech Communication*, *17*(3-4), 249–262.

Avula, S., & Arguello, J. (2020). Wizard of oz interface to study system initiative for conversational search. In *Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 447–451).

Avula, S., Chadwick, G., Arguello, J., & Capra, R. (2018). Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 52–61).

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). Query recommendation using query logs in search engines. In *International conference on extending database technology*

(pp. 588–596).

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, *5*(1), 133–143.

Belkin, N. J. (1982). Models of dialogue for information retrieval. In *Proceedings of the 4th international research forum in information science. boras, sweden: Skrifter fran högskolan i boras* (pp. 15–36).

Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, *4*(2-3), 111–129.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, *9*(3), 379–395.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: model and applications. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 609–618).

Bordes, A., Boureau, Y., & Weston, J. (2017). Learning end-to-end goal-oriented dialog. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*. OpenReview.net. Retrieved from `https://openreview.net/forum?id=S1Bb3D5gg`

Braslavski, P., Savenkov, D., Agichtein, E., & Dubatovka, A. (2017). What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 345–348).

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Bruce, B. C. (1975). Generation as a social action. In *Theoretical issues in natural language processing*.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gasic, M. (2018). Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 5016–5026).

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D.,

... Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 4516–4525). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-1459

Cai, F., & De Rijke, M. (2016). *Query auto completion in information retrieval*. Universiteit van Amsterdam [Host].

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, *19*(2), 25–35.

Chen, Y.-N., Celikyilmaz, A., & Hakkani-Tur, D. (2018). Deep learning for dialogue systems. In *Proceedings of the 27th international conference on computational linguistics: Tutorial abstracts* (pp. 25–31).

Choi, E., Kitzie, V., & Shah, C. (2012). Developing a typology of online q&a models and recommending the right model for each question type. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 815–824).

Coden, A., Gruhl, D., Lewis, N., & Mendes, P. N. (2015). Did you mean a or b? supporting clarification dialog for entity disambiguation. In *Sumpre-hswi@ eswc*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cohen, P. R. (2018). *Back to the future for dialogue research: A position paper.*

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive science*, *3*(3), 177–212.

Cole, C. (2011). A theory of information need for information retrieval that connects information to knowledge. *Journal of the American Society for Information Science and Technology*, *62*(7), 1216–1231.

Crestani, F., & Du, H. (2006). Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and*

Technology, *57*(7), 881–890.

Dalton, J., Ajayi, V., & Main, R. (2018). Vote goat: Conversational movie recommendation. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 1285–1288).

Das, M., De Francisci Morales, G., Gionis, A., & Weber, I. (2013). Learning to question: leveraging user preferences for shopping advice. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 203–211).

De Boni, M., & Manandhar, S. (2003). An analysis of clarification dialogue for question answering. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics* (pp. 48–55).

De Boni, M., & Manandhar, S. (2005). Implementing clarification dialogues in open domain question answering. *Natural Language Engineering*, *11*(4), 343–362.

Dervin, B., & Nilan, M. (1986). Information needs and uses. *Annual review of information science and technology*, *21*, 3–33.

Diaz, F. (2016). Pseudo-query reformulation. In *European conference on information retrieval* (pp. 521–532).

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A. H., … Weston, J. (2016). Evaluating prerequisite qualities for learning end-to-end dialog systems. In Y. Bengio & Y. LeCun (Eds.), *4th international conference on learning representations, ICLR 2016, san juan, puerto rico, may 2-4, 2016, conference track proceedings.*

Du, H., & Crestani, F. (2004). Spoken versus written queries for mobile information access: An experiment on mandarin chinese. In *International conference on natural language processing* (pp. 745–754).

Duan, N., Tang, D., Chen, P., & Zhou, M. (2017). Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 866–874).

Efthimiadis, E. N. (1996). Query expansion. *Annual review of information science and technology (ARIST)*, *31*, 121–87.

El Asri, L., Schulz, H., Sarma, S. K., Zumer, J., Harris, J., Fine, E., … Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings*

of the 18th annual sigdial meeting on discourse and dialogue (pp. 207–219).

Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., & Ziviani, N. (2005). Concept-based interactive query expansion. In *Proceedings of the 14th acm international conference on information and knowledge management* (pp. 696–703).

Ghosh, S. (2019). Exploring result presentation in conversational ir using a wizard-of-oz study. In *European conference on information retrieval* (pp. 327–331).

Grice, P. (1961). The causal theory of perception. In *Proceedings of the aristotelian society* (Vol. 35, pp. 121–153).

Hashemi, H., Zamani, H., & Croft, W. B. (2020). Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (pp. 1131–1140).

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 2019–2028).

Hassan Awadallah, A., Gurrin, C., Sanderson, M., & White, R. W. (2019). Task intelligence workshop@ wsdm 2019. In *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 848–849).

He, Y., & Young, S. (2005). Semantic processing using the hidden vector state model. *Computer speech & language*, *19*(1), 85–106.

