EXPLORING INTELLIGENT FUNCTIONALITIES OF SPOKEN CONVERSATIONAL SEARCH SYSTEMS

by

SOUVICK GHOSH

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

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE DISSERTATION

Exploring Intelligent Functionalities of Spoken Conversational Search Systems

by Souvick Ghosh

Dissertation Director: Chirag Shah

Conversational search systems often fail to recognize the information need of the user, especially for exploratory and complex tasks where the question is non-factoid in nature. In any conversational search environment, spoken dialogues by the user communicate the search intent and the information need of the user to the system. In response, the system performs specific, expected search actions. This is a domain-specific natural language understanding problem where the agent must understand the user's utterances and act accordingly. Prior literature in intelligent systems suggests that in a conversational search environment, spoken dialogues communicate the search intent and the information need of the user. The meaning of these spoken utterances can be deciphered by accurately identifying the speech or dialogue acts associated with them. However, only a few studies in the information retrieval community have explored automatic classification of speech acts in conversational search systems, and this creates a research gap. Also, during spoken search, the user rarely has control over the search process as the actions of the system are hidden from the user. This eliminates the possibility of correcting the course of search (from the user's perspectives) and raises concerns about the quality of the search and the reliability of the results presented. Previous research in human-computer interaction suggests that the system should facilitate user-system

communication by explaining its understanding of the user's information problem and the search context (referred to as the system's model of the user). Such explanations could include the system's understanding of the search on an abstract level and the description of the search process undertaken (queries and information sources used) on a functional level. While these interactions could potentially help the user and the agent to understand each other better, it is essential to evaluate if explicit clarifications are necessary and desired by the user.

We have conducted a within-subjects Wizard-of-Oz user study to evaluate user satisfaction and preferences in systems with and without explicit clarifications. However, the results of the Wilcoxon Signed Rank Test showed that the use of explicit system-level clarifications produced no positive effect on the user's search experience. We have also built a simple but effective Multi-channel Deep Speech Classifier (MDSC) to predict speech acts and search actions in an information-seeking dialogue. The results highlight that the best performing model predicts speech acts with 90.2% and 73.2% for CONVEX and SCS datasets, respectively. For search actions, the highest reported accuracy was 63.7% and 63.3% for CONVEX and SCS datasets, respectively. Overall, for speech act prediction, MSDC outperforms all the traditional classification models by a large margin and shows improvements of 54.4% for CONVEX and 18.3% over the nearest baseline for SCS. For search actions, the improvements were 32.3% and 2.2% over the closest machine learning baselines. The results of ablation analysis indicate that the best performance is achieved using all the three channels for speech act prediction and metadata features only when predicting search actions. Individually, metadata features were most important, followed by lexical and syntactic features.

In this dissertation, we provide insights on two intelligent functionalities which are expected of conversational search systems: (i) how to better understand the natural language utterances of the user, in an information-seeking conversation; and (ii) if explicit clarifications or explanations from the system will improve the user-agent interaction during the search session. The observations and recommendations from this study will inform the future design and development of spoken conversational systems.

Preface

Parts of this dissertation are based on work previously published by the author in (Ghosh, 2019a, 2019b, 2019c; Ghosh, Rath, & Shah, 2018)

Acknowledgements

The road to Ph.D. has been a fantastic journey for me, where I have had the pleasure of meeting the stalwarts in my area of work, being mentored by some amazing researchers, and making great friends for life. As I am writing the final piece of my dissertation work, I would like to thank everyone who has supported me for who I am and encouraged me to be a better person and researcher. Without them, this Ph.D. would have never materialized.

First, I would like to thank my advisor, Professor Chirag Shah, who has been a source of constant support and encouragement. I am grateful for his help in navigating every aspect of the Ph.D. life, from publication to academic services, and ultimately the job market. He has been the friend, philosopher, and guide that any doctoral student would be lucky to have. I consider myself blessed to have an amazing dissertation committee comprising Professor Nicholas Belkin, Professor Katherine Ognyanova, and Dr. Vanessa Murdock. All of them have provided me with invaluable advice and guidance that helped me in completing the dissertation presented here. I am especially thankful to Professor Belkin, who has always been a source of inspiration and knowledge. Thank you!

I am eternally thankful to Rutgers University and the School of Communication and Information (fondly referred to as SC&I, and pronounced Sky). I would like to thank Professor Marie Radford, Professor Ross Todd, Professor Nina Wacholder, Professor Sunyoung Kim, Professor Vivek Singh, and Professor Jeffrey Lane, whose inputs helped me grow as a researcher. I am thankful to Assistant Dean Sharon Stoerger, Professor Lilia Pavlovsky, and Undergraduate Program Director Michael Doyle, who believed in my abilities to teach both undergraduate and graduate-level courses. Their feedback helped me grow as an instructor and prepared me for a career in academia. I am

also thankful to Ph.D. Program Director Jennifer Theiss and student counselors Alli Machiaverna and Danielle Lopez, who provided crucial administrative support at all times. I am grateful to the numerous funding support I received from the Rutgers School of Communication and Information, NSF, and IMLS at various stages of my doctoral journey. My work on Searching and Learning – which was the first step in this dissertation work – was supported by the IMLS grant LG-81-16-0025-16.

I would like to thank various members of the InfoSeeking Laboratory for being wonderful colleagues and friends. It is a long list, and I may be missing some: Dr. Matthew Mitsui, Dr. Dongho Choi, Jiqun Liu, Soumik Mandal, Jonathan Pulliza, Manasa Rath, Yiwei Wang, Shawon Sarkar, and Shannon Taber. Thank you all! I have received amazing social support from my colleagues in Doctoral Students Association and my friends Maria and Pete. I cherish the happy hours and intellectual discussions with the friends I met at various conferences and travels.

I have kept some of the most special people for the last. I would like to thank my parents and my brother, Satanu, for being the wind beneath my wings. They have showered me with unconditional love and support over the years and helped me overcome the most difficult times. My heartfelt gratitude goes out to my partner, Alejandra, who has been my best friend and supported me during some of the toughest periods in my Ph.D. life. A special shout goes out to everyone who listened to my complaints patiently and celebrated my achievements with me.

Dedication

This dissertation is dedicated to my family for all their love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Outline

> Good design, when it's done well, becomes invisible. It's only when it's done poorly that we notice it. Think of it like a room's air conditioning. We only notice it when it's too hot, too cold, making too much noise, or the unit is dripping on us. Yet, if the air conditioning is perfect, nobody say anything and we focus, instead, on the task at hand.
>
> Jared Spool

1.1 Introduction

August 21, 2016. As the Emirates flight landed at JFK airport, New York City, I had hundreds of questions running in my mind. I was in a country I have never been before, starting a new phase of life as a Ph.D. student. I was used to driving on the other side of the road as I grew up in a British Commonwealth country. The social, cultural, and political understanding was also vastly different. Navigating the academic program seemed more laborious than the coursework itself. The situation is not unique to me. As a matter of fact, every student feels like it at some point in their life. So does a traveler when visiting unknown countries. Although situations may differ significantly, every human being encounters numerous situations in life where he feels puzzled and lost. He does not know how to navigate those situations and what actions he should take. That feeling of being lost, where our knowledge seems to fail us, and when we

frantically look for answers to our questions form the motivation behind searching.

1.1.1 Motivation behind Searching

Every person, through his interaction with different knowledge resources (people, books, or the Web), develops a view of the world around him. These views can be thought of as certain typifications that help the person to model the world around him and explain the different types of phenomena (Schutz & Luckmann, 1973). However, as we live our daily lives, we encounter certain anomalies – called anomalous states of knowledge by Belkin (1980) - which does not fit with the existing model of the world and leads to a problematic situation. The situation could be resolved by searching for new information, which helps us to create new knowledge structures or refine the existing ones. Dervin (1992), in her sense-making model, views active information seeking as a way to mitigate the gap (or uncertainty) between the desired and observed situations. The knowledge resources could be the folklores or scrolls preserved by elders in a tribal society or the books guarded by the friendly neighborhood librarians. The World Wide Web, which has entirely changed how information is stored, managed, and delivered, comes without the traditional "gatekeepers." This has ushered in a modern era of information and its retrieval, where the delivery of information is as important as the information itself.

Information Retrieval systems consist of three components – the user (or the searcher), the knowledge resource, and an intermediary who acts as a bridge between the information searcher and the knowledge resource. The knowledge resource has traditionally been textual, containing texts which are represented and organized in ways for fast and efficient access. The searcher, once stimulated by a problem, approaches the intermediary who tries to retrieve some content from the knowledge resource, which will help the searcher to satisfy his information needs. In this thesis, searcher, system, and speech, whenever used, refer to the information seeker, the intermediary (or the agent), and the audio channel, respectively.

One of the very first definitions of Information Retrieval using automated systems was proposed by Mooers (1951):

> "Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines are employed to carry out the operation." (p.25, Mooers, 1951)

Traditional IR systems consist of two streams of activities – the system performs several processes like acquisition, representation, and organization of the documents (and other objects). In contrast, the user performs processes like problem formulation, representation, and query construction (as described in Figure 1.1 from Belkin and Croft (1992)). The search interface (often browser-based), which popularly consists of a text box, allows the user to enter the textual query, which is then reduced to a collection of search terms (or keywords). The intermediary (the retrieval system) matches the terms in the queries with those in the documents using different scoring algorithms and presents potentially relevant objects to the user (usually in the form of a ranked list). The user feedback, if any, is incorporated through query reformulations. In an interactive environment, the user and the intermediary perform multiple rounds of query reformulations and retrievals to mediate the information need. The initial problem, the information need, and the knowledge structures of the user are dynamic and keeps changing during an information-seeking episode (Brooks & Belkin, 1983).

1.1.2 Shortcomings of Traditional IR Systems

The major shortcomings of traditional information retrieval systems could be attributed to the following factors (Begany, Sa, & Yuan, 2015):

1. The user may not know the exact nature of the information problem and what he needs to do to solve the problem (Belkin, 1980);
2. It is difficult for the user to find terms that are accurate in describing their information need. Also, they need to make sure that these terms appear in the documents in the knowledge resource (the database);
3. Traditional search interfaces are suitable for shorter queries (which may not be sufficient to describe the information problem)

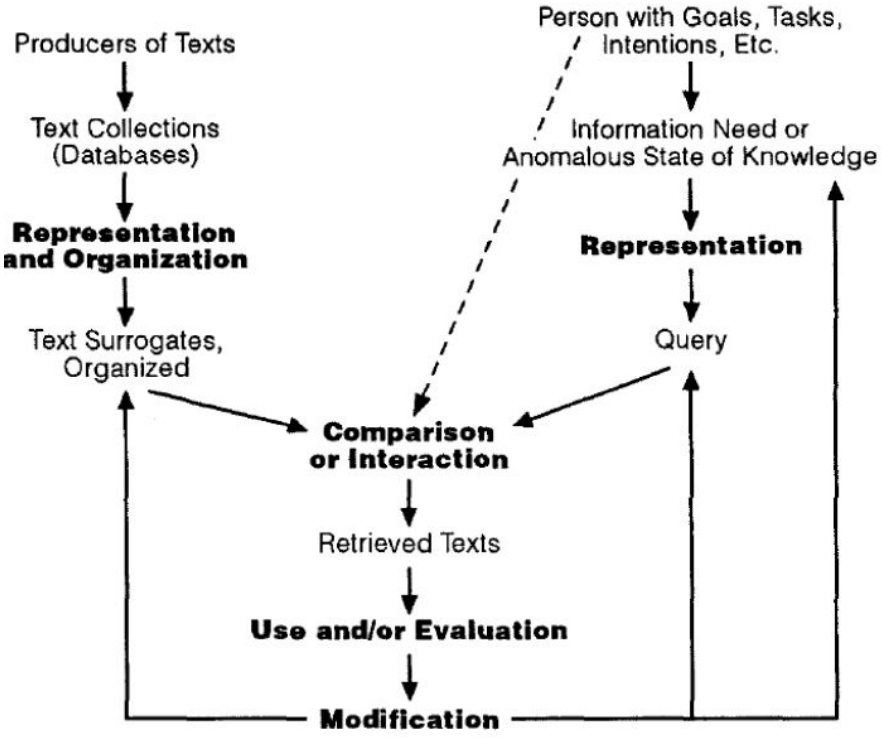


Figure 1.1: Structure of an IR System (Belkin and Croft, 1992).

Research in IR has acknowledged the limitations, as mentioned above, of traditional IR systems. The last few decades have witnessed the massive proliferation and subsequent digitization of information. Searching for information, therefore, has evolved from traditional text-based to more multimodal approaches that promote and support natural interactions. The initial focus – which was to organize and rank the documents – has gradually shifted to make the search systems more accessible, user-centered, and human-like. The current state-of-the-art information retrieval (IR) systems are interactive, provide recommendations, and summarize results in addition to retrieving relevant information.

1.1.3 Conversational Search Systems

As humans, we possess five basic senses: visual (sight), auditory (hear), gustatory (taste), olfactory (smell), and cutaneous (touch). The sense organs detect various forms of signals around us and send the raw data to our brain for processing. These

channels of information help us perceive the world around us. Humans are more reliant on visual stimuli as it allows us to process the largest amount of information with the least cognitive effort. However, from a communication perspective, in-person spoken conversation is the natural and most popular way of exchanging information between two or more humans. We talk to communicate and express ourselves. Therefore, in a situation of information need, where our existing knowledge structures have failed us, it is only natural that we talk to express our problematic situation. Co-presence is the natural mode of talking, which means that apart from verbal communication, para-linguistic and non-verbal communication is also taking place, which invokes senses other than hearing.

Conversational Search is a branch of interactive information retrieval where the searcher enjoys the freedom to explain his information problem to the system (or agent), in natural language. The use of natural language dialogues allows the user to explain better and the system to better understand the knowledge gap of the user. The user-agent interaction is not limited and can go over multiple turns, which should enable the system to build the necessary context, resolve ambiguities through clarifications, and retrieve documents (or information) which are most suited to the needs of the user. The recent popularity of conversational systems can be partially attributed to the limitations as mentioned earlier of the existing retrieval systems. Also, the ubiquity of mobile devices, substantial improvements in automatic speech recognition, the emergence of deep neural networks, and a focus towards more user-centered systems have encouraged the researchers to build dialogue-based retrieval systems. Such systems are significantly more complicated but they are a solution to the existing problems in information retrieval.

Conversational search systems mimic the human-human interactions which occur between a human information seeker and the provider (e.g., librarian-patron conversations). The conversations between the user and system could be in the form of text (as in the case of chatbots) or audio (as in the case of personal assistants). The use of natural language dialogues, over multiple turns, is the reason why these systems are called "conversational." In this research, we focus on voice-based (or spoken) conversational

search systems only.

A recently published statistic [1] highlights that 20% of all searches on Google and 20% of all mobile queries are by voice. Also, 65% of the users who have Amazon Echo or Google Home prefer using voice commands over keyboard inputs. The size of the market for virtual assistants has tripled from 3 billion USD in 2017) to 8.56 billion USD (in 2019) worldwide [2]. The estimated number of users of some type of digital assistants is also projected to reach 1.8 billion by 2021 (Figure 1.2). When asked about technology adoption, 55% of the users felt that they were more comfortable expressing themselves over voice rather than typing their queries. Also, many of them preferred not to touch their smartphones or other devices.

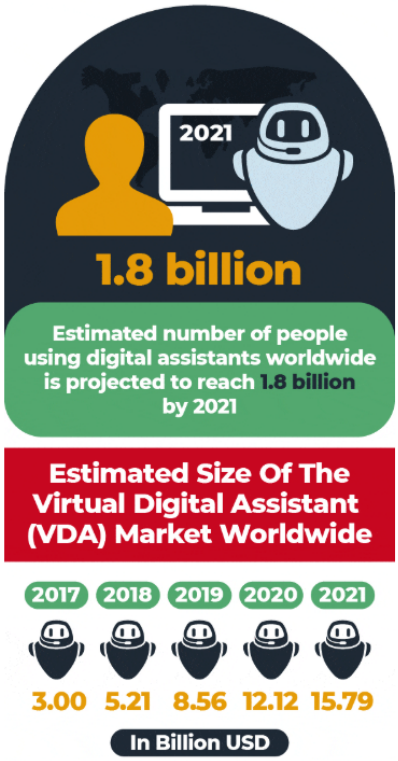


Figure 1.2: Size of VDA Market Worldwide.

Although text-based conversational chatbots are common, we limit our research to an audio-only environment. The advantages and use cases of voice-based systems differ

[1] https://99firms.com/blog/voice-search-statistics/#gref

[2] https://mobiteam.de/en/the-rise-of-virtual-digital-assistant-usage-statistics-and-trends/

from those of chatbots. The limited display capability of mobile devices – phones, wearables, and smart devices – makes it hard to type queries and read the search results, and hence, users prefer talking to the system and listening to the answers (Chang et al., 2002; Najjar, Ockerman, & Thompson, 1998; Trippas, Spina, Sanderson, & Cavedon, 2015b; Turunen, Hakulinen, Rajput, & Nanavati, 2012). Voice commands allow hands-free and eyes-free operation and, therefore, eliminate the need to type. Therefore, the user can multitask, as is observed while driving, cooking, or exercising (Frummet, Elsweiler, & Ludwig, 2019; Ghosh, 2019b; Guy, 2016) when typing is difficult, erroneous, or risky. Voice-based systems are also better suited for people with visual or manual impairment (Guy, 2016; Sahib, Al Thani, Tombros, & Stockman, 2012), dyslexia (Klemmer et al., 2000) or people with limited literacy skills (Trippas et al., 2015b). Although textual interfaces provide autocorrect suggestions, speaking to the system eliminates the need to spell complex, difficult, or foreign words correctly.

1.1.4 Challenges for Conversational Search Systems

Despite all the affordances provided by conversational systems, such systems are still in a nascent stage and need extensive research on multiple aspects. For example, existing state-of-the-art personal assistants perform exceptionally well for simple tasks like setting reminders or alarms, calling or texting a contact, and getting weather or traffic updates. However, as is familiar with any emerging technology, many challenges must be overcome before we can use conversational systems for search purposes.

First, even with existing high-performance computers, it is hard to maintain a long conversation (in real-time) over multiple turns without losing contextual information, something which ongoing research is striving to achieve. As such, the current state-of-the-art conversational retrieval systems are better suited for short conversations, simple queries, factoid questions, and non-exploratory searches. Second, conversational systems are often spoken, and the transient and linear nature of speech severely limits the functionalities of such interfaces. Thus, information is required to be transmitted in smaller chunks (short audios or limited results) (Lai & Yankelovich, 2002; Trippas,

Spina, Sanderson, & Cavedon, 2015a) to prevent overloading the users' short-term memory (Turunen et al., 2012). So, spoken conversational systems cannot present complex structures like images, graphs, videos, and the search engine result pages and associated hyperlink support (Trippas, Spina, Cavedon, Joho, & Sanderson, 2018). Reading the search engine results page (SERP) reduces user's satisfaction and often leads to search failures. Spoken conversational systems do not allow scanning of results, query modifications (Sa & Yuan, 2019), or looking at prior results. Lastly, such systems suffer in noisy environments, such as outdoors, where the environmental noise may superpose with the voice searches and responses, thus causing recognition and comprehension errors (Turunen et al., 2012). Non-native speakers of English, often with accents, also find it challenging to communicate with conversational systems as recognition errors increase manifold for words that are difficult to pronounce.

If we view searching as a learning process, then the spoken conversation may not be suitable for all types of learners or users. As the VARK model of education suggests that there are four primary types of learners: visual, auditory, reading/writing, and kinesthetic (Fleming, 2001). So, for educational and learning purposes, the preference of the search system will depend on the user's subjective preference for the learning medium. An ideal solution will be to use a multimodal interface that will support both textual queries and human-computer dialogue and use displays for presenting visual objects and the results of complex queries.

While developments in automatic speech recognition and generation have made it possible for artificial systems to listen and speak, a human-human conversation is more than that. It involves the correction of wrong utterances (on both sides), understanding of conversational implicatures (Grice, 1989) and non-verbal cues, adhering to conversational principles (Tannen et al., 2005), staying on-topic, and responding in a fashion that suits the style and sensibilities of the listener. As most users are not good at explaining their problems in few and exact words (Landis & Koch, 1977) and the existing state-of-the-art systems are still in a developmental stage, most of the modern-day conversational assistants fail to meet the lofty standards expected of them Luger and Sellen (2016). Therefore, current voice-based personal assistants are widely regarded

as task-based systems (Luger & Sellen, 2016) incapable of performing complex searches and engaging in multi-turn dialogues.

1.2 Research Problem

Most of the popular information retrieval systems and search engines are non-conversational. The popularity of handheld and mobile devices has led to the ubiquity of conversational search systems, where the one-shot query-response system is replaced by an interactive dialogue-based system. Spoken conversational search systems allow the users to talk to the search agent instead of typing the queries. The result could be presented back as audio, or a combination of audio and text, therefore, mitigating the limited visual display in mobile and handheld devices. The development of conversational systems, spoken or otherwise, has tremendous implications for user satisfaction and search experience. Not only is conversation the natural mode of communication for humans, but it also allows the user to interact with the agent using natural language. The existing query formulation and modification techniques, result presentation strategies, and evaluation metrics are catered towards traditional and visual search processes. Research communities have tackled different challenges in conversational systems from both user- and system-perspectives, exploring the different facets of such systems, which include but are not limited to context-building in conversations, query creations and modifications, intent recognition, and results presentation. However, it has also raised questions about the nature of interactions that occur between the conversational search system and the user: more specifically, what are the intelligent functionalities expected of such systems as they converse with the user.

There is a widely accepted research gap in determining the types and extent of interaction desirable in conversational search systems. While conversations help the intermediary understand the long- and short-term contexts of the user, there should be an option to backtrack if the conversation is following a wrong search direction. While using most of the existing systems, the user rarely has control over the search process as the actions of the system are hidden from the user. This eliminates the possibility of correcting the course of search (from the user's perspectives) and also creates

concern about the quality of the search and the reliability of the results presented. Previous research in human-computer interaction suggests that the system should facilitate user-system communication by explaining its understanding of the user's information problem and the search context (which is often referred to as the system's model of the user). Such explanations could include the system's understanding of the search on an abstract level and the description of the search process undertaken (queries and information sources used) on a functional level. While these interactions could potentially help the user and the agent to understand each other better, it is essential to evaluate if explicit clarifications are necessary and desired by the user.

Also, the current state-of-the-art conversational search systems often fail to recognize the information need of the user, especially for exploratory and complex tasks where the question is non-factoid in nature. In any conversational search environment, it is of utmost importance that the agent understands the utterances by the user and performs the appropriate search activity. This is a domain-specific natural language understanding problem where the user's utterances guide the agent's action. Spoken dialogues communicate the search intent and the information need of the user (searcher) to the agent (intermediary). In response, the agent performs specific, expected search actions. Prior literature in intelligent systems suggests that any conversation can be represented as a state-transition diagram where the edges represent the speech or dialogue acts, and the nodes represent the conversational states. Speech acts have been studied extensively in philosophy, speech, and dialog communities. They convey the meaning of the utterances on a functional level and could be used to understand what the user wants. Only a few studies in the information retrieval community have explored automatic classification of speech acts in conversational search systems, and this creates a research gap. Therefore, it is essential to develop insights on: (i) if explicit clarifications or explanations from the system will improve the user-agent interaction during the search session; and (ii) how to better understand the natural language utterances of the user, in an information-seeking conversation.