Heilman, M., & Smith, N. A. (2010). Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 609–617).

Hemphill, C. T., Godfrey, J. J., & Doddington, G. R. (1990). The atis spoken language systems pilot corpus. In *Speech and natural language: Proceedings of a workshop held at hidden valley, pennsylvania, june 24-27, 1990.*

Hendahewa, C., & Shah, C. (2013). Segmental analysis and evaluation of user focused search process. In *2013 12th international conference on machine learning and applications* (Vol. 1, pp. 291–294).

Huang, C.-K., Chien, L.-F., & Oyang, Y.-J. (2003). Relevant term suggestion in interactive web

search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, *54*(7), 638–649.

Ingwersen, P. (1982). Search procedures in the library—analysed from the cognitive point of view. *Journal of documentation*, *38*(3), 165–191.

Jia, J. (2009). Csiec: A computer assisted english learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, *22*(4), 249–255.

Jiang, D., Leung, K. W.-T., Yang, L., & Ng, W. (2015). Query suggestion with diversification and personalization. *Knowledge-Based Systems*, *89*, 553-568. doi: https://doi.org/10.1016/j.knosys.2015.09.003

Johansson, P. (2004). *Design and development of recommender dialogue systems* (Unpublished doctoral dissertation). Institutionen för datavetenskap.

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating query substitutions. In *Proceedings of the 15th international conference on world wide web* (pp. 387–396).

Jung, H., Oh, C., Hwang, G., Oh, C. Y., Lee, J., & Suh, B. (2019). Tell me more: Understanding user interaction of smart speaker news powered by conversational search. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–6).

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, *2*(1), 26–41.

Kiesel, J., Bahrami, A., Stein, B., Anand, A., & Hagen, M. (2018). Toward voice query clarification. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 1257–1260).

Kuhlthau, C. C. (2004). *Seeking meaning: A process approach to library and information services* (Vol. 2). Libraries Unlimited Westport, CT.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., & Dolan, B. (2016, August). A persona-based neural conversation model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 994–1003). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16

-1094

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837.

Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–9.

Luger, E., & Sellen, A. (2016). " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5286–5297).

Lurcock, P., Vlugter, P., & Knott, A. (2004). A framework for utterance disambiguation in dialogue. In *Proceedings of the australasian language technology workshop 2004* (pp. 101–108).

Marchionini, G. (1997). *Information seeking in electronic environments* (No. 9). Cambridge university press.

Mitra, B. (2015). Exploring session context using distributed representations of queries and reformulations. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 3–12).

Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 1121–1124).

Mosig, J. E., Vlasov, V., & Nichol, A. (2020). Where is the context?–a critique of recent dialogue datasets. *arXiv preprint arXiv:2004.10473*.

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of documentation*, *33*(1), 1–14.

Paisley, W. J. (1968). Information needs and uses. *Annual review of information science and technology*, *3*(1), 1–30.

Pesendorfer, E.-M., & Koeszegi, S. T. (2006). Hot versus cool behavioural styles in electronic negotiations: The impact of communication mode. *Group Decision and Negotiation*, *15*(2), 141–155.

Qiu, Y., & Frei, H.-P. (1993). Concept based query expansion. In *Proceedings of the 16th*

annual international acm sigir conference on research and development in information retrieval (pp. 160–169).

Qu, C., Yang, L., Croft, W. B., Trippas, J. R., Zhang, Y., & Qiu, M. (2018). Analyzing and characterizing user intent in information-seeking conversations. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 989–992).

Quintano, L., & Rodrigues, I. P. (2008). Question/answering clarification dialogues. In *Mexican international conference on artificial intelligence* (pp. 155–164).

Radlinski, F., Balog, K., Byrne, B., & Krishnamoorthi, K. (2019). Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the annual sigdial meeting on discourse and dialogue.*

Radlinski, F., & Craswell, N. (2017). A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 117–126).

Rao, S., & Daumé III, H. (2018). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2737–2746). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-1255

Rao, S., & Daumé III, H. (2019). Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 143–155). Association for Computational Linguistics.

Rha, E. Y., Mitsui, M., Belkin, N. J., & Shah, C. (2016). Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–9.

Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 583–593).

Ruthven, I. (2019). The language of information need: Differentiating conscious and formalized information needs. *Information Processing & Management*, *56*(1), 77–90.

Saracevic, T., Spink, A., & Wu, M.-M. (1997). Users and intermediaries in information retrieval: What are they talking about? In *User modeling* (pp. 43–54).

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information processing & management*, *26*(6), 755–776.

Searle, J. R. (1985). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2018). A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, *9*(1), 1–49.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth aaai conference on artificial intelligence.*

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social q&a. *Library & Information Science Research*, *31*(4), 205–209.

Shenouda, W. (1990). Online bibliographic searching-how end-users modify their search strategies. In *Proceedings of the asis annual meeting* (Vol. 27, pp. 117–128).