Throughout the rest of the thesis, we use conversational search systems (CSS) to define systems that provide a more human-like interaction (Arguello, Choi, & Capra,

2017) to the user who has the freedom to speak to the system (voice requests) instead of typing. We also take the liberty of assuming that the user is also the searcher while the system is the search intermediary or personal assistant.

1.3 Layout of the Thesis

This thesis is divided into three parts and a total of nine chapters.

- **Part-I – Introduction and Thesis Overview** :

  This part comprises chapters 1, 2, and 3, where we explain traditional and conversational search, the different challenges, the motivations behind this research, the research questions, and related works.

  - **Chapter 1 – Introduction and Outline:**

    We highlight the motivation behind searching, explain the challenges for conversational search systems, and describe the research problems tackled in this thesis. We also provide the layout of the thesis at the end of this chapter.

  - **Chapter 2 – Related Work:**

    In this chapter, we discuss the previous research done in the domain of conversational search systems. We point out the implications of the prior work and how they motivate the design of this thesis.

  - **Chapter 3 – Research Questions:**

    We present the research questions which guide the overall direction of this research.

- **Part-II – Methodology** :

  This part comprises chapters 4, 5, and 6. We discuss the research methodology, which includes the user study design, data collection and processing, and data analysis.

  - **Chapter 4 – Experimental User Study Design:**

    We discuss the experimental set up used, the design of the search tasks, and

the set of protocols that were developed for the study. We also specify the data collection procedure.

- **Chapter 5 – Transcription and Thematic Analysis:**

  We describe the details of the qualitative coding process and the themes developed for annotation. We provide descriptive statistics for the two datasets used in this study.

- **Chapter 6 – Development of the Deep Neural Classifier :**

  We report the details of the deep neural model used for the prediction of speech acts and the search actions. This includes the model architecture and the hyperparameters used for prediction.

- **Part-III – Results and Discussion** :

  This part consists of chapters 7, 8, and 9. We present the results and discuss our findings concerning the research questions. We suggest design recommendations and outline possible directions for future research in conversational search systems.

  - **Chapter 7 – Exploring the Role of Clarifications in User-Agent Information-seeking Dialogues:**

    In this chapter, we report the findings from our user study. We analyze the user feedback statistically and assess the effect of clarification on the user-agent conversation. We also use our user study to make design recommendations.

  - **Chapter 8 – Towards Natural Language Understanding of Spoken Conversational Search Systems:**

    In this chapter, we answer our second research question on natural language understanding of voice-based conversational search systems. We report the different performance metrics of the deep neural classifier and perform ablation analysis to show the impact of different types of data on the prediction performance.

  - **Chapter 9 – Conclusion and Future Work :**

    We conclude the thesis by providing a summary of our research, the practical

implications of our findings, and recommendations for future research.

The thesis has four appendices. Appendix A contains details of the different documentations prior to the study. This includes the Institutional Review Board Approval, Recruitment Letter, Consent Form, and Information Sheet for the users and the Wizard. Appendix B presents the questionnaires: pre-test, pre- and post-task, and exit interview. Appendices C and D show the post-study documentations and additional statistical charts, respectively.

1.4 Chapter Summary

In this chapter, we introduce the readers to the topic of this thesis, which is, Conversational Search Systems. We highlight the key concepts in information retrieval and how the motivations behind searching can be explained from the perspective of information science. The typifications of the world around us, often influenced by different social and cultural factors, explains how human beings create, modify, and restructure knowledge. Our lack of knowledge motivates us to search and make sense of uncertain concepts and situations. Searching is, therefore, an act of learning and sensemaking. Next, we follow the evolution of search from libraries to traditional browser-based systems to conversational systems. While traditional search systems have aspired to provide interactivity and search functionalities similar to those of a human intermediary, conversational systems provide the much needed natural language support and voice-based input mechanism. We briefly discuss the advantages and the challenges for conversational systems and outline the major problems that this dissertation addresses. We end the chapter by providing the layout of the chapters.

Chapter 2

Background and Related Works

Searching for information within, or using systems, has long evolved from libraries to web spaces and has kept evolving to more convenient and user-centered approaches. User profiling, coupled with intent recognition, has led to research in information fostering (Shah, 2018) and proactive search assistance. However, as mentioned in the previous chapters, searching is a problematic activity when the user has to define his information problem using a few keywords. To this end, conversational search approaches provide the much-needed freedom to the user who can explain his search requirements and resolve ambiguities, which may occur to the search system. The user can talk to the system to present his query, in natural language, and explain the information problem through multiple rounds of back and forth dialogues.

In this chapter, we discuss literature relevant to various aspects of conversational search systems (CSS). We discuss the different theoretical frameworks in information retrieval and the ones which apply to conversational search systems. We define spoken conversational search, the different categories of conversational search systems, and their properties. We also discuss previous research papers that highlight the major challenges for conversational systems and propose potential solutions. Finally, we provide an overview of the methodological approaches adopted by researchers in this domain, ranging from human-centered experimental design to algorithmic development. Our goal is to show the interdisciplinary nature of this area of research so that wheels are not necessarily reinvented on all sides.

2.1 Conversational Search Systems

In this section, we define what Conversational Search Systems are, the types, and properties of such systems.

2.1.1 Definition

If we investigate the history of IR, we can find evidence of conversational systems defined as dialogue-based, spoken, or discourse-oriented. H. C. Bunt (1989) is credited for coining the word *information dialogue* for the type of dialogues observed in simple information systems that provided factual information. The term dialogue, when used in the context of IR systems, refers to the multiple rounds of negotiation or clarification which occurs between the dialogue partners (user and intermediary). Such interactions aim at developing a constructive solution to the initially vague information problem and hypothesizing the information need of the user (Stein, Gulla, & Thiel, 1999).

Although conversational search systems are often defined as artificial systems that can interact, understand, and respond in natural language (Laranjo et al., 2018; Ram et al., 2018), it overlooks the complexity of a human-human conversation. Conversations are interactive and incremental, explanatory and educational for both parties involved, involves multiple rounds of turn-taking, and are expeditious (Joho, Cavedon, Arguello, Shokouhi, & Radlinski, 2018). As such, it could be useful to use prior research works in linguistics and apply them in the context of information seeking dialogues.

Radlinski and Craswell (2017) provided a formal definition of conversational search systems:

> "*A conversational search system is 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.*" (Radlinski and Craswell, 2017, p. 120).

2.1.2 Properties of Conversational Search Systems

In recent years, the term "conversation" has gained popularity in the context of information retrieval systems. The Oxford English Dictionary[1] defines a conversation as "a talk, especially an informal one, between two or more people, in which news and ideas are exchanged" while Wikipedia defines conversation[2] as "an interactive communication between two or more people". If we look closely at both the definitions from the perspective of IR, we can conclude that a conversation must be interactive, should involve two or more participants, and should lead to an exchange of knowledge or information. The term conversational defines the interaction mechanism between the user and the intermediary and highlights the human-like properties aspired of future search systems (Arguello et al., 2017).

Table 2.1: Properties of Conversational Search Systems (Radlinski & Craswell, 2017)

Property	Description
User Revealment	The system assists the user in discovering and expressing the information need using natural language. The user can also define his short- and long-term preferences.
System Revealment	The system reveals itself to the user. This could include the mode of operations, the capabilities and the limitations, and the functioning. This helps in managing the user expectations on what the agent can and cannot do.
Mixed-Initiative	The search is not necessarily controlled by the system or the user. Both the system and the user can take the initiative as appropriate.
Memory	The system should maintain a context of previous utterances by the user. This is similar to short-term memory and would allow the system to understand the references to past statements, contradictions, and backtracking.
Set Retrieval	The system should be able to explain the utility of the sets of complementary items.

To develop an effective conversational system, we need to allow for natural interaction (in the form of dialogues) between the user and the system. Such interactions, although sophisticated and unconstrained, would allow for an accurate understanding

[1] https://en.oxforddictionaries.com/definition/conversation

[2] https://en.wikipedia.org/wiki/Conversation

of the user's information problem and the knowledge gap. To generate appropriate responses, we must understand the intention and information needs of the user. In recent work, the desirable properties of such systems have been highlighted by (Radlinski & Craswell, 2017) (Table 2.1): the user and the system should negotiate the user's information need in conjunction with the user's long-term goals and the limitations of the system while keeping track of context and old utterances.

These properties can be related to the work of Belkin (1984), who suggested that both the user and the intermediary need to develop cognitive models of each other in an effective information transfer environment. To have successful communication, the participating individuals need to collaborate, create models of the other, and negotiate it to perfection (Hollnagel, 1979). The functions of an intelligent search interface, illustrated by Belkin, Seeger, and Wersig (1983) and Belkin, Brooks, and Daniels (1987), is presented in Figure 2.2.

Table 2.2: Functions of an intelligent agent (Belkin, 1987).

Name of the Function	Description
Determine Problem State	The agent must determine the position of the user in problem treatment process, e.g., if the user is yet to formulate the problem, if the problem has been well specified, and so on.
Determine Problem Mode	The appropriate mechanism capability should be determined. For example, if it involves reference retrieval.
Determine User Model	Generate description of the user. This could include the type, goals, beliefs, etc. of the user. e.g., if the user is a graduate student, if the search is to write a paper.
Describe Problem	Generate description of the type, topic, structure of the problem along with contextual information
Determine Dialogue Mode	Decide the appropriate dialogue type for situation, e.g., natural language, menu.
Build Relevant World	Choose and apply appropriate retrieval strategies to knowledge resource, e.g., best match, gap filling.
Generate Response	Determine propositional structure of the response which should be given to the user as appropriate to the situation.
Analyze Input	Convert input from user into structures usable by functional experts, e.g. query to SQL statements
Generate Output	Convert propositional response from the agent to a form which is appropriate to the user and his situation. For example, a picture may be converted to a description if no displays are available.
Explain	Describe mechanism of operation, capabilities, etc. to the user as appropriate. This could also include other information like query terms, information sources, length of utterance allowed, and so on. (Similar to System Revealment)
Secondary Communication	This is essential to build and sustain the communication. The agent reassures the user that it understands the user (or not). This helps in building the context, and to create models of each other.

Apart from the properties already stated above, a conversational agent should respond in real-time, the response should be incremental and engaging and should have a moral character. For example, the agent should not provide wrong information on purpose or be biased (e.g., to promote the sale of a particular product). Additionally, it should be transparent about the sources of information used, the query used (if applicable), and why some information was preferred over the others (for the sake of brevity, prior choices, and so on).

2.1.3 Types of Conversational Search Systems

Although our work is focused on voice-based conversational search systems, there could be various categories of conversational systems based on their media of communication, application areas, ways of functioning, and modes of development.

Existing conversational systems can be divided into two categories: text-based systems (chatbots) and voice-based systems (personal assistants). While there are some hybrid systems that allow the user to switch between the two modalities, the use cases suggest that users prefer either typing or speaking but not both at the same time. We would like to highlight that the results may be presented over text, audio, or on-screen, depending on the availability and affordances of the devices. While voice-based conversational systems allow the user to use speech as input, text-based conversational systems require the user to type the dialogues in natural language. Therefore, chatbots cannot be used for multitasking and handsfree operations. Also, while speech supports terms that are easy to pronounce, text modality supports searching by using the terms easy to type. However, text-based systems allow for more extended dialogues and can return lists of results, including images and videos. Unlike the transient nature of speech, the text output allows the user to scan the results, and process them slowly, and refer to at a later point in the conversation. This leads to less cognitive load for the user. Lastly, it also provides better support for exploratory and collaborative tasks in conversational environments, as the system can handle multiple users, their intents, and contextual information.

The use of personal assistants or chatbots can reduce the workload on the user and

increase productivity. Chatbots have been employed for providing search assistance in open- and closed- domains (Brandtzaeg & Følstad, 2017; Mallios & Bourbakis, 2016), for example, vacation planning (Shiga, Joho, Blanco, Trippas, & Sanderson, 2017), tour guidance (Kopp, Gesellensetter, Krämer, & Wachsmuth, 2005), flight booking services (Dubiel, Halvey, Azzopardi, & Daronnat, 2018), or as conversational partners (Radziwill & Benton, 2017). While chatbots provide conversational capabilities in webpages (popular in healthcare and banking industries), certain situations warrant a handsfree- and eyes-free operation. Persona-based conversational models (Li et al., 2016; Y. Zhang, Chen, Ai, Yang, & Croft, 2018) are also popular as they capture the styles and backgrounds of individual users. A separate line of research explored the use of embodied conversational agents (Bickmore & Cassell, 2005; Cassell, Sullivan, Churchill, & Prevost, 2000).

While task-oriented interactions help the user in accomplishing specific tasks (Wen et al., 2016; Williams & Zweig, 2016), non-task-oriented dialogues can be chitchat or informational (W. Wang, Huang, Xu, Shen, & Nie, 2018). Some examples of task-oriented systems would be digital personal assistants who can set alarms, play music, and place an online order at the request of the user. Depending on how they function, dialogue systems can be categorized as rule-based (if they use templates to generate responses), retrieval-based (if they seek the response in the collection) (Ji, Lu, & Li, 2014; Wu, Wu, Xing, Zhou, & Li, 2016; Yan, Song, & Wu, 2016; Yan, Zhao, & E, 2017) or generation-based (response generation using sequence to sequence models) (Bordes, Boureau, & Weston, 2016; Dhingra et al., 2016; Li et al., 2016; Tian et al., 2017; L. Yang et al., 2018). Generation-based models are more flexible than rule-based models, which are constrained by the heuristic templates, or retrieval-based models, which are dependent on the archive (i.e., similar responses must be present in the collection). Some hybrid models(Qiu et al., 2017) combine both the strategies.

Conversation is not only crucial for IR systems but also for building question-answering and recommendation systems. If an automatic sales agent could speak to the user, and offer relevant recommendations, it could improve the user experience and the

sales volume for the online retailers. Kenter and de Rijke (2017) considered conversation as a machine reading task and developed a standalone question-answering system to answer "common-sense" context-sensitive questions. Some of the other works (Fukui et al., 2016; Gupta, 2017; Zhou, Tian, Sukhbaatar, Szlam, & Fergus, 2015) have also focused on on multimodal aspects, like answering questions from an image using natural language (e.g., How many cars can be seen in the picture?). Visual question answering research works use insights from both natural language processing and computer vision. Sun and Zhang (2018) attempted to integrate conversational and recommender systems to build a personalized agent that focused on the long-term preference of the buyer using a deep reinforcement learning framework. The authors claimed that such a system is better suited to capture the intent and preferences of the user. Christakopoulou, Radlinski, and Hofmann (2016) addressed the problem of cold start in recommendation by using human-like dialogues to understand the users' preferences.

2.1.4 Challenges and Limitations

In a traditional search system (which is mostly browser-based), the user enters a query in a text box. This query serves as an expression of the user's information need. The system (who acts as the intermediary) uses the query to retrieve relevant documents that are expected to solve the user's problematic situation. The single-shot keyword-based approach is not an ideal way to elicit the information need of the user. Instead, by allowing the user to talk, when he expresses the search situation in natural language, the system can have a better understanding of the user, his information need, the search context, and the user preferences.

Table 2.3: Example of User-Agent Conversation

Speaker	User-Agent Information-seeking Dialogues
User:	Hey, Joanna,
Agent:	Hi I am Joanna. How may I help you today?
User:	What Artificial intelligence conferences are in early. 2020
Agent:	Okay. Please Let me look into it. Give me a few minutes
Agent:	The top of artificial intelligence conferences in early 2020 are 1) Conference on Computer Vision and Pattern Recognition, 2) International World Wide Web conferences and 3) Association for the Advancement of Artificial Intelligence.
User:	Where are they located?
Agent:	Okay, please let me look into it. Give me a few minutes.
Agent:	The top artificial intelligence conferences in early 2020 are: 1) Conference on computer vision and pattern recognition held in Seattle, USA, 2) International World Wide Web Conference is held in Taipei, Taiwan, and 3) Association for the Advancement of Artificial Intelligence in New York, New York, U.S.A.
User:	Bye Joanna

The interactions between the user and the agent are incremental and occur over multiple turns, thus allowing both the user and the system to develop accurate models of each other and build context to the search situation. If we look at the user-agent interactions in the example of Table 2.3, there are many challenges that must be overcome.

The departure from a text-based interface to spoken dialogue queries involves many challenges, in terms of characteristics, implementation, and evaluation. First, the agent needs to recognize the user utterances correctly. The agent may face problems understanding the accent of non-native speakers. Also, during the early stages of the search, the users are unsure of what they are looking for, and hence, the problem statement may be vague and too complicated. As the users are no longer constrained by the search box and keywords (which is a limitation of traditional text-based systems), they can express their information problem in a long and verbose manner. Also, the users may state multiple problems in a single turn, make multiple moves, and provide feedback (Trippas et al., 2018). Users also tend to ask random and unanswerable questions, often without context, which makes it extremely hard for an artificially intelligent agent to respond satisfactorily.

Therefore, the system will need to assign importance to specific sections of the conversation and determine the appropriate system response, like clarifying, asking a follow-up question, showing (or speaking from) some documents, or displaying the class hierarchy (in multimodal systems) (Z. Liu, Niu Z.and Nie, Wu, & Wang, 2017). Unlike traditional text-based systems (which present the result on display), conversational search systems operate in personal assistants and mobile devices, which come with small or no display. As speech is transient, linear, and temporal (Kotti, Papangelis, & Stylianou, 2017), users do not have the option to revisit the results or scan over them.

The linear nature of speech increases the complexity of the system manifold from an implementation perspective. For a scenario involving multiple goals, the system should be optimized using multiple rewards functions and must be able to balance the trade-offs between the length of response, diversity, and personalization (Stein & Maier, 1995). The response time between conversational turns should be kept minimal to enhance naturalness (Kenter & de Rijke, 2017); Otherwise, the user may feel frustrated and quit the search session. Also, for questions that are similar semantically, the system must not provide different answers. Lastly, the system should allow users to backtrack in case of wrongly formulated or misrecognized (errors of ASR) queries. The backtracking would require the system to model negation and query states, which would add more steps to the search process (Begany et al., 2015).

For a traditional browser-based system, the response generated by an IR system is usually in the form of the search engine results page (SERP) containing a ranked list of documents. While such a system expects the user to click on any of the hyperlinks on SERP, a conversational system is often constrained by the inadequacy of display support. Therefore, it is ineffective for such systems to display the list of results (on a small screen) from the search engine results page (SERP). Alternatively, reading out the entire list over audio is insufficient and cognitively demanding for the user (Thomas, McDuff, Czerwinski, & Craswell, 2017; Trippas et al., 2018). Instead, in similar situations, the system should perform answer aggregation, assimilation, or summarization of the SERP results.

While the audio-channel is good for factoid questions (e.g., Who is the President

of the USA?), which can be answered in a single sentence, more complex queries (e.g., What can you tell me about the life of current US President?) are answered with a SERP. However, if the user intention and the information problem are simple enough, the user may expect the final answer, obtained directly from within a document, without the SERP. To answer such a question directly, the intermediary (system) will need to create a model of the user and his intent through conversation. It could also use historical search patterns to model similar users, their intents, and queries. By modeling the user question (voice query), the user profile, and the query, similar documents can be clustered to answer the question. Such clusters would allow the user to have greater coverage of information space (Trippas et al., 2015b), better understand his information need, and assess relationships between the search results (Pu, 2010). The system response, presented in the form of audio-summaries, will allow the user to select between the results of the SERP page (in case of a SERP summary) in subsequent turns or access the information directly (in case of document summary). However, presenting the information not only requires the integration of the search engine with the documents (Trippas et al., 2018), but also the information must be transmitted in smaller chunks (short audios, limited results) (Lai & Yankelovich, 2002; Trippas et al., 2015a) to prevent overloading the users' short-term memory (Turunen et al., 2012).

For evaluating the conversational search systems, it is essential to develop newer evaluation paradigms and metrics. Such evaluations will be based on quantitative and qualitative factors and are likely to be user-dependent and costly (Stein & Maier, 1995). The evaluation will also require preparation of gold standard datasets and assessment of the system performance not only from an information retrieval perspective, but also from the perspectives of artificial intelligence (recognition of voice, modeling of intent, intelligence), communication (clarity, adequacy, and better dialogue management), and user-centered designing (user comfort, convenience, and aesthetics).

Overall, a conversational system should be more robust, precise, effective, meaningful, engaging, and interactive (Jadeja & Varia, 2017) than a text-based system. The interactions should not only be effective functionally but be more human-like in

terms of style (Thomas, Czerwinski, McDuff, Craswell, & Mark, 2018), diversity, emotions, ethics, and morality (Z. Liu et al., 2017; Radlinski & Craswell, 2017). If the conversation involves multiple users (if the device or system is used by multiple users simultaneously), the conversational system should adapt accordingly (to the speaker, his language, background knowledge, and characteristics). The system should be able to derive the relationship between different users and respond to the users individually or in a group. The system should also model personality and keep track of different users, their attributes, and states. In an open-domain conversation, the conversational search agent must maintain coherence between successive turns and take the initiative like a human intermediary (Bowden, Oraby, Wu, Misra, & Walker, 2017). As many of these characteristics are desirable but too complicated to execute, the conversational agent should gracefully reveal the services offered and the limitations to the user (Radlinski & Craswell, 2017).

2.2 Conceptual Frameworks

In this section, we discuss different conceptual frameworks for informational retrieval and specifically conversational systems.

2.2.1 Theoretical Models in IR

Many researchers have proposed different theoretical frameworks for research and development in user interfaces that support information seeking (Marchionini & White, 2007). The popular models of information search (Belkin, 1980; Ellis, 1989; Marchionini, 1997; Saracevic, 1997; Wilson, 1999) explored how the users performed a search in a specific environment, either online or in-person (Hearst, 2009). Searching for information is motivated by an information problem or need. Belkin (1980) explained the information need from a cognitive viewpoint while Taylor (1962) conceptualized information need as a fluid process with four stages: visceral, conscious, formalized, and compromised. Information seeking has been envisioned as an evolving process (Bates, 1989) with multiple stages or steps (Ellis & Haugan, 1997; Marchionini & White, 2007;

Saracevic, 1997) and involves both the user and the intermediary.

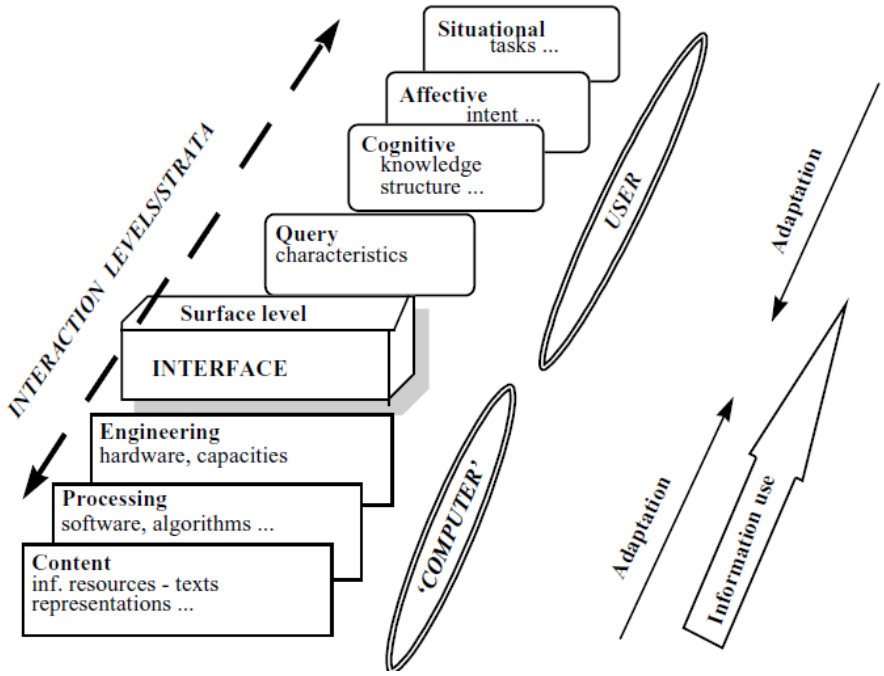


Figure 2.1: Stratified Model (Saracevic et al., 1997).

Marchionini and White (2007) divided the information-seeking process into various subprocesses – recognize the information need, accept the information need, formulate the problem question, expression of the problem in a way understood by a search system, examination of the results returned by the search system, and reformulation of the problem and its expression. Ideally, the search engine (or intermediary) should assist the user in different stages of this process from translating the information need to query to evaluating the results (Hearst, 2009). Saracevic (1997) proposed the stratified model (Figure 2.1) from both the user- and the system- perspectives. According to this model, there are several connected levels – content, processing, and engineering levels on the system side and query, cognitive, affective and situational levels on the user side – and the user-computer interaction occurs at the interface level. The strata are not independent of each other and the weakest point of interaction between user and computer can compromise from achieving the best possible outcome of the search. Ellis and Haugan (1997) divided the process of information-seeking into eight

functional categories – surveying, chaining, monitoring, browsing, distinguishing, filtering, extracting, and ending. Bates (1989) suggested that the information query is an ever-evolving and modifying process that is not satisfied by the final retrieved set but by bits of information at each stage of the search process (Figure 2.2). Several other studies looked at information seeking from cognitive perspectives (Ingwersen, 1996) and as part of everyday life (Savolainen, 1995).

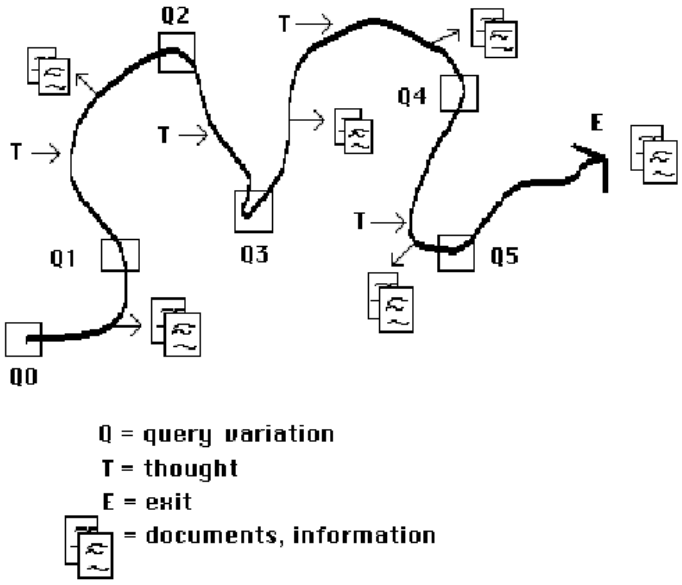


Figure 2.2: Berrypicking Model (Bates, 1989).

2.2.2 Frameworks for Conversational Search

Some of the early research in conversational search explored how dialogues can be incorporated in retrieval systems (Oddy, 1977; Sitter & Stein, 1992; Stein & Maier, 1995). According to Winograd, Flores, and Flores (1986), in an information-seeking episode, the searcher-intermediary dialogue is akin to a conversation and can be represented using a state transition network. In other words, any information-seeking dialogue can be modeled based on the different speech or dialogue acts. This formed the basis of the Conversation for Action (CfA) model – a theoretical model which could simulate any conversation – where each node represents one of the dialogue states, and the arcs are speech acts which help to transition from one dialogue state in the conversation to

the other. The CfA model combines the philosophy of language with interpretations in context. The authors argue that the meanings of utterances are socially constructed, and the behavioral expectations of the participants control the flow of conversation. Later, the CfA model was modified and extended by Sitter and Stein (1996) through their analysis of some generic dialogue scenarios. The authors proposed the Conversational Roles (COR) model, which categorizes the dialogue acts based on the intention of the participants and aims to guide the users through different stages of IR by formulating a dialogue plan. An alternate approach has been adopted by Belkin, Cool, Stein, and Thiel (1995) who conceptualize Information Retrieval as interactive information seeking and use case-based reasoning (CBR) (Riesbeck & Schank, 1991; Schank, Kass, & Riesbeck, 2014) to model the human-computer interaction in an IR system. To understand the information problem and to generate effective interactions, the authors propose various information seeking strategies (ISS) and example scripts to model the pattern of human-computer interaction. The interaction patterns differ significantly for different information-seeking strategies and could be modeled using the COR model. In a separate work, Daniels et al. (1985) performed discourse analysis of human-human information interactions and proposed a problem structure that could be used to guide human-computer information-seeking dialogue. The authors presented the goal hierarchy (Table 2.4) and how the goals are instantiated as the foci of the dialogue.

Table 2.4: Goal Hierarchy for Document Retrieval Problem (Daniels et al., 1985)

Level	Goal
1	User leaves the system
2	User is satisfied
3	Appropriate response to user
4	Appropriate search formulation
5	Subgoals to achieve Level 4 goals
6	Subgoals to achieve Level 5 goals

The proposed frameworks in the early 1990s preceded the recent growth in conversational search systems. Some of the latest works in this domain (Azzopardi et al., 2018; Trippas, Spina, Cavedon, & Sanderson, 2017a; Vakulenko, Markov, & de Rijke, 2017) attempted to develop frameworks capable of explaining the information-seeking

dialogues and the associated cognitive functions. Trippas et al. (2017a) suggested a turn-based framework for spoken environment (Figure 2.3), Azzopardi et al. (2018) provided an extensive list of actions and interactions in conversational search (Figure 2.4) while Vakulenko et al. (2017) proposed the QRFA model to show the conversation flow in information seeking episodes (Figure 2.5). These frameworks create the platform for automatically identifying the different dialogue patterns and the corresponding search actions witnessed during human-human information-seeking dialogues.

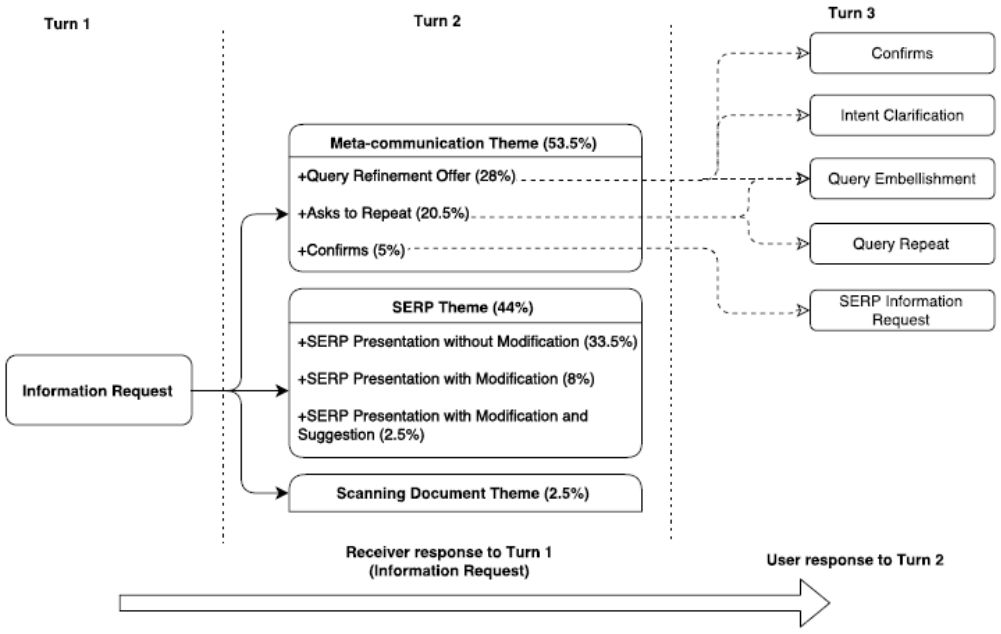


Figure 2.3: Interaction Theme Map (Trippas et al., 2017b)

In recent research, attempts have been made to develop newer frameworks to explain conversational information-seeking dialogues or to bridge the terminology gap between different conversational datasets and the associated cognitive actions (Azzopardi et al., 2018; Trippas et al., 2017a; Vakulenko et al., 2019). To this end, Trippas et al. (2017a) suggested a turn-based framework for the spoken environment, Azzopardi et al. (2018) provided an extensive list of actions and interactions in conversational search while Vakulenko et al. (2019) proposed the QRFA model to show the conversation flow in information seeking episodes.

		User	Agent		
	Query Formulation [11]	**Reveal** Disclose Non-Disclose Revise Refine Expand	**Inquire** Extract Elicit Clarify	User Revealment [10]	Memory [10]
Set Retrieval [10]	Result Exploration [11]	**Inquire** List Summarize Compare Subset Similar	**Reveal** List Summarize Compare Subset Similar	System Revealment [10]	
		Navigate Repeat Back More ... Note	**Traverse** Repeat Back More ... Record		
Mixed Initiative [10]		**Interrupt**	**Suggest**		
		Interrupt	 Recommend Hypothesize		
		Interrogate Understand Explain	**Explain** Report Reason		

Figure 2.4: Actions and Interactions (Azzopardi et al., 2018)

	Proactive		Reactive	
User	**Q**uery	Information	**F**eedback	Positive
		Prompt		Negative
Agent	**R**equest	Offer	**A**nswer	Results
		Understand		Backchannel
				Empty

Figure 2.5: Functional Annotation Schema (Vakulenko et al., 2019)

2.3 Approaches to Conversational Systems Research

Research in IR, and in Conversational Search Systems have followed two distinct approaches – user-centered and algorithmic. The former focuses on the user behavior and preferences while the latter is concerned with building predictive models and systems in a data-driven way. In this section, we try to highlight the breadth of this domain by providing an overview on the research approaches. The first part discusses research involving users while the second part elaborates on the applications of machine learning algorithms in conversational search systems.

2.3.1 User-centered Research

Human information-seeking behavior on the Web is governed by the affordances of browser-based search systems. The primary actions of the user can be categorized as query input, result selection recommendations, and item selection (White, 2016). The search system has become more interactive over the years. It not only responds with the ranked list of relevant documents, but also completes the query automatically, recommends alternate queries, and provides short summaries or snippets of results. Research in spoken conversational search systems has focused on various aspects of user-centered development and design: input mechanism (interface design, query behavior, and spoken dialogues), information processing (dialogue analysis and user intent detection), and result presentation (result summarization, system clarification, prosody). This line of research requires a controlled environment, involves users or participants, focuses on detecting the user's behavioral patterns and preferences, and makes design recommendations.

Interface Design

Although speech is the most natural way of human-human communication, yet typed-in text emerged as the primary mode for human-computer communication. Any conversational system can be thought of as an interface or platform through which the users can interact with computer applications like databases in spoken natural language. With

the improvements in automatic speech recognition (Xiong et al., 2018) and machine learning (L. Yang et al., 2018), the development of advanced voice-based search interface has become possible. In the early 1990s, the available interfaces were limited to small interactive programs that could handle simple requests over the phone. Currently, spoken interfaces are being used for question-answering systems, virtual digital assistants (McTear, Callejas, & Griol, 2016), and domain-specific tasks (Walker, Passonneau, & Boland, 2001). Interacting with computer systems through speech still remains unnatural (Klemmer et al., 2000; Turunen et al., 2012). Newer technologies like Google Home and Amazon Alexa have a high acceptance rate because of novelty but they leverage one of the major advantages of voice-based conversational systems: they support operation without touching or looking at the device (Cohen & Oviatt, 1995; McTear et al., 2016; Yankelovich, Levow, & Marx, 1995).

Research in spoken search interfaces (Gibbon, Moore, & Winski, 1997; McTear et al., 2016; Varges, Weng, & Pon-Barry, 2009) explored how to translate dialogues to database queries, the optimum amount of information to be presented back, and if providing a summary of irrelevant options increases user's confidence (Demberg & Moore, 2006) One of the most useful properties of any interface is to provide real-time, context-aware feedback for any user action (Griol, Carbo, & Molina, 2013), such as highlighting the query terms in the search results (Hearst, 2009). Ajmera et al. (2011) proposed to acoustically highlight the keywords in audio that contain the query terms while Tuuri, Eerola, and Pirhonen (2011) used prosodic non-speech audio feedback for physical training application. A beeper-style sound stimulus was introduced to express four different meanings, and the perceived expression by a group of users was noted. It was found that prosodic characteristics provide useful information to correlate between communicative function and acoustic descriptions (Tuuri et al., 2011). Yankelovich et al. (1995) conducted a detailed study for designing the speech user interface and recommended that the system-level dialog must be short and informative. Winterboer, Tietze, Wolters, and Moore (2011) proposed a user model-based summarize and refine (USMR) system and highlighted the importance of discourse markers. Higashinaka et al. (2014) discussed the challenges of open-domain conversational systems. Such

systems are hard to develop as there are no bounds on the topics, the utterances, and the direction of the conversation. Sugiyama, Meguro, Higashinaka, and Minami (2013) proposed that system utterances could be generated by using a combination of salient words in the user utterance and related words gathered from social media.

Query Behavior

The differences between voice-based and text-based search systems have been assessed by several researchers (Arguello, Choi, & Capra, 2018; Crestani & Du, 2006; Guy, 2016; Sa & Yuan, 2019; X. Yuan, Belkin, & Sa, 2013; X. J. Yuan & Sa, 2017). Most of the studies collect data through controlled laboratory-based observational studies or crowdsourcing (Arguello et al., 2018). Guy (2018) suggested spoken queries are closer to natural language than typed in queries. In a separate study, Crestani and Du (2006) concluded that spoken queries are more prolonged than written queries but do not have any positive effect on the information retrieved. Sa and Yuan (2019) investigated the partial query modification patterns in spoken conversational systems using a Wizard-of-Oz study while X. J. Yuan and Sa (2017) assessed the user query behavior for different task types using spoken and textual interfaces.

Guy (2018) extended the previous work to explore how voice-based search differed from text-based search. The author suggested that spoken queries are mostly used for audio-visual retrieval, on the go, and therefore, the browsing behavior of people is different when using spoken queries. Studies (Yi & Maghoul, 2011) have also shown that voice queries are typically longer but use twice the number of stopwords and hence, are not linguistically richer when compared to typed-in queries. The findings of Sahib, Tombros, and Stockman (2012) claimed otherwise, arguing that voice-based queries are about 2.5 words long, which is less than typed in mobile queries (2.9 words) and typed queries in desktop search (3.1 words).

Experimental Study Designs

There have been several studies that attempted to understand the user behavior and preferences for conversational agents (Arguello et al., 2017; Avula, Chadwick, Arguello,

& Capra, 2018; Begany et al., 2015; McDuff, Thomas, Czerwinski, & Craswell, 2017; Thomas et al., 2017; Trippas et al., 2018, 2017a, 2015b; X. Yuan, Belkin, Jordan, & Dumas, 2011). Most of these studies use crowdsourced workers or Wizard of Oz techniques (Avula et al., 2018; Thomas et al., 2017; Trippas et al., 2018, 2017a, 2015b) and almost all of them monitor the interaction patterns between the searcher and the agent (Dubiel et al., 2018; Teevan, Alvarado, Ackerman, & Karger, 2004; Trippas et al., 2018; Vtyurina, Savenkov, Agichtein, & Clarke, 2017).

To understand the types of conversations that take place between the user and the intermediary, it is essential to observe the searcher-intermediary interactions during the task. The collected data usually includes audio and video signals along with transcripts of the dialogues. The collected datasets (Thomas et al., 2017; Trippas et al., 2018) are useful to researchers for understanding the patterns of human-human interactions that occur between the searcher and the intermediary. The research outcomes provide insights on ideal human-computer interactions and search tactics (Gibbon et al., 1997) for effective voice-based searching.

Some other studies (Larson, Jones, et al., 2012; Trippas et al., 2017a; Zarisheva & Scheffler, 2015) discuss protocols for transcribing the audio and annotating the utterances. Annotating the dialogues of a conversation is based on the hypothesis that each distinct class will provide an insight into the user and system behavior during the conversation (Reithinger & Maier, 1995). Various studies (Allen & Core, 1997; H. Bunt, 2009; Kim, Chern, Feng, Shaw, & Hovy, 2006; Qu et al., 2018; Searle & Searle, 1969) have explored the process of annotating and developed different annotation schemas and classification taxonomies for spoken conversational systems. The annotation schema could focus on the effectiveness of the conversation (Kim et al., 2006), user intent (Qu et al., 2018), or functionality (Trippas et al., 2017b). H. Bunt (1999, 2009) proposed a dialogue system called Dynamic Interpretation Theory (DIT) to categorize four different types of human-human dialogue. DIT uses various communicative factors like semantic, social, linguistic, cognitive, and physical contexts to create an annotation schema.

Studies have also been conducted to evaluate and improve user satisfaction in conversational systems (Jadeja & Varia, 2017; Kiseleva & de Rijke, 2017; Kiseleva et al., 2016; Mehrotra, Awadallah, Kholy, & Zitouni, 2017). Thomas et al. (2018) explored user-specific conversational styles to improve user engagement. Others have worked on querying by voice (Utama, Weir, Binnig, & Çetintemel, 2017), reformulations of spoken queries (Nogueira & Cho, 2017), characteristics of spoken query (Guy, 2016). Some research works have also explored identifying user intent through query suggestions (Radlinski & Craswell, 2017), clarifications (Aliannejadi, Zamani, Crestani, & Croft, 2019), and negative user feedback (Bi, Ai, Zhang, & Croft, 2019).

2.3.2 Applications of Machine Learning

Conversational systems, as the name suggests, involve dialogue between the user and an automated system. However, such interactions are not restricted to audio-channels only. For example, recent works on conversational systems have not only focused on IR systems but also on language modeling, question-answering, chatbots, and image captioning systems. Although retrieval is the most important part of conversational search systems, yet it requires applications from other areas like natural language processing, speech recognition, artificial intelligence, cognitive science, and so on. Deep neural models have a profound influence on all the above developments because of their ability to process massive volumes of raw data to discover the complex, non-linear relationship between input and output hierarchically. In this section, we discuss the studies in conversational systems that involve various applications of machine learning.

The architecture pipeline of a spoken conversational system has been highlighted by Kotti et al. (2017) in their work (Figure 2.6). The process starts with recognizing speech (voice-query), then processing and understanding it in the context of the user and his situation, and generating the response. As most retrieval systems work best with text, an efficient, conversational system should identify the spoken utterances and convert the spoken query into text using automatic speech recognition (ASR) (Hinton et al., 2012). Next, the system needs to develop an understanding of the natural language (Natural Language Understanding or NLU), manage the states of the conversation (Dialogue

Management), and generate an appropriate response (Natural Language Generation or NLG). The response generation involves identifying the domain of conversation, user intent, and user-, language-, and intention-modeling. Finally, the textual response needs to be converted back to audio (Text-to-Speech or TTS) for the user. The final stage would be to evaluate the quality and usefulness of the user-system interaction (Quality Evaluation).

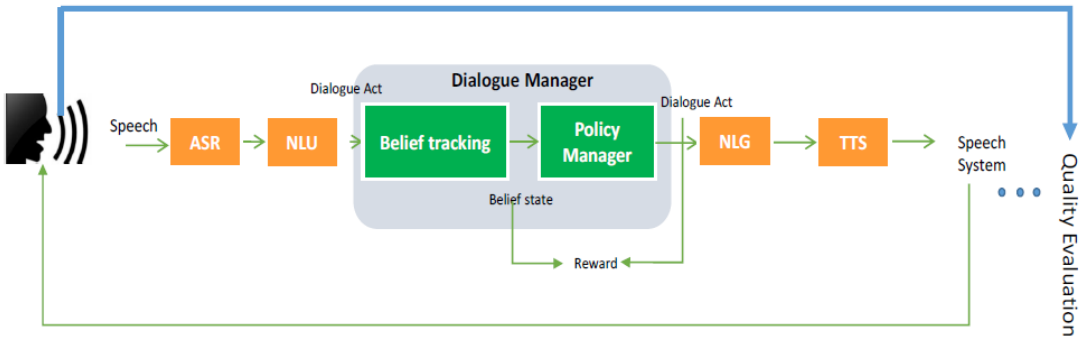


Figure 2.6: Architecture of Spoken Conversational Systems (Kotti et al., 2017).

Machine Learning (ML), and Deep Neural Networks (DNN) specifically, have provided immense support in automatic speech recognition and natural language processing, understanding, and generation. The applicability of deep neural networks in language modeling has been proven by the works of Bengio (2012), Mikolov, Karafiát, Burget, Černockỳ, and Khudanpur (2010), and Mikolov and Zweig (2012). By representing words in a document using embeddings, it is possible to capture the context and semantic relationship between words. Word embeddings successfully capture the grammatical structures and the domain information from the training corpora. As we move hierarchically from words to sentences to documents, the sparsity could be tackled using different vector manipulation techniques. Ganguly, Roy, Mitra, and Jones (2015) applied generalized language modeling to reduce the vocabulary gap and enhance retrieval efficiency. The authors used word embeddings to find transformational probability between terms present in a query and a document. In separate work, Zamora-Martínez, Espana-Boquera, Castro-Bleda, and De-Mori (2012) augmented the neural language model with a cache model to capture long-term dependencies.