Sitter, S., & Stein, A. (1992). Modeling the illocutionary aspects of information-seeking dialogues. *Information Processing & Management*, *28*(2), 165–180.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., ... Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Spina, D., Trippas, J. R., Cavedon, L., & Sanderson, M. (2017). Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology*, *68*(9), 2101–2115.

Spink, A., & Ozmultu, H. C. (2002). Characteristics of question format web queries: An exploratory study. *Information processing & management*, *38*(4), 453–471.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science*, *48*(8), 741–761.

Stein, A., & Maier, E. (1995). Structuring collaborative information-seeking dialogues. *Knowledge-Based Systems*, *8*(2-3), 82–93.

Stoyanchev, S., Liu, A., & Hirschberg, J. (2014). Towards natural clarification questions in dialogue systems. In *Aisb symposium on questions, discourse and dialogue* (Vol. 20).

Sun, Y., & Zhang, Y. (2018). Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 235–244).

Szpektor, I., Gionis, A., & Maarek, Y. (2011). Improving recommendation for long-tail queries via templates. In *Proceedings of the 20th international conference on world wide web* (pp. 47–56).

Taylor, R. S. (1962). The process of asking questions. *American documentation*, *13*(4), 391–396.

Taylor, R. S. (1967). Question-negotiation an information-seeking in libraries. *College and Research Libraries*, *29*(3).

Taylor, R. S. (1982). Value-added processes in the information life cycle. *Journal of the American Society for Information Science*, *33*(5), 341–346.

Thomas, P., McDuff, D., Czerwinski, M., & Craswell, N. (2017). Misc: A data set of information-seeking conversations. In *Sigir 1st international workshop on conversational approaches to information retrieval (cair'17)* (Vol. 5).

Trienes, J., & Balog, K. (2019). Identifying unclear questions in community question answering websites. In *European conference on information retrieval* (pp. 276–289).

Trippas, J. R., Spina, D., Cavedon, L., Joho, H., & Sanderson, M. (2018). Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 32–41).

Trippas, J. R., Spina, D., Cavedon, L., & Sanderson, M. (2017). How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 325–328).

Trippas, J. R., Spina, D., Sanderson, M., & Cavedon, L. (2015). Results presentation methods for a spoken conversational search system. In *Proceedings of the first international workshop on novel web search interfaces and systems* (p. 13–15). New York, NY, USA:

Association for Computing Machinery. doi: 10.1145/2810355.2810356

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In *Sigir'94* (pp. 61–69).

Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. (2017). Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems* (pp. 2187–2193).

Walker, M. A., Passonneau, R., & Boland, J. E. (2001). Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th annual meeting on association for computational linguistics* (pp. 515–522).

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2001). Clustering user queries of a search engine. In *Proceedings of the 10th international conference on world wide web* (pp. 162–168).

White, R. W. (2016). *Interactions with search systems*. Cambridge University Press.

Williams, J. D., Henderson, M., Raux, A., Thomson, B., Black, A., & Ramachandran, D. (2014). The dialog state tracking challenge series. *AI Magazine*, *35*(4), 121–124.

Williams, J. D., & Young, S. (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, *21*(2), 393–422.

Wilson, T. D. (1981). On user studies and information needs. *Journal of documentation*, *37*(1), 3–15.

Wilson, T. D. (1994). Information needs and uses: fifty years of progress. *Fifty years of information progress: a Journal of Documentation review*, 15–51.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of documentation*, *55*(3), 249–270.

Yan, R., Song, Y., & Wu, H. (2016). Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp.

55–64).

Yang, H., Guan, D., & Zhang, S. (2015). The query change model: Modeling session search as a markov decision process. *ACM Transactions on Information Systems (TOIS)*, *33*(4), 1–33.

Yang, L., Zamani, H., Zhang, Y., Guo, J., & Croft, W. B. (2017). Neural matching models for question retrieval and next question prediction in conversation. *NeuIR '17*.

Yerbury, H., & Parker, J. (1998). Novice searchers' use of familiar structures in searching bibliographic information retrieval systems. *Journal of information science*, *24*(4), 207–14.

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, *101*(5), 1160–1179.

Yuan, X., Belkin, N., Jordan, C., & Dumas, C. (2011). Design of a study to evaluate the effectiveness of a spoken language interface to information systems. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–3.

Yuan, X. J., & Sa, N. (2017). User query behaviour in different task types in a spoken language vs. textual interface: A wizard of oz experiment. In *Proceedings of isic, the information behaviour conference, zadar, croatia* (Vol. 22).

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020* (pp. 418–428).

Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., & Craswell, N. (2020). Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 3189–3196).

Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., & Chen, J. (2020). Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *ACL 2020*, 109.

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 177–186).

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., & Zhou, M. (2017). Neural question generation from text: A preliminary study. In *National ccf conference on natural language processing and chinese computing* (pp. 662–671).