Retrieval models work well only when there is a large amount of data to select candidate responses. Yan et al. (2016) mined context using previous utterances and obtained the relationships between Query-Reply, Query-Posting, and Query-Context pairs. The responses were ranked using a convolutional neural network (CNN). In separate work, F. Yang, Mukherjee, and Dragut (2017) used the Dual-LSTM chain model to rank the responses and to suggest the next utterance simultaneously. Both the models proposed were trained using public posts and their associated replies available on the Web. Wu, Wu, et al. (2016) used a sequential matching network for response-utterance matching and an RNN to obtain the relationship between utterances. This model was trained using the Ubuntu Dialog Corpus (Lowe, Pow, Serban, & Pineau, 2015), which contained around one million context-response pairs obtained from the chat logs of Ubuntu Forum. The latest work by L. Yang et al. (2018) used a retrieval-based module with additional modules for knowledge extraction and pattern matching.

Sordoni et al. (2015) trained the model (Recurrent Neural Network Language Model) end-to-end using social media data. The model was data-driven, context-sensitive, and open-domain. On the other hand, Serban, Sordoni, Bengio, Courville, and Pineau (2016), Serban, Sordoni, et al. (2017), and Serban, Klinger, et al. (2017) developed different generative models for goal-oriented (Serban et al., 2016) and non-goal-oriented dialogues (Serban, Klinger, et al., 2017). Serban et al. (2016) used RNN and n-grams to build a hierarchical recurrent encoder-decoder. The open-domain dialogue model was pretrained on a large corpus of question-answer pairs, and the response was generated word by word. Serban, Klinger, et al. (2017) used a hierarchical sequence-to-sequence framework with multiresolution RNNs to build the conversational model. Two sequences were generated, the high-level coarse tokens and the low-level natural language words. This conversational model, which was built for the technical support domain, generalized well to new examples. In a different work (Serban, Sordoni, et al., 2017), the authors used twitter conversations to train an RNN architecture with latent variables. The latent variables modeled the complex dependencies in the sub-sequences and helped in the generation of long utterances. However, generation-based models are often criticized for generating dull responses which are not informative to the user.

Dialogue systems may comprise single or multiple turns. Single-turn systems (Shang, Lu, & Li, 2015; H. Wang, Lu, Li, & Chen, 2013; M. Wang, Lu, Li, & Liu, 2015; Xing et al., 2017) are easier to build, and models question-response pairs, but the lack of context makes them ineffective for long conversations. While Shang et al. (2015) used an encoder-decoder framework, M. Wang et al. (2015) matched two short texts to produce the response. They mined matching patterns using a dependency tree and then used those patterns to train the deep neural network. H. Wang et al. (2013) used a retrieval-based approach to match the candidate responses. The authors trained the neural network on a Chinese microblog service. Multi-turn systems (Serban, Klinger, et al., 2017; Serban et al., 2016; Serban, Sordoni, et al., 2017; Sordoni et al., 2015) can preserve the context by storing the entire conversation session as a vector. Use of RNN (Shang et al., 2015; Sordoni et al., 2015) is typical for modeling context and temporal information in multi-turn dialogue systems. Some other works have focused on different aspects of conversational IR systems: The work of Williams et al. (2014) and Zhao and Eskenazi (2016) tackles dialog states using deep reinforcement learning; Z. Chen, Yang, Zhao, Cai, and He (2018) used CRF-structured network with end-to-end training to identify dialogue acts; Kotti et al. (2017) used CNN to predict unsuccessful dialogues in the early stages of the conversation. Ren, Malik, Ni, Ke, and Bhide (n.d.) used deep learning (LSTM with attention) and word embeddings to understand the questions in a multiturn conversation. C. Liu, Xu, and Sarikaya (2015) used the deep neural network for understanding spoken language. W. Wang et al. (2018) enhanced the conversational experience by increasing the coverage (by extending the topic to related ones) or focus of the conversation (by selecting important keywords). The authors used an encoder-decoder framework with recurrent neural networks (with attention) and multilayer perceptrons for encoding and RNN for decoding. Boroş and Dumitrescu (2015) designed a high-quality neural text-to-speech system (using autoencoders) for smart devices that require minimal memory and no internet support. The research proves that DNNs could be efficiently applied to mobile devices, with minimal decline in performances.

While task-oriented interactions help the user in accomplishing specific tasks (Wen

et al., 2016; Williams & Zweig, 2016), non-task-oriented dialogues can be chitchat or informational (W. Wang et al., 2018). Some examples of task-oriented systems would be digital personal assistants who can set alarms, play music, and place an online order at the request of the user. Both Wen et al. (2016) and Williams and Zweig (2016) proposed task-oriented systems, which operate in a text-based environment (chatbots) and can be trained end-to-end. While Wen et al. (2016) use CNN feature extractor with a Jordan-type RNN belief tracker, Williams and Zweig (2016) used Long Short-Term Memory (LSTM). Chitchat systems (Shang et al., 2015; Xing et al., 2017; Y. Zhang et al., 2018) are purely conversational without any goal orientation and are mostly open-domain (e.g., ELIZA). Informational systems, on the other hand, help the user in exploring some topics through conversation. Shang et al. (2015) and Xing et al. (2017) generated responses for short textual conversations. The former implemented an encoder-decoder framework using RNN, while the latter used a topic-aware sequence-to-sequence model. Persona-based conversational models are also popular as they capture the styles and backgrounds of individual users: Li et al. (2016) used LSTMs while Y. Zhang et al. (2018) used memory augmented neural network to develop persona-based chatbots.

Machine (and deep neural) learning has been a popular choice for solving various problems in conversational IR (Jadeja & Varia, 2017; Li et al., 2017; Yan & Zhao, 2018). However, implementing a deep neural conversational IR model has many challenges (Jadeja & Varia, 2017). A single objective function may not be enough to capture the different goals of the conversational system. Recent works (Shang et al., 2015; Sordoni et al., 2015) in recurrent neural networks have been more successful in modeling context in conversations using natural language (Vinyals & Le, 2015). While CNNs use fixed-size windows to handle variable-length word sequences and identify local patterns, RNNs can capture information in a word sequence. Y.-N. Chen, Hakkani-Tür, and He (2015) trained the model to learn intents and utterances using tri-letter vector embeddings. Jadeja and Varia (2017) highlighted the different challenges and suggested how Deep Reinforcement Learning can be used for implementing the conversational IR models. Li et al. (2017) proposed an adversarial framework: the generator to produce responses

sequences and discriminator to differentiate between human and machine-generated dialogues (reward function). Microsoft Research has developed a specialized deep learning architecture called deep semantic similarity model (DSSM) and convoluted-DSSM for conversational web searches (Deng, 2016).

Some of the recent works use deep neural networks to answer complex questions (Kenter & de Rijke, 2017), predict the success of dialogues (Kotti et al., 2017), improve contextual awareness (Yan et al., 2016), reformulate multiturn questions (Choi et al., 2018; Gao, Galley, Li, et al., 2019; Ren et al., n.d.), and present exploratory search results as interactive stories (Vakulenko et al., 2017). Much work has also been done on conversational recommendation (Michalski, Charlin, & Pal, n.d.; Micoulaud-Franchi et al., 2016; Sun & Zhang, 2018; Y. Zhang et al., 2018) where the authors use end-to-end frameworks for e-commerce, movie, music, healthcare, and banking industries.

2.4 Chapter Summary

The interdisciplinary nature of conversational search as a domain has led to overlapping research in the fields of human-computer interaction, machine learning, and information science. In this chapter, we have described the background and research works related to this thesis. We discuss the seminal papers that define conversational search systems and the properties expected of such systems. As "conversational" is a broad term covering any natural language dialogue-based interaction, conversational search systems could be categorized into text-based chatbots or voice-based personal assistants. Similarly, they may or may not be task-oriented and could involve question-answering and recommendations. Next, we discussed some of the papers which contribute towards building a conceptual framework for both information retrieval systems and conversational systems.

As spoken conversational systems are still in a nascent stage, numerous challenges need addressing soon. Research in conversational systems has explored context-building in conversations, user query behavior, user intent recognition, and results presentation. However, there are some open research problems in natural language understanding

and user-system communication. For example, current state-of-the-art models perform poorly as the search tasks get more complex. To exhibit human-level cognition while searching, search systems must understand the natural language utterances by the user and respond accordingly. Also, the user-system communication is problematic in the absence of system-level clarifications. As the search actions of the system are not revealed to the user, it is possible for the system to incorrectly interpret the user's information problem and therefore, pursue a wrong search direction.

Research in conversational search systems could also be categorized as user- or system-oriented depending on the research methodology. The choice of methodology is motivated by the problem being investigated by the researcher. User-oriented research concentrates on the end user and explores different aspects of system design and interfaces. By conducting user studies, the researchers observe user behavior in controlled or naturalistic setting and evaluate system usability and user preferences. System-oriented research focuses on the algorithmic development, which includes (but is not limited to) the use of machine learning to process large-scale data and develop insights. We describe both of these research approaches as they inform our choice of methodology and study design.

Chapter 3

Research Questions

In this chapter, we discuss the theoretical structures which guide the direction our dissertation research and the research questions that follow from those structures. In the following sections, we briefly recapitulate the research gaps and discuss the theoretical models which furthers our understanding of the problem and helps us in proposing the research questions.

3.1 Facilitating User-system Conversation

Over the last few years, we have witnessed many research efforts towards the development of voice-based personal assistants. While the existing state-of-the-art personal assistants perform satisfactorily for simple tasks, they are insufficient as search systems. As the search tasks get more complex, it necessitates a greater effort to understand the user's information problem. The effort could be in the form of higher cognition (as in a human-human conversation), more contextual awareness (through better knowledge representation), or longer conversations.

Most of the existing conversational search systems fail to identify the information need and the search context of the user. This can be attributed to the lack of effort from the agent to explicitly clarify its understanding of the user with the user himself. As the user has minimal control over the search process, and the actions of the system are hidden from the user, it leads to vicious cycle where every iteration leads the search away from the preferred direction (of the user). Many conversational search episodes lead to failures as the user is unhappy about the clarity of communication, the quality of the search performed, and the reliability of the information returned.

To address this problem, we rely on the theoretical understanding of discourse as

proposed by prior works (Brooks & Belkin, 1983; Hollnagel, 1979). Conversational search dialogues are mostly goal-oriented, where the user's goal is to solve his problem using the information provided by the system (or intermediary). The system's goal is to provide the most relevant and useful information available to it. The search situation is not well-defined (more so for complex search tasks) as the goals of the user can differ significantly based on his status, socio-cultural position, and self-image (Brooks & Belkin, 1983). However, the subjective nature of the user goal can be handled through co-operative dialogue, which iteratively enhances the knowledge of the system about the needs of the user.

The types and models of interaction required in intentional discourse have been discussed in Hollnagel (1979), where the focus is more on the functional analysis of discourse rather than linguistic. Failures in human-human communication are often due to the inability to understand the other person. This is exacerbated by any reluctance to clarify our understanding of the conversational partner explicitly. Therefore, Hollnagel (1979) proposed the joint cognitive system, where the author stressed the importance of shared knowledge to achieve an effective and successful communication. Hollnagel (1979) highlighted that any successful communication must involve both the parties having a clear model of each other (Figure 3.1). This model – which is an abstract level of understanding – involves knowledge of the preferences, long- and short-term contexts, and personality of the conversational partner. We believe that in any user-system information-seeking conversation, the user should clarify his search preferences, search context, and information need to the agent, and the agent should reveal its actions, limitations, and capabilities to the user (Radlinski & Craswell, 2017).

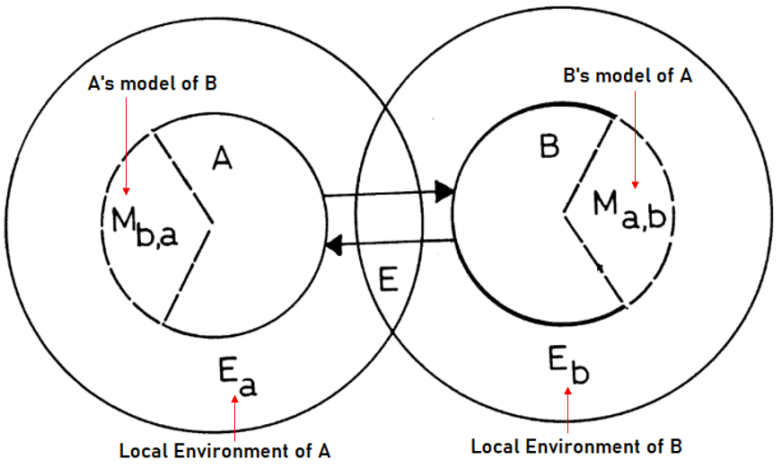


Figure 3.1: Paradigm for Communication (Hollnagel, 1979).

The framework proposed by Hollnagel (1979) has been used previously to develop a model of clarity (Belkin, 1988). In his work, Belkin (1988) defines clarity as a state where the user develops an understanding of the components and functioning of the system he is interacting with. The authors also highlight that such clarity can be achieved through "overt explanation." In a futuristic system with high levels of cognition, the clarity is guided by the requirements of interaction and can be achieved by explaining the ongoing search process, the intended actions (by the system), and the information resources (Belkin, 1988). While the system should explain its model of the user on-demand or during the search process, both the user and the system can also engage in secondary communication by providing more information about each other, the search task and the context.

While these interactions could potentially help the user and the agent to understand each other better, it is essential to evaluate if explicit clarifications are necessary and desired by the user. In our work, we envision a system that does not possess human-like cognition and is feasible in the next few years. Therefore, in our work, we explore the influence of explicit system-level explanations, which are cognitively less complex and are provided unprompted by the system. The clarification provided is limited to the search system's model of the user and the user's information need at different points in time, and includes explanations about the queries and information sources used by the

system. We investigate if such clarifications can facilitate the user-agent communication and improve the user's search experience.

Therefore, we formulate our first research question as:

> *For moderately complex tasks, can we determine the influence of explicit system-level clarification on the user's search experience?*

3.2 Improving System's Understanding of the User's Search Problem

Current state-of-the-art conversational search systems often fail to recognize the information need of the user, especially for exploratory and complex tasks where the question is non-factoid in nature. In any conversational search environment, it is important that the agent understands the utterances by the user and performs the appropriate search activity. This is a domain-specific natural language understanding problem where the user's utterances guide the agent's action. Spoken dialogues communicate the search intent and the information need of the user (searcher) to the agent (intermediary). In response, the agent performs specific, expected search actions.

To better understand the natural language utterances of the user during an information-seeking conversation, we situate our research using the theoretical frameworks developed in intelligent systems, discourse, and dialog communities (Sitter & Stein, 1992, 1996; Stein & Maier, 1995; Winograd et al., 1986). During spoken conversational search sessions, the user considers the system as a conversational partner, and the system is expected to interpret the user's commands, and intentions as an intelligent cognitive being should. However, the user has the freedom to explain his problem using natural language and can be verbose, vague, and incoherent. The user may not plan the conversation, and there may not be any specific goal to achieve. Also, the users change the goals and strategies during the search session (Stein & Maier, 1995). Therefore, Winograd et al. (1986) represented the searcher-intermediary conversation (during information search) using a state transition network. The authors highlighted that any information-seeking dialogue could be modeled based on different speech or dialogue

acts. This formed the basis of the Conversation for Action (CfA) model – a theoretical model which could simulate any conversation – where each node represents one of the dialogue states, and the arcs are speech acts which help to transition from one dialogue state in the conversation to the other. The CfA model is designed to achieve computer-supported "negotiations" between humans. The model combines the philosophy of language with interpretations in context. The meanings of utterances are socially constructed, and the behavioral expectations of the participants control the flow of conversation.

Later, the CfA model was modified and extended by Sitter and Stein (1996) through their analysis of information-seeking dialogue scenarios. The authors proposed the Conversational Roles (COR) model, which categorizes the dialogue acts based on the intention of the participants and aims to guide the users through different stages of IR by formulating a dialogue plan. The participants assume the roles of information seeker and provider in alternate turns. The COR model focuses on the interpersonal nature of dialogues and facilitates the acquisition of knowledge about the conversational partner incrementally and collaboratively (Stein et al., 1999). COR considers not only the illocutionary aspects of conversation but also the communicative and functional aspects. By using the local discourse structures and analyzing the different dialogue acts and their interrelationships, the user and the agent can formulate conversational tactics (Belkin et al., 1995).

Stein et al. (1999) aimed to represent all possible scenarios in a human-human dialogue through this model - acceptance and refusal of conversational roles, requests, promises, offers, negotiations, and evaluation - by constructing a recursive state-transition-network (Figure 3.2). The COR model highlighted the various roles which the searcher and the intermediary assume during search tasks and formed the basis for various prototypical IR systems like MERIT, CORINNA, MIRACLE, and SPEAK!

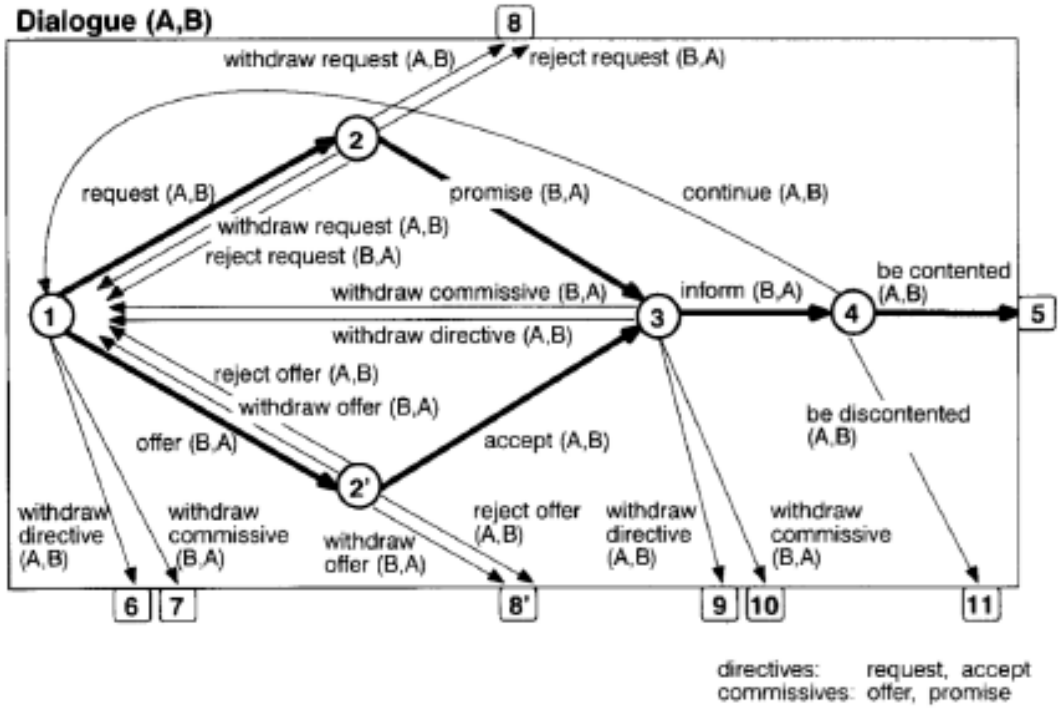


Figure 3.2: Conversational Roles Model (Sitter & Stein, 1992).

Therefore, to develop conversational search systems that possess higher-order cognitive capabilities to understand the natural language utterances of the user, a viable strategy would be to identify the speech (or dialogue) acts in the dialogue. By using the discourse structures present in conversational dialogues, it could be possible to identify the functional aspect of it (the goal or intention of the user). Our second research question – which is motivated by the COR model and focused on improving natural language understanding in conversational search systems – could be stated as follows:

> *How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?*

Information-seeking dialogues are task-oriented, and therefore, our problem is reduced from open-domain to search task-specific scenarios. We use the theories in discourse analysis and human-machine interaction to further the natural language understanding of conversational search systems. The actions of the system – which could either be a spoken utterance or a search action – is governed by the spoken utterances of the user. In our work, we use the speech acts to identify the intention of the user

and infer the functional aspect of the utterance. Next, we predict the search function performed by the system (which is in response to the user's utterance)

3.3 Research Questions

The two research questions can therefore, be restated as:

- **RQ1:**
  For moderately complex tasks, can we determine the influence of explicit system-level clarification on the user's search experience?

- **RQ2:**
  How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?

Both of the research questions aim to address one common problem in conversational search systems, which is the inability of the system to understand the user. By answering both the research questions, we can improve the user-system interaction through better natural language understanding and communication techniques. While both the questions go beyond simple tasks involving factoid answers, the first research question addresses user-system interaction from a communications perspective. Explicit system-level clarifications can not only improve the users' search experience but also help the system to understand the users' information needs in a better way. The second research question will lead to an exploration of the problem from a systems perspective. By predicting the speech acts accurately – which convey the functional aspect of dialogues – and the search actions, we might help the system to understand the intention of the user and determine the appropriate search strategy.

3.4 Chapter Summary

In this chapter, we discuss the different theories which motivates this research and guides the overall direction of this dissertation research. By our review of the seminal work done in discourse analysis and intelligent systems, we identify the theoretical

underpinning on which are research questions are developed. We have formulated two research questions focused on the design and development of user-centered conversational search systems: the first from a human-computer interaction perspective and the second from the perspective of natural language understanding. The research questions dictate our choice of methodology and analysis.

Chapter 4

Experimental User Study Design

In this chapter, we discuss the proposed methodology to answer the research questions. We discuss the user study design, the experimental setup, the search tasks, and the set of protocols that were followed. The types of data collected and the data collection mechanism has also been explained.

4.1 Methodology

First, we will restate our research questions:

- **RQ1:**

  For moderately complex tasks, can we determine the influence of explicit system-level clarification on the user's search experience?

- **RQ2:**

  How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?

To answer the first research question, we needed user-system interaction data that contained a detailed record of the user- and the system-dialogues. However, as our research focus was on system-level clarifications, the interaction data should contain such clarifications. There was another caveat: the search task performed by the system must not be simple. To assess the influence of system-level clarification on the user's search experience, we needed to use data that contained some form of user feedback. This feedback could be collected post-search or during search using some evaluation strategy. To reduce the variances and confounding factors, we opted to design a laboratory-based, controlled, user study. The second research question could be answered by collecting

user-system spoken dialogues during a conversational search session. As we were not investigating any specific outcome variable, the user study data (collected for RQ1) was sufficient to answer the second research question.

Existing conversational search systems have limited cognition and cannot sustain long conversations or complex questions. Therefore, such systems are not suitable for inspecting the functionalities that do not yet exist. Most of the research in this domain relies on a human Wizard to simulate the role of the intermediary (which is a futuristic version of conversational search systems). Such Wizard-of-Oz experiments help us gain insight into the functionalities, expectations, and design of conversational agents of the future. An assessment of publicly available datasets revealed that the data collected was not suitable to answer the research questions which we proposed. Therefore, we performed a within-subjects Wizard-of-Oz experiment, which had two setups – one experimental and one control – to explore the role of clarification in user-agent dialogue.

4.2 Study Design

The user and the intermediary were asked to complete different search tasks in a laboratory setting. The experiment was a within-subjects design, in which the users performed three different search tasks using the two different systems:

1. The baseline or control system ($S_{control}$), where the interaction between the user and the intermediary follows the normal flow of conversation; and

2. The experimental system ($S_{experimental}$), where the intermediary explicitly stated its models (or understanding) of the user and the user's information need.

The baseline system ($S_{control}$) helped us to collect the conversational dialogues between the user and the agent, their interaction and behavioral patterns, and the web search activities of the agent in response to the user's questions. No explicit system-level clarifications were involved in this system. In the experimental system, every time the agent searched for or returned some information, it explicitly clarified its model of the user. As the model is an abstract representation, we instructed the Wizard to

clarify the query terms, information sources, and its understanding of the user's information need. The clarifications provided by the Wizard was cognitively simple and included system revealments. The Wizard revealed his actions to the user every time he searched and asked for confirmation to proceed with the search. While the baseline system served as the control, the experimental system helped us to answer our research question.

There were a total of three tasks (the details of the tasks are provided in the next section), the first being a warm-up task to familiarize users with the system. The next two tasks were of moderate complexity levels (developed using the Taxonomy of Educational Objectives). These tasks helped us in understanding the interaction patterns and search experiences of the users when seeking information. The data collected from both the systems were used to create the CONVEX dataset.

		Search Task	
		Conference	Perfume
System Used	Control	System: Control Search Task: Conference	System: Control Search Task: Perfume
	Experimental	System: Experimental Search Task: Conference	System: Experimental Search Task: Perfume

Figure 4.1: Task System Combination.

Configuration	Task Order	Search Task	System Used
1	1	Conference	Control
	2	Perfume	Experimental
2	1	Conference	Experimental
	2	Perfume	Control
3	1	Perfume	Control
	2	Conference	Experimental
4	1	Perfume	Experimental
	2	Conference	Control

Figure 4.2: Search Task, System Used, and Order of Presentation.

The combination of the tasks and the systems (Figure 4.1) were rotated for different users to avoid the influence of learning effects on statistical analysis. We also rotated the order in which these task-system combinations were presented. Figure 4.2 shows

the 2*2*2 matrix of the search task, the system used, and the order of tasks. We used the four configurations shown in the figure for our study. In all experimental setups, the user had to talk to the Wizard over the audio channel and explain his information need. The Wizard assessed the user's information need and controlled how the responses are presented back to the user. The various forms of response were facts, summaries, or follow-up questions, which were typed in by the intermediary and converted to audio using text-to-speech software. In our experiment, we do not inform the user that the intermediary is a human and not an automated system. For the warm-up and experimental tasks, the users were allowed a maximum of 10 and 20 minutes, respectively, after which they were asked to fill up a questionnaire and move to the next task. The Wizard was present for the entire duration of the study and was allowed a five-minute break after each task. The Wizard was also allowed to leave once the search tasks were completed, and the exit interview started. Participants were allowed to leave at any time, but to be compensated, they were required to either complete all the tasks or spend at least one hour in the laboratory engaging in the search task. Overall, the user and the agent had to converse with each other to resolve the user's information problem and complete the tasks successfully.

4.3 Search Tasks

In this section, we explain how the search tasks were conceptualized and created. We created backstories or simulated search situations (Borlund, 2000) to place the user in real-life information seeking scenarios. Such a situation comprises two parts: the backstory and the search task. The backstory provides context about the search task and situates the user in the simulated task and creates an information need. This promotes a more natural search behavior (Borlund, 2002). The search task serves to mitigate the knowledge barrier and satisfy the information need. The description of the task instructs the searcher on what to do without explaining how to do it. The searcher interprets the task in the context which is influenced by the backstory. According to (Borlund, 2003), the situation should have the following characteristics (White, Jose, & Ruthven, 2006):

1. realistic enough so that the user can relate himself to it;

2. topically interesting, and;

3. provides an imaginative context for the user to apply the situation.

4.3.1 Task Complexity

Bloom, Krathwohl, and Masia (1956) proposed the taxonomy of educational objectives to develop learning objectives. The goal was to foster more creative thinking in the classroom instead of rote learning. In this taxonomy, Bloom proposed a hierarchical order of cognitive domains - Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.

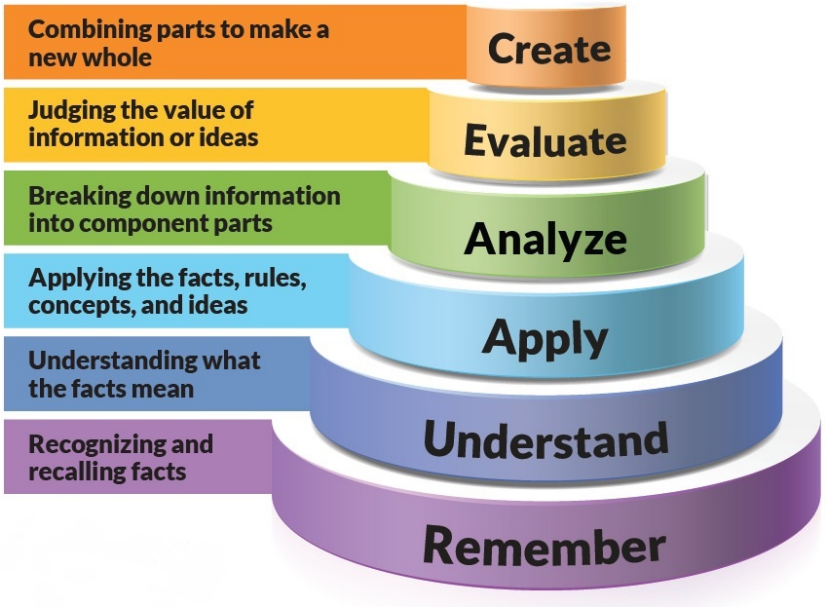


Figure 4.3: Taxonomy of Learning (Krathwohl, 2002)

Later, Anderson et al. (2001) and Krathwohl and Anderson (2001) revised the taxonomy to create a two-dimensional framework. The authors separated the noun and verb aspects of the original taxonomy to the knowledge dimension and the cognitive process dimension, respectively. The revised taxonomy (Figure 4.3) made significant modifications to how the cognitive process dimension - consisting of Remember, Understand, Apply, Analyze, Evaluate, and Create levels in a hierarchical order, from lowest to highest - interacts with the knowledge dimension (Factual, Conceptual, Procedural,

and Metacognitive). While the lowest levels involve additions of new information to our existing knowledge structures, the middle levels involve reshaping and accommodating knowledge, while the highest levels may require deleting and restructuring pre-existing knowledge. The importance of the taxonomy lies in the fact that it views educational goals as not only recalling facts but creating a more comprehensive understanding and application of the learned concepts.

4.3.2 Task Development

The first task was a warm-up task which familiarized the participants with the search system, the search process, and the intermediary. The next two tasks (on different topics) were assigned one after the other and involved the use of the baseline or the experimental system. The warm-up task was of low complexity, while the two main tasks were of moderate complexity levels. Our tasks, which simulated naturalistic search behavior among experimental subjects, were adopted from the literature (Byström & Järvelin, 1995; Kelly, Arguello, Edwards, & Wu, 2015; White et al., 2006; Wildemuth & Freund, 2012). While the warm-up task was from the Remember and Understand levels of Bloom's taxonomy, the main tasks were from Analyze and Evaluate levels (highlighted in Table 4.2 and Table 4.1). The tasks (adopted from White et al. (2006) and Kelly et al. (2015)) initiated a multi-turn conversation between the searcher and the intermediary. While present state-of-the-art systems handle simple tasks with factoid answers, it was important to develop the tasks which would be more complicated (but not too much) and yet be representative of the search tasks likely to be presented to the future systems.

The three search tasks are described in Table 4.1. Table 4.2 categorizes the tasks based on cognitive complexities and outcomes. The warm-up task required the users to find some factual information. They were expected to identify the facts and write them in the response. The experimental tasks were more complex as they involved a comparison of different results and the users were required to come up with a recommendation.

Table 4.1: Search Tasks

Tasks with Backstory
Warm-up Task (Level: Remember and Understanding)
It is May, and you are in North America. You have been sneezing every morning and have red, watery eyes. Investigate the following: (1) What could be the cause of that? (2) Medicines and/or Home remedies. In a few lines, state your findings.
Experimental Task 1 (Level: Analyze and Evaluate)
You are a researcher planning to submit your first paper in one of the top conferences in Artificial Intelligence and/or Natural Language Processing. However, you are not sure which conference would be the best option for you. To finalize the conference, you may want to consider the following factors: (1) Are you interested in Artificial intelligence or Natural Language Processing? Or both? (2) Writing the paper takes time. Are you looking for early- or late- 2020? (3) The location could be: North America, Europe, Asia. Which one would you prefer? Once you have made the decision, write about the choices which you made and which conference you picked and why (in brief)
Experimental Task 2 (Level: Analyze and Evaluate)
You want to gift a perfume to your dad for Christmas. You know that your dad loves the smell of Bergamot in perfumes. To decide, you may consider (1) Bergamot as one of the notes (scents) in the perfume. (2) How much you want to spend (for example, you may want a perfume under 100 USD, but your budget allows you to go as high as 150 USD) (3) You may add any seasonal preference. Once you have made the decision, write about the choices which you made, and which perfume you picked and why (in a few lines)

Table 4.2: Task Categorization and Cognitive Processes and Outcomes

Task	Learning Process	Mental Activities	Target Outcomes
T_{warmup}	Remember and Understand	*Identify*, Compile	*Fact*, List
T_{main}	Analyze and Evaluate	Compile, Describe, *Compare, Decide*	List, *Recommendation*

4.4 Procedure

This section highlights the steps involved in the study – from the perspectives of both the user and the agent (Figure 4.4)

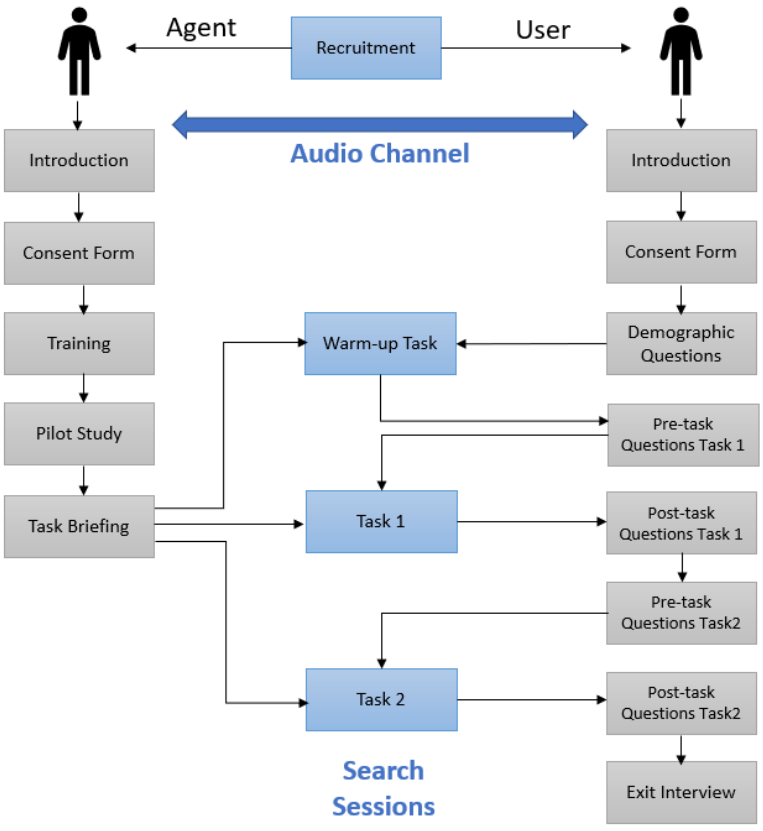


Figure 4.4: Experimental Procedure

1. Introduction

   Once the participants reported for the user study, they were given a brief overview

of the research procedure, study objectives, and the role which they were expected to assume during the search tasks. The participants were handed an Information Sheet (See Table A.2 for the Information Sheet), which explained the overall flow of the study. They were also informed that the system is in the prototypical stage and running in the server. This helped in creating a plausible theory as to why the system might be slower than existing virtual digital assistants, and why no devices were available.

2. Consent Form

   The participants were asked to sign the Consent Form[1], which allowed the researchers to collect data as per Institutional Review Board (IRB) guidelines.

3. Study Briefing

   The participants were briefed about the system and the key voice commands to operate the agent (See Table A.2). The Wizard was informed of the task-system combinations and handed the guidelines for reference (See Table A.3 and Table A.4). The study apparatus was checked to make sure that all the components were functioning as expected. Once the audio connection was set up, the participant (user) was invited into the test room.

4. Pre-test Questionnaire

   The user filled out the Pre-test Questionnaire (See Section B.1 for details). The questionnaire collected some basic demographic and background information about the user.

5. Warm-up Task

   The users were given a warm-up task (See Table 4.1 for details) to familiarize themselves with the functioning of the systems. The researcher was available to clarify any doubts which the user had about the study, the search tasks, or the functioning of the agent.

6. Pre-Task Questionnaire

[1]See Appendix Figures A.2, A.3, A.4, and A.5 for the Consent Form

For each search task, the users were required to answer a Pre-task Questionnaire (See Table B.2 for details).

7. Experimental Search Tasks

   Once we handed the task description, the users had a maximum of 20 minutes to complete the search task. The users activated the agent (Wizard) using the key phrase "Hi Joanna". They had to explain their information need to the agent, who performed the search and responded with the information. Users could end the search session at any time if they were satisfied with the information collected. The users had to write a response to the task question at the end of the search. Since there were a total of two tasks – one using the baseline and the other using the experimental system – the steps 6-8 were repeated two times. The description of the search tasks are provided in Table 4.1.

8. Post-Task Questionnaire

   For each search task, the users had to complete a Post-task Questionnaire (See Table B.3 for details).

9. Exit Interview

   At the end of the study, the researcher conducted a semi-structured interview (See Table B.4 for details) to assess the experience of the participants. The participants were compensated for their time and effort and paid 20 USD.

The entire experiment was conducted over 6-8 weeks. We recruited a total of 25 participants (users) and 1 Wizard for the study. The numbers do not include the pilot participants. The duration of each study was around one hour, which included two search tasks of 20 minutes (maximum) each and one warm-up task of 15 minutes (maximum).

4.5 Wizard of Oz Setup

One of the major challenges of conducting the study was the difference between human-human and human-machine conversation. The socio-cultural aspects of conversation

govern how human beings take turns in conversation, maintain levels of patience and politeness, and avoid interrupting the conversational partner. However, dialogue strategies are much different when humans talk to machines. The lack of cooperation from the machine is often a key factor that causes the human-machine conversation to become awkward and unsuccessful, with the human partner being stiff and embarrassed to continue the conversation beyond a certain point (Morel, 1989). The existing state-of-the-art automated voice agents can perform only simple searches and answer factoid questions. Also, the users expect such systems to provide the answer in a form suitable (short and simple, no videos, images, SERPs, or hyperlinks) for the audio environment. Fraser and Gilbert (1991) refers to human-computer dialogue as "formal, baby talk, telegraphic or computerese" due to the lack of intelligence on the part of the computer. Richards and Underwood (1984) also noted that users tend to be brief and concise, speak slowly to aid in voice recognition, use simpler vocabulary and more keywords, and take fewer turns.

This made it necessary to simulate the role of intermediary using a human operator who would observe the social norms of conversation and perform intelligent functions, but follow a set of rules to deceive the user into thinking that he is interacting with an intelligent artificial intermediary. This was the only way to ensure that we captured an authentic human-computer interaction without developing the actual state-of-art system. For the success of our study, it was necessary to simulate the role of the conversational search agent using a human operator. We identified two important requirements for the Wizard (human intermediary):

1. The intermediary should observe the social norms of conversation and perform intelligent search functions in real-time.

2. The intermediary should follow a set of rules so that the user is deceived into thinking that he is interacting with an intelligent artificial intermediary.

We recruited a human operator who was skilled in searching efficiently. The actions of the human intermediary were restricted to maintain the pretense: *the user thought that he was interacting with an intelligent automated agent who is smarter than the*

existing state-of-the-art but not as smart as a human replacement. It was the only way of ensuring that we captured an authentic human-computer interaction without developing the actual state-of-the-art system.

4.6 Observational Study and Experimental Site

The user study contained several search sessions, using three entities - the user (or searcher), the agent (or intermediary), and the search system. The user – who had no access to the internet or any other online or offline information source (apart from a spoken communication channel) – was presented with an information need (through a simulated backstory and a search task). The agent, on the other end of the communication channel (audio only), with access to a networked computer, assisted the user with the search task.

Any user study has several use cases, and the design of the study is governed by the research questions and the type of data to be collected. A Wizard of Oz study can be of two types (Bernsen, Dybkjaer, & Dybkjaer, 1996): a controlled experiment where the participants are instructed to perform some predefined tasks using a simulated system under artificial laboratory settings. Such a study, while not realistic, helps in collecting data that is reproducible as the number of variables is significantly reduced. The other type of Wizard-Of-OZ study would be an uncontrolled experiment where the participants interact with the system in a natural environment in an unsupervised manner. It is a field study where the results and interactions are realistic but not reproducible owing to a large number of variable factors. The success of our study depended on our ability to deceive the searcher by using a human intermediary in place of an artificially intelligent search system. Existing conversational agents like Siri or Alexa, works with a noise-free background. If we allowed the Wizard to be in a natural environment, the presence of any background noise could reveal the human intermediary and affect the interactions. Also, the hardware required for a remote study (to record search logs of the Wizard, the searcher-intermediary audio interactions, and the real-time conversion of text-to-speech) was considerably complex and difficult to operate for the participants. As such, in our study design, the searcher and the intermediary

were spatially separated. They were located in different rooms and were not able to see each other or communicate using gestures. The entire study was conducted in a laboratory-based controlled environment (Figure 4.5).

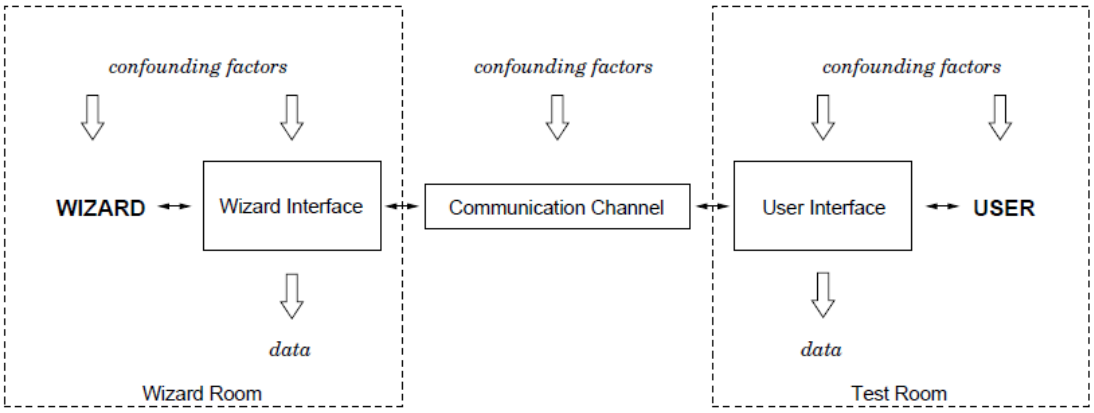


Figure 4.5: Experimental Setting (Petrik, 2004).

4.6.1 Test room for Wizard/Intermediary

The test room for the Wizard contained a networked computer that the Wizard used for searching information. We created a storyboard with the guidelines and conversation scripts. The mock system was simulated using a high-quality Bluetooth speaker (which played the user's audio to the Wizard), a microphone (to transfer the audio response from the Wizard to the user), and a networked computer with keyboard and mouse (see Table 4.3). The Wizard searched on the computer and typed in the response in textual form. The text was then converted to speech using Amazon Polly[2], which is a state-of-the-art text-to-speech system. The audio was played on the speaker, recaptured using the microphone, and transmitted to the user (in the test room) using the audio channel. The microphone served two purposes: (i) It helped us transmit the high-quality audio played on the speaker to the user (on the other side of the audio channel); and (ii) On occasions where the user interrupted the Wizard, the microphone allowed us to end the agent dialogue using the mute button. The audio channel between the user and the Wizard was established using Google Voice[3]. Google Voice also allowed us to record

[2] `https://aws.amazon.com/polly/`

[3] `https://voice.google.com/`

the entire conversation between the user and the Wizard. The room had to be quiet to allow the Wizard to concentrate on the search task (which was cognitively challenging in nature). The microphone was turned off at all times when it was not transmitting. This was done to eliminate any environmental noise, which, if transmitted to the user, would have revealed the true nature of the intermediary. The computer was running Windows 10 operating system. Although multiple browsers were available, the Wizard used the Google Chrome browser (Version 80.0.3987.122 64-bit) throughout the study.

4.6.2 Test room for User/Searcher

The test room for the searcher was also under laboratory settings to control confounding factors. While we did not use any unorthodox hardware, the room contained only the mock system. Before the start of the study, the researcher created a ruse by presenting a background for the study (Appendix Table A.2, points 7 and 8). This was done to convince the user that he was interacting with an artificially intelligent system and not a human intermediary. The researcher elaborated on the ruse and explained that an advanced prototype was running on the server and was connected to the wireless headphone. We also explained that any voice-based hardware devices (like Amazon Alexa or Google Home) were absent as the software was in a research and development stage. The wireless Bluetooth headphone transmitted the user's commands to the Wizard and the entire conversation was recorded using the record functionality of Google Voice (see Table 4.3). We also kept a backup audio recorder in the room for any unexpected scenario.

4.7 Variables

In any user study, the number of variables should be limited in order to investigate the research questions properly. The experimental variables could be divided into three different categories:

1. Control Variables

Table 4.3: Components of the Mock System

Component	Description of the Hardware/Software Used
User Interface	
Headphones	Used by the user for voice commands and listening to the responses from the Wizard. The user believed that the headphone was connected to one of the computers in the room. The noise cancellation properties made sure that the user was not distracted by any surrounding noise.
Phone (Hidden)	Used to set up an audio call over Google Voice.
Wizard Interface	
Computer	Used by the Wizard to search online. The computer had Windows 10 OS installed. The Wizard used the Chrome browser (Version 80.0.3987.122 64-bit) for searching online.
Text-to-Speech Software	A web-based instance of Amazon Polly, a state-of-the-art text-to-speech software, was used to translate the Wizard's typed-in text to voice.
Speaker	Played the voice (dialogues and commands) of the user.
Microphone	The voice output from Amazon Polly was transmitted to the user using the microphone.
Others	
Communication Channel	Audio call over Google Voice. Google Voice provided the call recording facility, which allowed us to record the conversation between the user and the Wizard.
Text-to-Speech Software	Converted the typed-in responses from the Wizard into speech. Allowed different voices and accents. We used the default (Joanna) voice.
Online Diary	Contained the dialogue scripts which could be accessed by the Wizard.
Kaltura	Screen recording tool which was used to capture the search actions of the Wizard.

Control variables affect the outcome or response of the experiment. In any experimental setting, there could be various predictors that influence the outcome variable. Therefore, the researcher needs to reduce the variance by fixing all the predictors except the control variable. The control variable is set by the researcher for different experimental settings. In our study, the *System Used* was the control variable.

2. Response Variables

   Response variables help in measuring the outcome of the experiment. These variables are directly related to the research questions. For the first research question, we were evaluating the user's search experience using the experimental and the control systems. Therefore, the response variables were the eight system-level differentials (Q4-11) in the post-task questionnaire which collected user feedback on the system experience.

3. Confounding Variables

   These variables are noise and do not belong to either control or response variables. These variables (all predictors except control variables) influence the outcomes but are not planned while designing the experiment. The experiment should be designed to reduce the confounding variables as much as possible. Two examples of confounding variables in our study are *Search Task* and *Task Order*.

Fraser and Gilbert (1991) performed an in-depth analysis of the different variables involved in a Wizard-of-Oz study. In Table 4.4, we have provided a detailed description of the variables involved in our study.

4.8 Recruitment and Roles

The recruitment of the participants (both the user and the intermediary) was a critical aspect of the study. The user should be a representative sample of the target population (for the system), while the intermediary should be selected to closely emulate an ideal version of a future voice-based conversational search agent. Therefore, it was important

Table 4.4: Experimental Variables and Values

Variable	Definition	Values
Scope	The amount of simulation involved in the experiment: a full system or any specific component of the system	Full system with a focus on system response (with or without explicit clarification)
Task	The range of tasks which are simulated in the study	Search tasks which are moderately complex
Searcher parameters	The details of the participants who play the role of the searcher: it could include different demographic features, search expertise, previous experience with voice-based search agents, and domain expertise. Some personal factors like user personality, patience, affinity towards technology could act as confounding factors.	Gender: Male/Female
		Search Experience: Users were required to have prior experience in browser-based searching (Self-reported, not verified or enforced)
		Search Skill: Not set
		Experience with voice-based personal assistants: Preferred (Self-reported, not verified or enforced)
		Domain Knowledge: Not Set
		Knowledge of the Wizard: No
Wizard parameters	Different parameters relevant to the person playing the role of Wizard. Once again, personal traits like personality could be confounding factors.	Amount of Training: 2 search sessions with the researcher, followed by 2 pilot studies
		Knowledge of the Task: Yes
		Topic Knowledge: Yes
		Level of experience with the search tasks: Expert
Communication Channel	The modes of communication from the Wizard to user and from user to Wizard.	Wizard to User: Text converted to Audio User to Wizard: Audio

to decide on the different criteria that governed the recruitment of the participants during the planning stage of the study.

4.8.1 Users/Searchers

As voice-based search agents are becoming ubiquitous, the future system should be used by people of all age groups and professions. This was an advantage as we did not have to identify any particular group of users who are more likely to use such a system. Although the age of the participants was not an influencing factor in our study, we aimed to keep the variance low by targeting students from a public university with similar search skills and experience. We circulated the recruitment letter throughout campus (See Table A.1 for details), over university electronic mailing lists, and on online forums. In the recruitment letter, we mentioned our preference for participants who were: (i) fluent speakers and listeners of North American English, (ii) proficient in using the internet and search functions, (iii) familiar with voice-based conversational search systems. The preferred language skills and search expertise of the participants were advertised but not assessed by the researcher. Instead, we relied on self-reporting by the participants. Although previous studies (White et al., 2006) have shown that experienced and inexperienced users perform search differently, we targeted the users who are familiar with the basics of search systems.

The recruited participants were asked to assume the role of the searcher. Based on the number of different experimental settings (one experimental and one control system in our case), we recruited N = 25 users for our study. Although our initial plan was to recruit 20 participants, we ended with 20% more participants to account for the outliers in the data. The number of users recruited was deemed to be sufficient to perform the relevant statistical tests with requisite statistical power. Participants were allowed to leave at any time, but to be compensated, they were required to either complete all the tasks or spend at least one hour in the laboratory engaging in the search task. The compensation amount was fixed at 20 USD.

4.8.2 Intermediary/Wizard

The role of the intermediary was most vital for the success of our WOZ study. Based on prior literature, the Wizard not only needs to be an experienced searcher who will be able to perform searches and provide results in real-time but also "a con man" (Price, Dahlstrom, Newton, & Zachary, 2002) to deceive the user into thinking that the intermediary is non-human. Modern search systems replicate the traditional role of librarians as a search intermediary. However, for our research question, the Wizard did not need to be an expert searcher (like a reference librarian). As we were exploring the role of system-level clarifications in the human-system information-seeking conversations, our Wizard needed to perform searches in real-time, follow the protocols, and closely emulate a search agent few years into the future. Also, the system-level clarifications were cognitively less complex and limited to explanations about the queries and information sources used by the system. Therefore, we hired a computer science undergraduate student who was proficient with searching online and had considerable experience in performing voice searches. For every user study session, the Wizard had to be present for the entire duration of the study. The Wizard was allowed a five-minute break after each task. The compensation of the Wizard was fixed at 20 USD for every user study session.

4.9 Wizard Training and Protocol

We were concerned about the task and topic learning effects on the Wizard over subsequent search sessions and users. This could have potentially influenced the interaction patterns and outcomes of the experiment. Therefore, to eliminate the influence of the learning effects, we trained the Wizard before the actual experimental sessions. This ensured that the skill and the behavior of the Wizard did not change during the course of the study. The Wizard was provided with the task description to search online and prepare notes about the search topic and possible search directions. The training familiarized the Wizard with the working of the experimental setup and improved his task- and topic-knowledge. In the pre-experimental stage, we also prepared a script that

contained predefined dialogues for various search situations. The script had templates to standardize the vocabulary and acted as a guideline for the Wizard to replicate the simple vocabulary common in artificially generated speeches. For example, every time the user started a search session, the Wizard responded with "Hi, I am Joanna, how may I help you today?". Similarly, search sessions ended with "It is always great talking to you, bye!" The details of the conversational script and guidelines to the intermediary have been provided in the following sections. These documents govern the behavior of the Wizard, more specifically, what he should and should not do.

We allowed the Wizard a few weeks to familiarize himself with the protocol, the scripts, and the tasks. Next, we conducted two search sessions with the Wizard to evaluate his performance when simulating the experimental and the control system. While the dialogue scripts contained the general structure of the system-level responses, the blank slots were filled up by the Wizard during the original conversation. The guidelines and conversational scripts were essential to minimize the influence of the human intermediary while at the same time increasing the response time and reducing the cognitive workload on the Wizard. Besides, it made the results more reproducible and accurate and reduced the variances and degrees of freedom within the WOZ experiment.

We conducted two pilot studies: one with a researcher who was familiar with the details of the study and the other with an end-user. All the components of the study – the experimental systems, the data collection devices, the conversational script, the guidelines for the intermediary, the tasks, and the questionnaires – were reviewed by peers and tested in a pilot study prior to the actual study. Based on the feedback received, we reevaluated some technical and design details and made minor adjustments to the script and the protocols. The pilot sessions also allowed the Wizard to get familiar with the search task and the mock system.

While we present the conversational script for the Wizard in the next subsection, the other guidelines and checklists for the Wizard can be found in Table A.3 and Table A.4.

4.9.1 Conversational Script for the Wizard

The conversational script of the Wizard is presented in this subsection. The script was generated through observation of searcher-intermediary interaction in differently publicly available datasets, SCS and MISC (Thomas et al., 2017), which allowed us to understand the different categories of dialogue and their functions in human-human conversation. After identifying the categories, we standardized the vocabulary of the dialogue by closely following the speech patterns used in existing state-of-the-art systems (*Write Out a Script with Conversational Turns*, n.d.). In this section, we show the different stages in conversation, the speaker, and the intended functional effect.

- Key phrases to start the conversation

  Function: Triggers the start of the conversation

  Speaker: User

  - Hi, Joanna
  - Hey, Joanna

- Greetings (Initial Response)

  Function: Initiates the user to state his information problem

  Speaker: Wizard

  - Hi, I am Joanna, how may I help you today?
  - Hi, I am Joanna, what kind of information are you looking for?

- Response to the user's information request

  Function: The Wizard can respond to the user's utterance in the following four ways: Positively, when he acknowledges the user's question; negatively, when he has trouble understanding the user or the question; request more information, or provide explicit clarification about the Wizard's model of the user.

  Speaker: Wizard

  - Positive response to the user's question
    - Okay, please let me look into it. Give me a few minutes.

        * Please give me a moment to search.
    - Negative response to the user's question
        * Information Request too long or complex
            · I am having a problem processing the request. Can you restate the question slowly?
        * Problem with hearing
            · I could not hear you properly. Can you please repeat it?
        * More information is required
            · Can you tell me more about what you are looking for?
            · Do you have any specific preferences?

- Presenting Results

  Function: The Wizard should summarize the top-3 results on SERP or performs disambiguation of the domain. He could also ask the user if he is interested in a specific result or may read from any document.

  Speaker: Wizard

    - SERP Summary
        * This is what I found: <summary of top-3 SERP>
    - Selecting Document
        * Do you want me to look into any of the results in particular? I can read from: <say the name of top-3 sites in a using their website name and extension e.g.: Quora.com for www.quora.com>
    - Read from Document
        * This is from: <name of website>: <some answer>

- Clarification of the user model before issuing the query

  Function: This applies to the experimental system, where the Wizard provides some simple clarifications to the user. The clarifications could include the Wizard's understanding of the user's information need, expressed by explicitly describing the query words or the information sources used. System clarifications

allow the user to understand the search actions performed by the Wizard.

Speaker: Wizard

  - Based on what you said, I am entering the query: <query used>
  - To clarify, you are looking for <information object>. Is that correct?
  - I am reading the information from the: <name of the website>. Is that alright?
  - Based on what you said, I understand that you are looking for <information object>. Am I correct?

- After the last search

  Function: To make sure that the conversation does not end abruptly. It also allows the user to ask more questions when in doubt.

  Speaker: Wizard

  - Did you find what you were looking for?
  - Is there anything else I can help you with?

- Repeat the utterance

  Function: The user requests the Wizard to repeat the last utterance. This could be due to several factors like a problems in hearing, the replay functionality, note-taking or cognitive reasons.

  Speaker: User

  - Joanna, can you repeat?

- End phrase to terminate search session

  Function: The user notifies the Wizard that he is ending the search session.

  Speaker: User

  - Bye, Joanna

- Ending the conversation

  Function: Ends the search session.

  Speaker: Wizard

– It is always great talking to you. Bye.

4.10 Data Collection

Different forms of data lead credibility to the experimental results and allow for a more comprehensive evaluation in the later stages of data analysis. We collected four different types of raw data during our study: the users' background and demographic information, the spoken utterances by the user and the intermediary, the survey responses by the user before and after the tasks to assess the search experience, the search actions of the Wizard (the screen recordings), and the exit interview.

4.10.1 Pre-test Questionnaire

Before starting the search session, the registered participants were given an overall description of the objective and aim of this research, then asked to sign the online consent form and to fill out a questionnaire on basic demographic and background information. The questions inquired about the basic profile (i.e., age, gender, and academic backgrounds), self-reported language proficiency, self-reported web search experience, search frequency, and search skills. The participants were also asked to report their frequency of use of intelligent personal assistants (Siri, Cortana, Amazon Alexa). For all the three self-reported items, we used a 5-point Likert scale, where 1=novice and 5=expert. The user filled out the questionnaire on a paper form before entering the test room. The exact questions can be found in Table B.1.

4.10.2 Pre-Task Questionnaire

Before each task, the participants were asked to answer a set of questions about the task topic, knowledge of the topic, topic interest, and perceived difficulty of the task. In Table B.2, we present the list of questions present in the pre-task questionnaire.

4.10.3 During Task

During each of the search tasks, we recorded the search activities undertaken by the Wizard and the conversation between the user and the Wizard. Using a screen capture tool, Kaltura, we recorded the search activities of the Wizard. Recording the Wizard's screen and search activities provided us with the details of queries input in response to the user dialogues, the pages visited, the amount of time spent on each page, and the results presented back. The audio channel was captured using Google Voice, which recorded the interaction – the information-seeking dialogues - between the user and the Wizard.

After each task, we requested the participants (users) to write the answer to the search task provided. The response – a few lines about what the user was looking for, what he found, and his selection of perfume or conference – helped the researcher assess if they were able to complete the task successfully.

4.10.4 Post-Task Questionnaire

The participants also answered questions about their search experience. We used a set of questions derived from the User Engagement Scale (O'Brien & Toms, 2010) to measure how the user engagement varied with the two systems. Our questions, measured on a five-point Likert-scale, captured the details on novelty, involvement, and feelings of reward, success, and engagement (O'Brien & Toms, 2010; Thomas et al., 2017). These questions are relevant to the use of new technology and were worded as in Table B.3: We reverse-coded one question to make sure that we can identify if the participants select random answers.

4.10.5 Exit Interview

At the end of the study, the participants were asked to attend a brief in-person semi-structured interview, where they answered a few open-ended questions about their overall experiences. The questions listed in Table B.4, assessed the overall experience of the users with the two prototypical systems and the tasks.

4.11 Implications of the User Study

Our study served three purposes:

1. The user study helped us answer our first research question, which explored the role of system-level clarifications on the search experience of the user. The questionnaire data allowed us to assess how explicit system-level clarifications influenced the user's search experience when interacting with a conversational search system.

2. We created a new dataset called CONVEX (CONVersation with EXplanations) using the user study data. We collected richer data, which included both user-agent information-seeking dialogues and the search actions undertaken by the intermediary. This dataset was used to build our classification model to predict the speech acts and search actions (to answer RQ2).

3. The feedback received from the users at the end of the study (as part of the semi-structured interview) allowed us to suggest useful features and make design recommendations.

4.12 Chapter Summary

To summarize this chapter, we explain our motivation behind the choice of methodology. None of the publicly available datasets were suitable to answer our research questions, which warranted the collection and analysis of new data. We describe the details of the Wizard-of-Oz experimental setup: how the study was designed, test rooms were prepared, and the different variables involved. We recruited people through advertisements on social media and electronic mailing lists. Twenty-five participants were recruited to play the role of searcher while one participant played the role of the intermediary or Wizard. In our study, the user was never informed that the system on the other side was not an artificially intelligent agent but a human. To maintain the pretense, the Wizard was trained extensively on how to conduct himself and what protocols to follow. We developed an exhaustive set of guidelines and model script to guide the actions of

the intermediary, which have been elucidated in this chapter. We also discuss the study design and the development of search tasks, the experimental procedure, and the data collection mechanisms.

Chapter 5

Transcription and Thematic Analysis

This chapter presents the basic statistics of the user study data. It also explains the thematic analysis performed to identify the different categories of speech acts and search actions, and the process of annotation.

5.1 Motivation

Our second research question was:

> *How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?*

To answer the research question, we collected user-system interaction data (which we call CONVEX data) using the user study described in the previous chapter. However, that was the first step towards creating a gold standard data which could then be used to train and evaluate automatic prediction models. First, we perform a thematic analysis to develop a set of qualitative codes for speech acts and search actions. We have used previous literature to identify the initial themes and revised them in subsequent rounds to develop labels for speech acts and search actions. Next, we hired multiple annotators to label the utterances in the CONVEX dataset developed by us. To test the validity of our model, we also annotated the publicly available Spoken Conversational Search (SCS) dataset.

5.2 CONVersation with EXplanation (CONVEX) Dataset

The CONVEX dataset was created using the user study described in the previous chapter.

5.2.1 Participants and Demographic Information

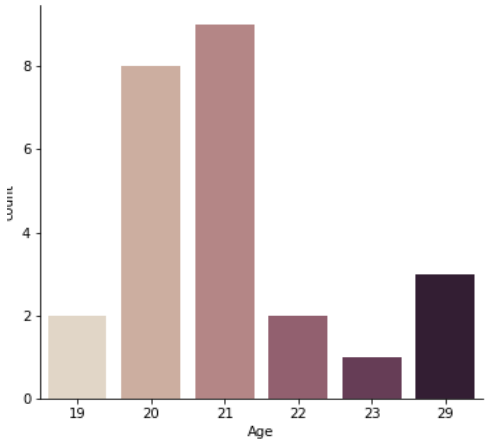
For our user study, we recruited a total of 26 participants (25 users and one intermediary). The user statistics can be found in Table 5.1. There were 20 females and six male participants, including the Wizard. The mean age of our participants was 21.64 with the maximum and minimum ages being 29 and 19 respectively (Median= 21.0, Variance= 8.15, Standard Deviation = 2.855). Twenty-two participants reported themselves to be native speakers of English while the remaining three identified Greek, Hindi, and Gujarati, respectively, as their first languages. The participants rated their English speaking and listening skills (which were essential for our study) and search skills on a 5-point Likert scale where 1=Novice and 5=Expert. Based on the self-reported scores, most of the participants considered themselves proficient in speaking and listening English (with means of 4.8 and 4.92 respectively). The online web search skill was high, with a mean of 4.6 and a median of 5. Almost all the participants had prior experience with voice-based personal assistants and identified their success rate to be between 1 and 5 (with a mean of 3.2 and a median of 3.0) when interacting with voice-based assistants.

5.2.2 Transcription

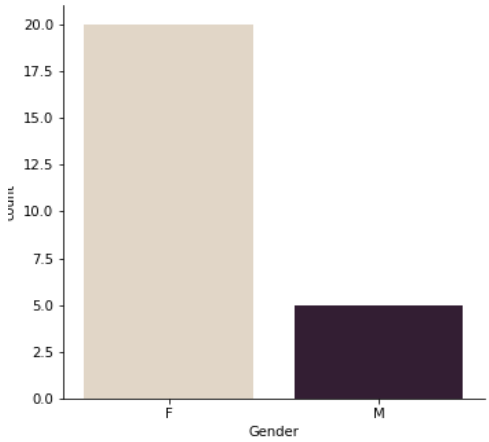
During our user study, we captured a total of 50 user-agent information-seeking conversations (two tasks for 25 user-agent pairs). Each search session was between 5 and 20 minutes, and each study led to audio and video recordings of length 20 to 60 minutes. Overall, we had more than 12 hours of audio and video to transcribe. For transcription, we followed the steps highlighted in previous works (McLellan, MacQueen, & Neidig, 2003; Thomas et al., 2017; Trippas et al., 2017b) and made necessary changes as required based on our data. The steps involved could be enumerated as follows:

1. Audio and Video Processing

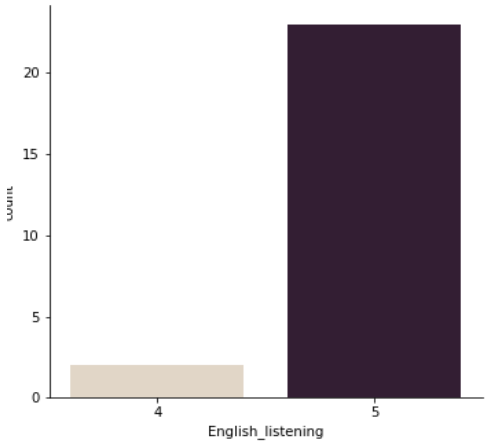
   We recorded the audio dialogues between the user and the Wizard using Google Voice. The video was captured using a screen capture tool, Kaltura. The video included the Wizard's dialogues while the audio files captured the dialogues for



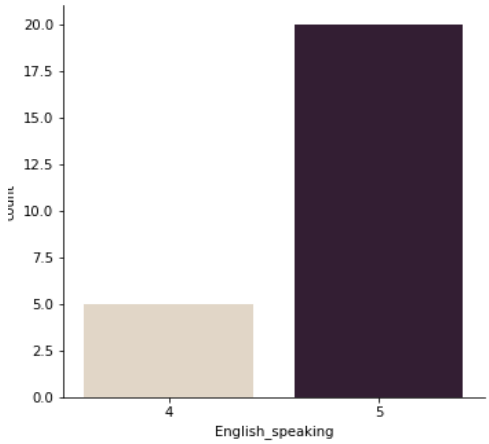
(a) Distribution of age.



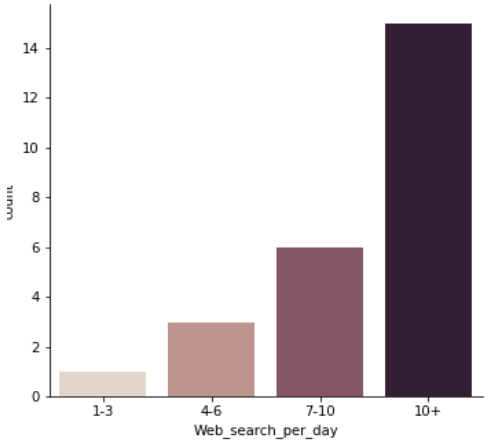
(b) Gender of participants.



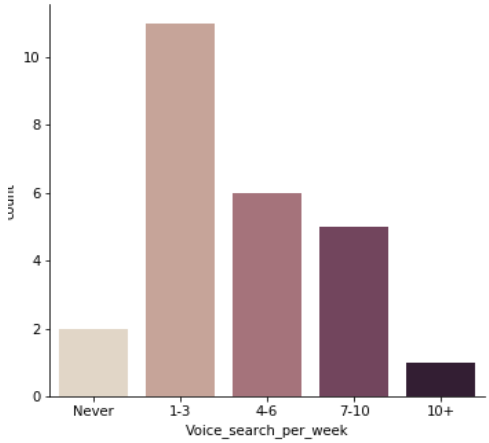
(c) English Listening Skills.



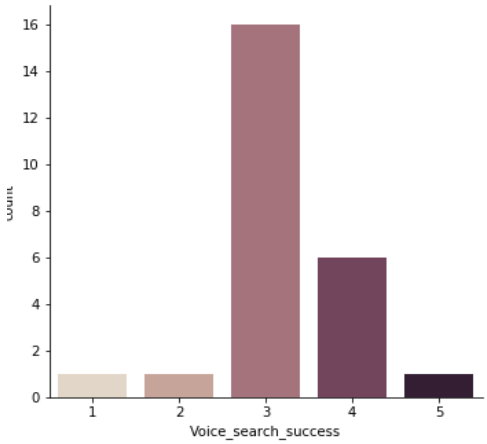
(d) English Speaking Skills.



(e) Web Searches per day



(f) Voice Searches per week.



(g) Voice Search Success

Figure 5.1: Demographic and Search Information

both. The first step in the transcription process was to synchronize the audio and video. We used an open-source video editing tool to obtain the synchronization (by overlaying the audio tracks using the first response from the Wizard). The recordings were trimmed to begin from the first user utterance and end with the last user/agent utterance.

2. Automatic Speech-to-Text Transcription

   After generating a high-quality audio file from the previous step, we used Amazon Cloud to store the audio files[1] and Amazon Transcribe[2] to automatically generate the text transcripts. The transcription also contained the timestamps and speaker identification.

3. Utterance Identification

   The utterances were identified based on two rules:

   (a) Any utterance must have a single speaker;

   (b) Two utterances should be separated by more than 10 seconds; and

   (c) A single utterance should be on a single topic. Any change in topic marks a new utterance.

[1]https://aws.amazon.com/s3/

[2]https://aws.amazon.com/transcribe/

We identified the utterances automatically and then manually inspected them for accuracy and topic identification.

4. Correct the automatically-generated text

   We came up with a basic set of rules for correction:

   (a) Preserve the originality of the transcription as possible, which means that the user- or agent- utterances should be as exact as possible. Do not correct the errors of either the user or agent, which could include grammatical errors or incorrect usage of words. Instead, focus the corrections on the imperfect text generated by the speech-to-text software.

   (b) Abbreviations should not be expanded.

   (c) Inaudible segments should be removed.

   (d) The text should be kept structured and consistent.

   (e) Transcription should not involve subject-matter experts but should be generalized and replicable.

5.2.3 Thematic Analysis and Annotation Schema

In an information-seeking dialogue, both the searcher and the intermediary take turns to speak. An utterance is a continuous and uninterrupted sequence of speech by one of the two participants. The end of an utterance is marked when the other participant starts speaking or when the current speaker changes the conversational topic. Interjections are ignored when annotating utterances.

To identify the popular themes in the data, we performed qualitative coding using the following steps:

1. Evaluating the correctness of the transcription and the utterance segmentation:

   This was important to ensure that there was no overlap between the utterances.

2. Using the existing frameworks to come up with the initial themes or labels:

   The initial categories for Speech Acts were borrowed largely from the Philosophy of Language and the frameworks created by early scholars in conversational IR

(Stein et al., 1999; Winograd et al., 1986). In the Conversational Roles (COR) model (Figure 3.2), the authors aim to represent all the possible scenarios in a human-human dialogue through the following speech acts: acceptance and refusal of conversational roles, the requests, promises, and offers, negotiations, and evaluation. For identifying and annotating the search actions, we have used the framework proposed by Azzopardi et al. (2018) and Trippas, Spina, Cavedon, and Sanderson (2017c) to obtain initial themes. During the first phase of our coding process, we came up with 15 speech acts and nine different search actions. The initial themes are presented in Tables 5.1 and 5.2.

3. Employing independent annotators to label the utterances:

   Two independent annotators annotated the dataset using the search and speech labels developed in the last stage. The inter-annotator agreement was around 68% and 65% for speech and search actions, respectively.

4. Modifying the themes:

   As the inter-annotator agreement was low, we performed a second round of thematic analysis and redefined the codes to resolve the ambiguities. Some of the finer labels were merged to create broader categories.

5. Re-annotating the data:

   In this last step, three independent annotators were asked to re-annotate the utterances, and we ended up with an inter-annotator agreement of 90.2%. The set of codes was finalized as the agreement among the annotators was fairly high.

The final thematic codes (along with their descriptions) are presented in the following subsections. It should be noted that search actions are performed only by the intermediary, while speech acts are possible for both the user and the intermediary.

Table 5.1: Speech Acts: Initial Codes

Speech Act	Description
Request	The user makes an information request to the intermediary.
Offer	The intermediary offers some results or suggestions to the user.
Agent_Accept	The intermediary accepts the user's information request.
Agent_Reject	The intermediary rejects the information request by the user.
User_Accept	The user accepts the offer made by the intermediary.
User_Reject	The user rejects the offer made by the intermediary.
Withdraw	The user or the intermediary withdraws the request or offer.
Answer	The intermediary responds with the answer to user's question.
Inform	The user provides some contextual information about the search. This is not a question or information request.
Repeat	The user or the intermediary repeats the last utterance.
Clarify	The intermediary asks some clarifying questions to have better idea about the information needs of the user.
Instruct	The user instructs the intermediary on how to search.
Contented	The user states that he is satisfied.
Discontented	The user states that he is dissatisfied.

5.2.4 Themes for Speech Acts

A total of 16 broad themes were identified for Speech Acts. The coding scheme, along with the category descriptions and examples, is explained as follows:

- Speech Acts (S) – Categories and Examples

    1. Question or Seek

       Description: Includes the initial information request. It could also involve the situation when the user comes up with a new search request during the conversation.

        (a) Initial Information Request

            Example: Joanna I am looking for a men's perfume can you give me some options?

        (b) Information-seeking Questions

            Example: So, what's the price of Tom Ford Bergamot?

    2. Accept

       Description: The agent or user accepts the request of the conversational

Table 5.2: Search Actions: Initial Codes

Search Act	Description
Query Creation	The intermediary ceates the first query.
SERP scanning	The intermediary scans the search engine results page and provides an overall idea of what he found.
SERP Top Result	The intermediary explains what the top-result contains without opening the document.
Query Refinement	The intermediary refines the query by rewording it.
Query Shortening	The intermediary shortens the last query by removing words from it.
Query Expansion	The intermediary expands the last query by adding words to it.
Document scanning	The intermediary reads from inside the document.
Summary	The intermediary reads out a summary of results.
List	The intermediary provides a list of results.

partner. In the user-agent dialogues, such instances were identified by the presence of keywords such as 'Yes,' 'Ok' or any interjections, the meaning of which can be construed as an acceptance.

Example: Ok, please let me look into it. Give me a few minutes.

3. Reject

   Description: The user or agent can choose to reject the request of the conversational partner.

   Example: I will not be able to answer that question.

4. Counter

   Description: The agent knows the user's information need and suggests some changes in the query. The control moves from the user to the agent.

   Example: You would have to name a specific conference so I can check the deadline.

5. Offer

   Description: The agent knows the user's information need and offers to do something different from the user's request. The control moves from the user to the agent.

   Example: None of these are in Europe. Would you like me to query Top

conferences in A.I early 2020 in Europe?

6. Request to Simplify the Search Problem

   Description: The request by the user is too complex for the agent to process. The agent requests the user to simplify the problem.

   Example: hmm. That search has become too complex for me. Can we do it in steps?

7. Answer

   Speaker: Agent

   Description: The agent either informs the user of the result or answers the question asked. This act signals the transfer of control back to the user. It could either be the final answer to the user's problem or an intermediate step.

   (a) Reading from the Document

   Example: According to the sephora.com, Yves ST Laurent l'Homme Cologne Bleue is 116 U. S. Dollars and contains Bergamot, Marine accord and Cardamom scent.

   (b) Answering Follow-up Questions

   Example: Yes, the $8.95 shipping fees for the 2 to 3 day express shipping.

8. Clarify

   Description: The agent seeks clarifications from the user to get a better understanding of the user's information need.

   (a) Explicit Clarification of the User's Model

   Description: This action is specific to the experimental system. The agent clarifies its model of the user by describing the formulated query or the search action undertaken. This allows the user to suggest edits, restate information problems, or take control of the search situation.

   Example 1: Based on what you said, I'm running the query, Men's perfume Bergamot. Am I correct?

   Example 2: I have another article called Future of Artificial Intelligence

for 2020 You need to know from read.com. would you like me to read from that one?

(b) Follow-up Questions

Example: Can you tell me more about what you are looking for?

9. Inform

Description: The user provides some additional information related to the search, either as a clarification to the question being asked or to add context to the information problem.

(a) Declaring Preferences

Example: No, I want to know the price in US dollars

(b) Providing Additional Information

Example: Yes, and can you also put in as a gift for Christmas?

10. Evaluation

Description: The user could be contented or discontented.

(a) Satisfied with the Results

Example: Yes. I found what I was looking for. It was a great talk.

(b) Confirms Answer

Example: That's the one.

11. Instruct

Description: The user directly instructs the agent on how to perform the search. This could be done by defining keywords, queries, information sources etc. At this point, the control of the search is with the user instead of the agent.

(a) Suggest Alternate Source

Example: Can you use another source?

(b) Suggest Query Reformulation

Example: No. Query, how many artificial intelligence conferences is [sic] there are in the United States?

(c) Suggest Search Strategies

Example: Now filter by date.

12. Repeat

Description: The agent may ask the user to restate the information request or the user may ask the agent to repeat the last utterance or answer. This also includes the utterance, which is repeated on request.

(a) Repeat Information Request

Example: I'm having a problem processing the request. Can you restate the question slowly?

(b) Repeat Answer

Example: Joanna, can you repeat?

13. Confirmation

Description: The user confirms when the agent asks for clarification or feedback.

(a) Confirm Query

Example 1: Yes. that is ok.

Example 2: Yes.

(b) Decline

Example: No, I don't.

14. Courtesy

Description: The user or the agent follows the norms of polite conversation by being deferential.

(a) Thanks

Example: Thank you, Joanna.

(b) Polite Expressions

Example 1: Is there anything else I can help you with?

Example 2: No, I'm good, Joanna.

(c) Asking for Other to be Patient

Example: Ok. Please let me look into it. Give me a few minutes.

15. Greetings and Closing Rituals

    Description: Includes the key phrases spoken by the user to activate and end the search session. It also includes the opening and closing dialogues of the agent, which signals the beginning and end of the conversation.

    (a) Greeting

        Example 1: Hi Joanna.

        Example 2: Hi, I am Joanna. How may I help you today?

    (b) Closing Ritual

        Example 1: Bye Joanna.

        Example 2: It is always great talking. Bye.

5.2.5 Themes for Search Actions

A total of four top-level themes were identified for Search Actions. Search actions are performed only by the agent (Wizard in our case). The coding scheme, along with the category descriptions and examples, is explained as follows:

- Search Actions – Categories and Examples

1. Query Creation or Refinement

   Description: The agent creates a new query or modifies an existing query for subsequent search.

   (a) Query Creation

       Example: Google search : “bergamot and lavender cologne”.

   (b) Query Update

       Example: Google search : “bergamot and lavender cologne under 150$” (Query updated to include “under 150$”)

2. SERP Scanning

   Description: The agent scans the search engine results page (SERP) and provides a summary of top results from SERP. It may include summary snippets or answers provided by the search engine at the top.

(a) Reading Summarized Answer from SERP

Example: Google featured snippet to the query "what is the weather in may in new jersey": According to Google, average weather in May in Atlantic City, New Jersey, United States daily high temperatures increased by nine degrees Fahrenheit from 65 degrees Fahrenheit to 74 degrees Fahrenheit. Rarely falling below 55 degrees Fahrenheit or exceeding 84 degrees Fahrenheit.

(b) Reading Factual Answer from SERP

Example: Google's rich answer to the query: "238 pounds to USD": Tom Ford Private Blend Venetian Bergamot is 306.52 United States dollars.

3. Document Scanning

Description: The intermediary reads from inside the documents returned by query.

(a) Reading from Document in SERP

Example: [Reading from sephora.com] According to the sephora.com Yves ST Laurent l'Homme Cologne bleue is 116 U. S. Dollars and contains Bergamot, Marine accord and Cardamom scent.

(b) Reading from a Previously Opened Document

Example: [Reading from `https://www.healthline.com/health/allergic-rhinitis#risk-factors`] You could use a dehumidifier or a high-efficiency particulate air filter. Eye drops and nasal sprays can also help relieve from itchiness.

4. Organizing Answer from Multiple Documents

Description: The intermediary reads from inside the documents returned by the query.

(a) Reading from Inside Multiple Documents + Summarizing Answer

Example: [Combining answers from `https://www.cigna.com/individuals-families/health-wellness/hw/medical-topics/allergic-rhinitis-hw33436` and `https://www.healthline.com/health/allergic-rhinitis#`

`risk-factors`] There is no permanent cure for allergic rhinitis. One of the best things you can do is to avoid the things that cause your allergies. You can take antihistamines to treat allergies. You can also use decongestants to relieve a stuffy nose and sinus pressure. Eye drops and nasal sprays can help relieve itchiness. Your doctor may recommend immunotherapy, or allergy shots if you have severe allergies.

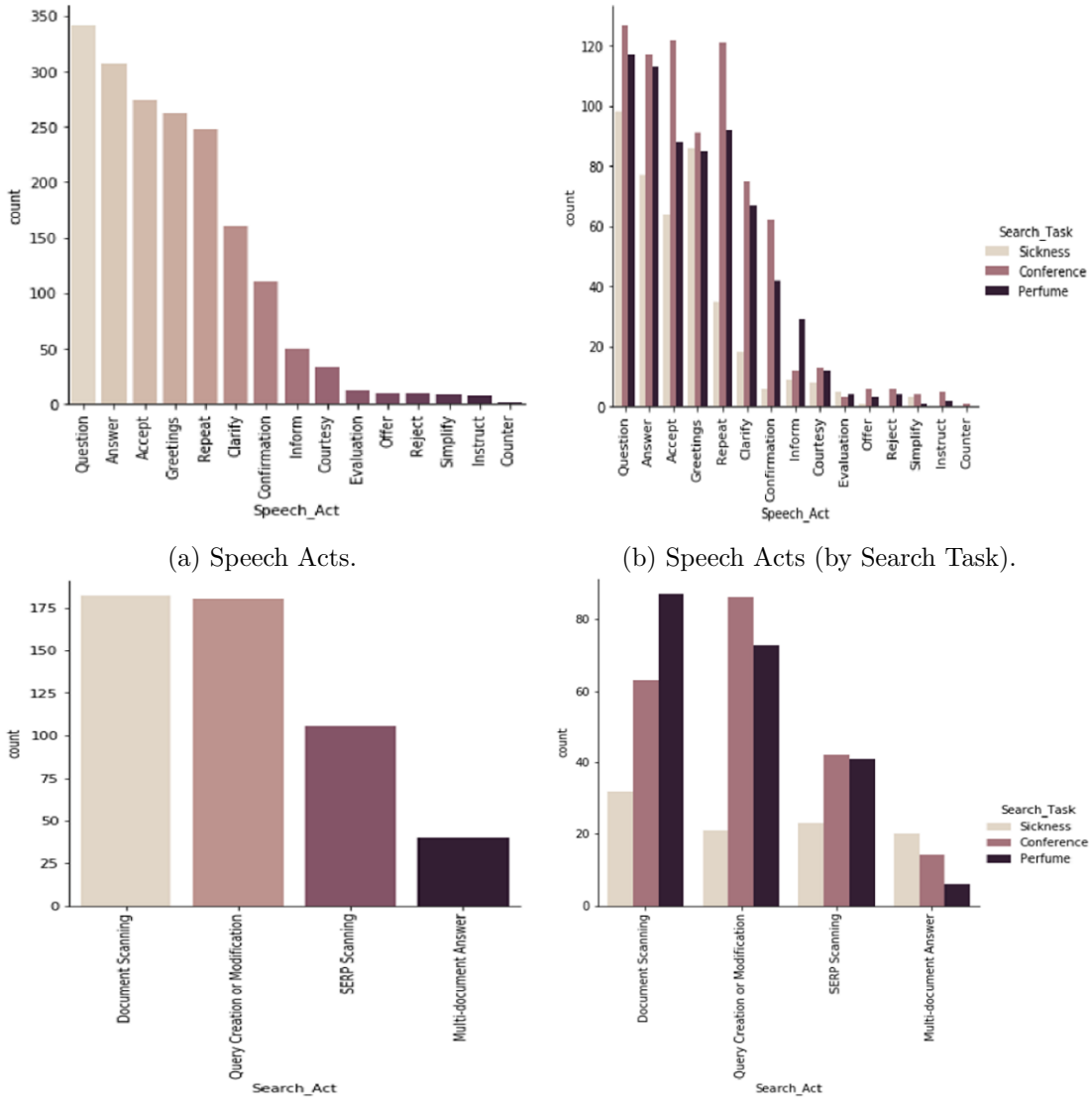
5.2.6 Statistics

The CONVEX data contained a total of 1834 utterances (with speech acts) from the user and the intermediary combined. The number of instances containing search actions was 509. Search actions were performed only by the intermediary. The distributions of the speech and search acts (overall and by search tasks) are shown in Figure 5.2. Figure 5.3 shows the number of utterances for each search task. Based on the visualizations, it can be assumed that the two experimental tasks had similar number of utterances, and were therefore, similar in task complexity. Five speech acts (Offer, Reject, Simplify, Instruct, and Counter) could be considered minority while 'Organizing Answer from Multiple Documents' was the only search act to have a low occurrence rate.

5.3 Spoken Conversational Search (SCS) Dataset

The publicly available SCS dataset (Trippas et al., 2018, 2017a, 2017b) was created through observation of participants engaged in information seeking over a spoken channel. This dataset was used to validate the performance of our predictive model. It contained a total of nine backstories and nine search tasks for simulated search sessions, adopted from the Taxonomy of Educational Objectives (Bloom et al., 1956; Krathwohl & Anderson, 2009). The user was provided with a backstory that was created using TREC topics: Q02, R03, and T04. The tasks belonged to three different levels of varying cognitive complexities – Remember, Understand, and Analysis levels of the taxonomy.

The conversations recorded were between the searcher and the intermediary (both



(a) Speech Acts.

(b) Speech Acts (by Search Task).

(c) Search Actions.

(d) Search Actions (by Search Tasks).

Figure 5.2: Statistics for CONVEX Dataset

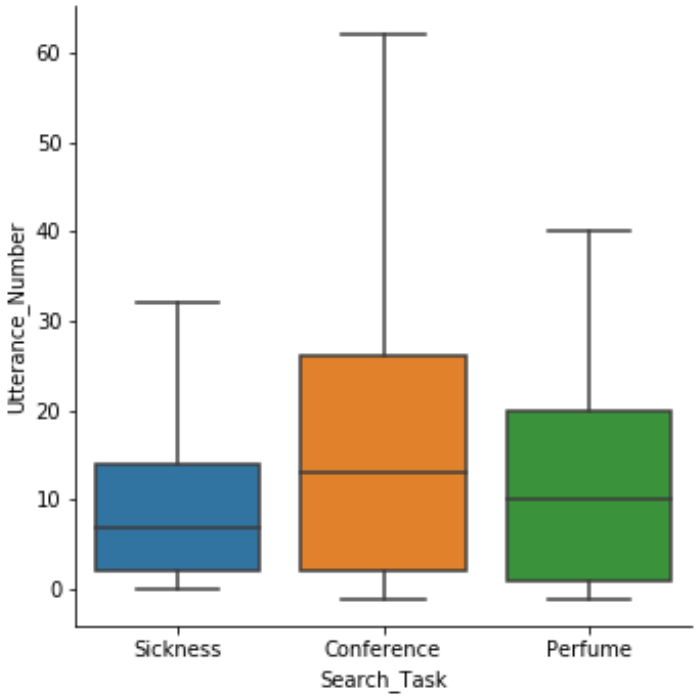


Figure 5.3: Frequency of utterances for search tasks.

roles played by the participants) in a laboratory-based user study with 13 pairs of participants located in the same room. Only the intermediary had access to a networked computer and had to rely on the searcher to explain the search task. There was a limit of ten-minute per task, and the searcher had to explicitly instruct the intermediary to stop the search and end the task. There were between six to sixty-nine turns in total, with one label per utterance. The dataset contained only the transcripts of the conversation with broad themes for the search actions performed by the intermediary. The researchers recruited participants to play the role of either the searcher or the intermediary. As the goal of the study was to observe the conversational patterns, none of the intermediaries were expert search agents. Instead, they were similar to the searchers (as is evident from the transcripts and recorded audio files) in search skills, knowledge of the database, and topic knowledge. Although the researchers report the different search actions performed thematically, they did not share the details of the search actions, like the pages visited, dwell time, click behavior, and so on. As such, in order to recode the search actions (using a different set of themes), the search activities

need to be inferred using the dialogues and existing codes.

5.3.1 Transcription

The publicly available dataset was already transcribed. We made some minor corrections to the transcripts and utterance segmentation.

5.3.2 Thematic Analysis and Annotation

We used the same set of themes developed for the CONVEX dataset. Once again, two independent annotators were hired to annotate the dataset with speech act and search action labels. The final inter-annotator agreement was above 90%.

5.3.3 Statistics

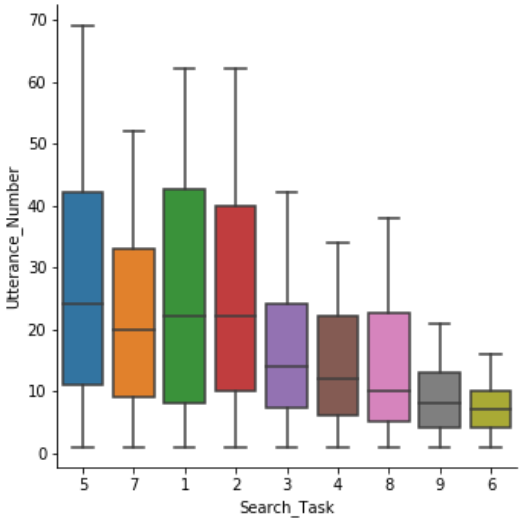
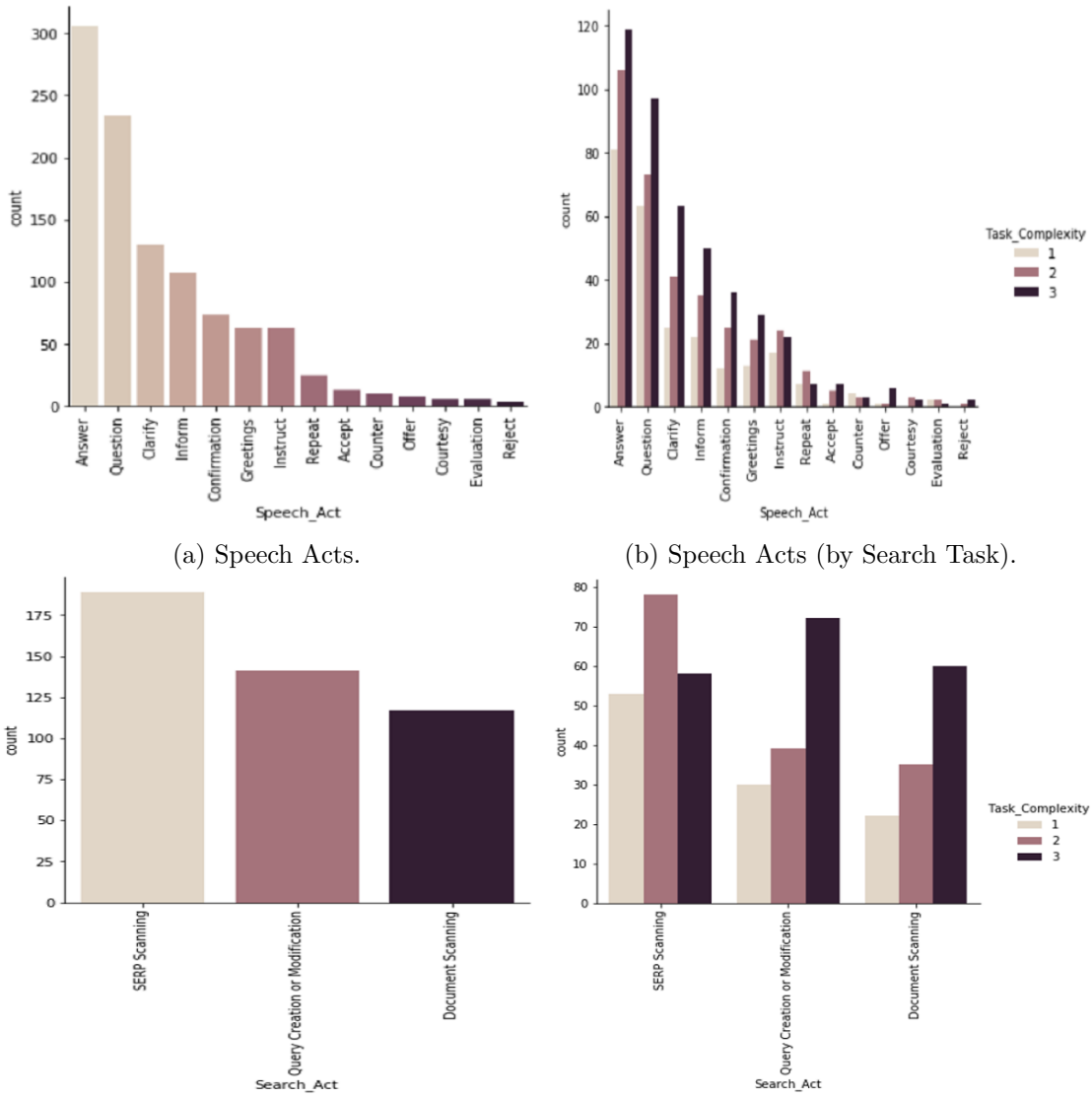


Figure 5.4: Frequency of utterances by search tasks.

The dataset contained information-seeking dialogues from the user and the intermediary. There were a total of 1043 utterances from the user and the intermediary combined. The number of instances with search actions was 447. The distribution of the speech and search acts (overall and by search tasks) is shown in Figure 5.5. Figure 5.4 shows the number of utterances for each search task. We observed that the SCS dataset did not contain any instances of 'Organizing Answer from Multiple Documents.'

As the publicly shared data included only broad search categories, the human annotators did not find enough evidence to differentiate between search actions involving single or multiple documents. Six speech acts were a minority in SCS. There were no instances of 'Simplify,' while the frequencies of occurrence of 'Reject', 'Evaluation', 'Courtesy', 'Offer', and 'Counter' were very low.



(a) Speech Acts.

(b) Speech Acts (by Search Task).

(c) Search Actions.

(d) Search Actions (by Search Tasks).

Figure 5.5: Statistics for SCS Dataset

5.4 Chapter Summary

In this chapter, we describe the two datasets – CONVersation with EXplanation (CONVEX) and Spoken Conversational Search (SCS) – statistically and qualitatively. We describe the user demographics for both the datasets and help the users visualize the distribution of utterances, search actions, and speech acts in the two datasets. The images help us compare the tasks by complexity and number of utterances. We explain in detail the transcription process and the subsequent thematic analysis. We also share the two codebooks which we developed for classification of the speech acts and the search actions. The codebooks contain a description of the themes and examples to familiarize the readers with the different categories of speech acts and search actions. The readers can also develop an idea of the majority and minority classes involved in the prediction tasks.

Chapter 6

Development of the Deep Neural Classifier

In this chapter, we discuss the suggested features for classification, and the details of the proposed deep neural model, MDSC.

6.1 Motivation

Our second research question was:

> *How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?*

To answer the second research question, we collected user-agent interaction data, which contains the search activities performed by the agent in addition to the spoken dialogues. Next, we annotated each utterance in the dataset with the corresponding speech act and search actions. The final step in answering the question was to develop a prediction model that could automatically identify the speech acts and search actions. Based on prior literature (Cai et al., 2017; Y. Liu, Sun, Lin, & Wang, 2016; Serban, Sankar, et al., 2017; W. Wang et al., 2018; Wu, Schuster, et al., 2016), machine learning – deep learning in particular – have shown commendable performance when tackling problems related to natural language, speech, and text. Therefore, we developed a multi-channel deep neural model for the prediction tasks, the details of which are presented in the following sections.

6.2 Prediction Task

Information seeking dialogues are not chit-chats but goal- or task-oriented conversations. Therefore, detecting speech patterns in information-seeking dialogues is a subset

of the original problem: open-domain machine conversation. Arguably, it is easier to generate speech in a task-oriented environment. However, in a search session, the search actions of the agent are as important as the spoken responses. Therefore, natural language understanding in conversational search agents involves connecting the speech component of user utterances to the system search action. In order to do so, we have used a modular approach which could be extended to any human-human or human-computer information-seeking dialogue. First, for each utterance, we identify the speech act it belongs to. In the next step, we use the speech act and other features to identify the appropriate search activity.

6.3 Features and Channels

Bi-directional models have proven effective in learning patterns and context in natural language and have been used extensively for natural language inference (Y. Liu et al., 2016), chatbots (YIN, 2019), and dialog systems (Khatri, Goel, et al., 2018; Peng, Fang, Xie, & Zhou, 2019). Prior research has also explored the use of different categories of features in conversational search systems (W. Wang et al., 2018). Therefore, to answer our research question, we have developed a simple, modular bi-directional LSTM model, MDSC, with three channels. For each of the three channels, we have used a different category of data. The input representations of each channel, x_i, are fed as input to the neural network. The feature representation function ψ maps each utterance into a vector of features:

1. A sparse representation of natural language features (both lexical and syntactic)
2. A dense vector representation of each utterance using word embeddings; and
3. A dense representation of the different dialogue metadata features.

In the absence of any of the channels (which could be due to missing data), the model could still be used and implemented with reasonable accuracy.

6.3.1 Channel 1: Lexical and Syntactic Features

The lexical and semantic features were generated from the utterances using the SpaCy API[1] and were subsequently processed and extracted. SpaCy is a high-speed, industrial-strength natural language processing library that works remarkably well in generating natural language processing features. Use of the SpaCy library has been reported in prior work on chatbots (Altinok, 2018; Khatri, Hedayatnia, et al., 2018) and natural language understanding systems (Bocklisch, Faulkner, Pawlowski, & Nichol, 2017). The different natural language features extracted at word- (or token-) level were:

1. Named Entity Type

   If a given token is a named entity, then this feature identifies the category of the named entity (which could be people, nationalities, companies, organizations, places, etc.)

2. IOB Code

   The IOB code of the named entity tag helps in identifying the position of the token in the named entity. If the token begins an entity, the code is B. Similarly, the IOB Code is I or O if the token is inside or outside an entity. If no entity tag is set, the code is kept blank.

3. Orthographic Features

   We identify the shape of the token by transforming the token string to get the word shape – capitalization, punctuation, digits. However, we consider only the first three characters to reduce the number of dimensions in later steps.

4. Alphabet

   A Boolean variable to identify if the token contains alphabetic characters.

5. Digit

   A Boolean variable to identify if the token contains numeric digits.

[1] https://spacy.io/

6. Punctuation

   A Boolean variable to identify if the token is a punctuation symbol.

7. Out-of-vocabulary Word

   A Boolean variable to identify if the token does not exist in the vocabulary. This feature could be important in identifying utterances containing wrong or misspelled words.

8. Stopword

   A Boolean variable to identify if the token is a frequently occurring word (or a stopword).

9. URL

   A Boolean variable to identify if the token is an URL.

10. Coarse-Grained Part-of-Speech

    The basic (or top-level) part-of-speech tags for the English language, which would include nouns, pronouns, verbs, adverbs, adjectives, conjunctions, etc. without classifying them into subcategories.

11. Fine-Grained Part-of-Speech

    The detailed information on the coarse-grained part-of-speech. For example, a verb (which is the base form) can be further categorized as VB Verb, VBD (Verb, past tense), VBG (Verb, gerund or present participle), VBN (Verb, past participle), VBP (Verb, no third-person singular present) or VBZ (Verb, third-person singular present).

12. Dependency Relation

    A single word can be used in different contexts, and its meaning differs based on the usage and relative positioning in the sentence. The relations in the dependency parse tree are often used for word sense disambiguation (Hale, 2003).

13. Character Offset

    Distance of the word from the beginning of the sentence. Each character counts as one unit in distance.

14. Sentiment of the token

    The sentiment of the words in the posted question could indicate an aggressive or abusive language. This feature could help in identifying the positivity or negativity associated with each utterance and could help in identifying the correct class.

15. Lexical Type

    Each word in the utterance can be grouped into different lexical types like a noun or a verb phrase.

For all the words in the utterance, we obtained the word-level one-hot representation of each feature and combined them hierarchically to obtain the sentence-level and then the utterance-level representations. For each utterance, we concatenated all the features to obtain a final vector of 194 dimensions (for CONVEX utterances) and 164 dimensions (for SCS utterances). For an utterance u, the feature extraction function ψ concatenates the respective vector representations of the natural language features using the merge function $\mathcal{M}$. For n words in the utterance, each feature $\mathcal{F}_j$ was represented by a vector $(v_1...v_k)$ of size k_{F_j}.

$$\psi(u) = [\mathcal{M}_{i=1}^{n}(\mathcal{F}_{ij}{}_{j=1}^{15})], \tag{6.1}$$

6.3.2 Channel 2: Word Embeddings

In any conversational dataset, the transcripts of the dialogue (or utterances) are often the most important (and also the easiest available) data. We have used the words in the utterances to generate word-level embeddings, which can identify the proximity of the words in the semantic space. To generate the word representations, we used the pre-trained GloVe[2] model with an utterance length of 100 words and a vocabulary of 25,000 words. The word embeddings help in capturing the rich linguistic context of the words (as each word is projected onto a 100-dimensional space based on their semantic

[2]https://nlp.stanford.edu/projects/glove/

proximity) (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

Each utterance (which contains a sequence of words) in the dataset was input to the embedding function $\mathcal{E}$, such that: $\mathcal{E} : \mathcal{V} \rightarrow \mathbb{R}^m$ (where $\mathcal{V}$ denotes the vocabulary set and m is the embedding dimension). For an utterance, u, which contained a total of n words (w), the feature extraction function ψ concatenates (||) the word embeddings of individual words (obtained using the embedding function) using the merge function $\mathcal{M}$.

$$\psi(u) = [\mathcal{M}_{i=1}^{n}(\mathcal{E}(w_i)], \tag{6.2}$$

However, word embeddings do not perform efficiently if used in isolation. Hence, we augmented our natural language features with word embeddings (each category implemented as a separate channel to ensure modularity).

6.3.3 Channel 3: Dialogue Metadata

Lastly, we combined all the dialogue metadata features which were available with the utterances. Although these features were specific to our dataset, it should not be difficult to derive them for any conversational data.

1. Utterance Number: The sequence number of the utterance in a given information-seeking conversation. For example, in the SCS data, the number of utterances in a conversation varied from 2 to 69;

2. Duration of the utterance: Length of the utterance in seconds;

3. If the speaker is intermediary;

4. System: The system which we were using (this was applicable only for the CONVEX dataset as it had two systems. SCS dataset had a single system)

5. Complexity of the task: In the SCS dataset, the task complexity was 1 (Remember), 2 (Understand), and 3 (Analyze). In CONVEX, the task was of either level

0 (low complexity for the warm-up task) or 1 (moderate complexity for experimental tasks).

6. Previous Speech Act: The speech act of the previous utterance.

7. Previous Search Act: The last search action performed by the intermediary.

8. Previous User Speech Act: The last speech act by the user.

For an utterance u, which contained the metadata features mentioned above, the feature extraction function ψ concatenates (||) the respective numeric, binary, or one-hot representations.

$$\psi(u) = [\mathcal{F}_{i_{i=1}}^{8}], \tag{6.3}$$

6.4 Output Classes for Speech Acts and Search Actions

We have used the final set of themes developed in Chapter 5 to label the output classes for our utterances.

As the number of minority classes were high for both CONVEX and SCS data, we merged some of the speech acts for the output class labels. The theme Reject was merged with Accept as they were both initial system responses. Similarly, Counter, Offer, and Simplify were merged as they were suggestions given to the user by the agent. The final class labels for speech act prediction are presented in the Table 6.1.

Table 6.1: Class Labels for Speech Acts

Class	Speech Acts	Description
S1	Question or Seek	Search Request by the User.
S2	Accept or Reject	The response of the agent to user's information request. Combines the minority theme Reject with Accept.
S3	Counter or Offer	The agent suggests query modifications or offers alternative search strategy. Combines minority themes Counter, Offer, and Request for Simplification.
S4	Answer	The agent responds with the answer to user's question.
S5	Clarify	The agent seeks clarification from the user.
S6	Inform or Declare	The user provides more information to the agent.
S7	Evaluation	The user could be contented or disconted.
S8	Instruct	The user instructs the agent on how to search.
S9	Repeat	The user or agent repeats last utterance
S10	Confirmation	The user confirms or rejects agent's action.
S11	Courtesy	The user or agent follow norms of polite conversation
S12	Greetings and Closing Rituals	Key phrases to start or end the search session.

For Speech Actions, all the four themes were used as output classes. SCS did not contain any instances of SR4. The output labels for search action prediction task is presented in Table 6.2.

Table 6.2: Output Labels for Search Actions

Class Label	Search Actions
SR1	Query Creation or Refinement
SR2	SERP Scanning
SR3	Document Scanning
SR4	Organizing Answer from Multiple Documents

6.5 MDSC Model Architecture and Implementation Details

For our model, we used a bidirectional RNN with Long short-term memory (LSTM) cells (Hochreiter & Schmidhuber, 1997) to encode the words and the context into a vector representation. Although some of the recent work in natural language understanding (Ahmadvand, Choi, & Agichtein, 2019) have used context explicitly, we used

the bidirectional LSTMs, which have shown significant improvements in performance when contextual information is involved. While regular LSTM cells in a feedforward network process text from left to right, bidirectional LSTMs (Schuster & Paliwal, 1997) analyze the text in both directions, from left to right and from right to left. As the amount of text increases, with the sequence of texts before and after, the system is able to recognize the patterns more accurately. Next, we had a max-pooling layer followed by a dense layer, with the rectified linear unit (ReLU) as the activation function and l1 and l2 regularizers. The last layer, which outputs the n-dimensional prediction vector, is another dense layer with softmax activation function (where the value of n is 12 and 3 or 4 while classifying speech acts and search actions respectively). All the hyperparameters were determined based on previous literature and experimental fine-tuning.

Given an utterance u, where the input sequence $x = (x_1, ..., x_{100})$, and the corresponding embedding $e = (e_1, ..., e_{100})$, our bidirectional recurrent neural network (RNN) computed the forward hidden vector sequence $\overrightarrow{h} = (\overrightarrow{h}_1, ..., \overrightarrow{h}_{100})$, backward hidden sequence $\overleftarrow{h} = (\overleftarrow{h}_1, ..., \overleftarrow{h}_{100})$, and the output vector sequence $y = (y_1, ..., y_{100})$ by iterating the forward layer from $t = 1$ to 100 and the backward layer from $t = 100$ to 1.

$$\overrightarrow{h_t} = \mathcal{H}(W_{e\overrightarrow{h}} e_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \tag{6.4}$$

$$\overleftarrow{h_t} = \mathcal{H}(W_{e\overleftarrow{h}} e_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{6.5}$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{6.6}$$

6.5.1 Dropout

Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) attempts to regularize the model so that it could learn diverse parameters. By masking specific parameters in the hidden units, it forces the model to learn more efficiently, using

different patterns every time. The bidirectional LSTM layer had a dropout of 0.25 and a recurrent dropout of 0.1.

6.5.2 Activation Function

In our model, we have used two different activation functions in different layers of our model: Rectified Linear Units (ReLU) and Sigmoid. ReLU was used in the hidden layers of the neural network and was replaced by sigmoid functions in the output layers.

6.5.3 Optimization Function

Optimization algorithms used in training deep learning models are different from the traditional optimization algorithms (Goodfellow, Bengio, & Courville, 2016). These algorithms do not directly influence the performance measure P. Instead, they aim at reducing a cost function J, which is expected to reduce P. Schaul, Antonoglou, and Silver (2013) compared different optimization algorithms in their paper by working on different learning rates to set hyperparameters. Adam – an adaptive optimizer – which uses mini-batches to adjust the learning rates of model parameters automatically was the preferred optimization algorithm. It is more robust in the choice of hyperparameters and considers estimates of both the first and second-order moments for bias correction (Goodfellow et al., 2016).

6.5.4 Loss Function

As we were dealing with a multi-class prediction for answering our research question – 12 output classes for speech acts and 4 for search actions – therefore, we have used categorical cross-entropy as the loss function for our models. For every instance i, and every possible class j, cross-entropy is calculated as:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} log(p_{ij}) \tag{6.7}$$

MDSC model using the three channels individually are shown in Figure 6.1, Figures 6.2, and Figure 6.3. For ablation analysis, the output of all the three channels was

combined together (as shown in Figure 6.4) and in pairs and input to another dense layer with softmax activation. The batch size was 32 for both search act prediction tasks (CONVEX and SCS). For speech act prediction, the batch size was 32 when analyzing CONVEX and 64 for SCS. The number of epochs was fixed at 300. We repeated our experiment for 30 times, working with different training and test data collections (randomized and selected using different seed values).

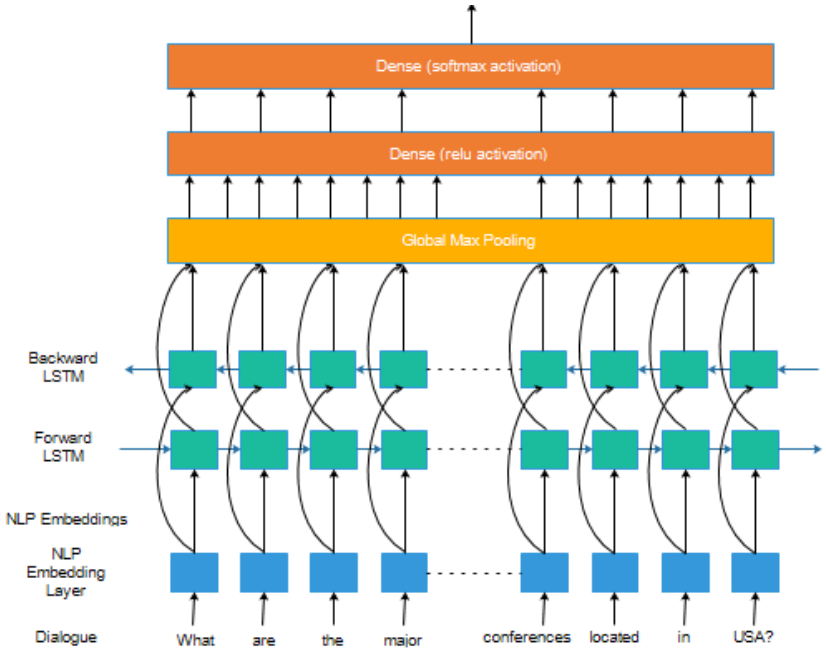


Figure 6.1: Bi-LSTM using Channel 1 (NLP Features).

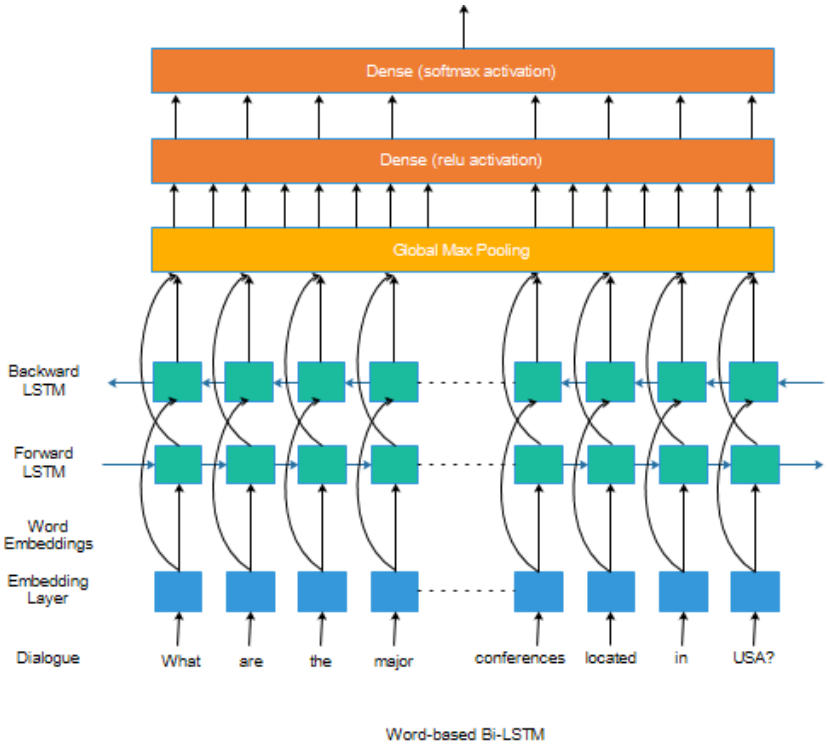


Figure 6.2: Bi-LSTM Model using Chanel 2 (Word-embeddings)

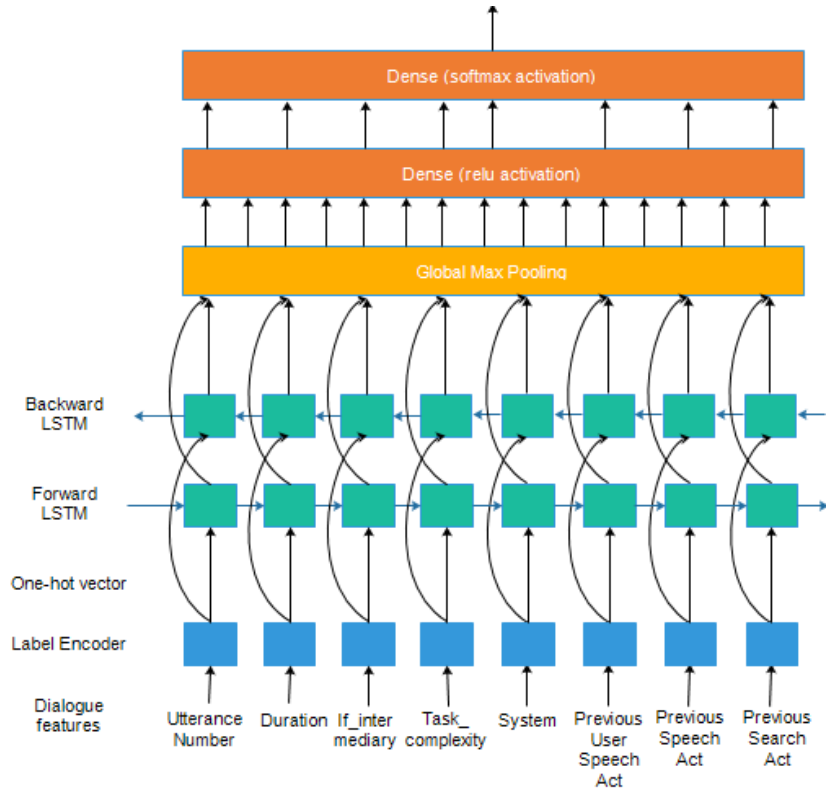


Figure 6.3: Bi-LSTM: Dialgue Metadata.

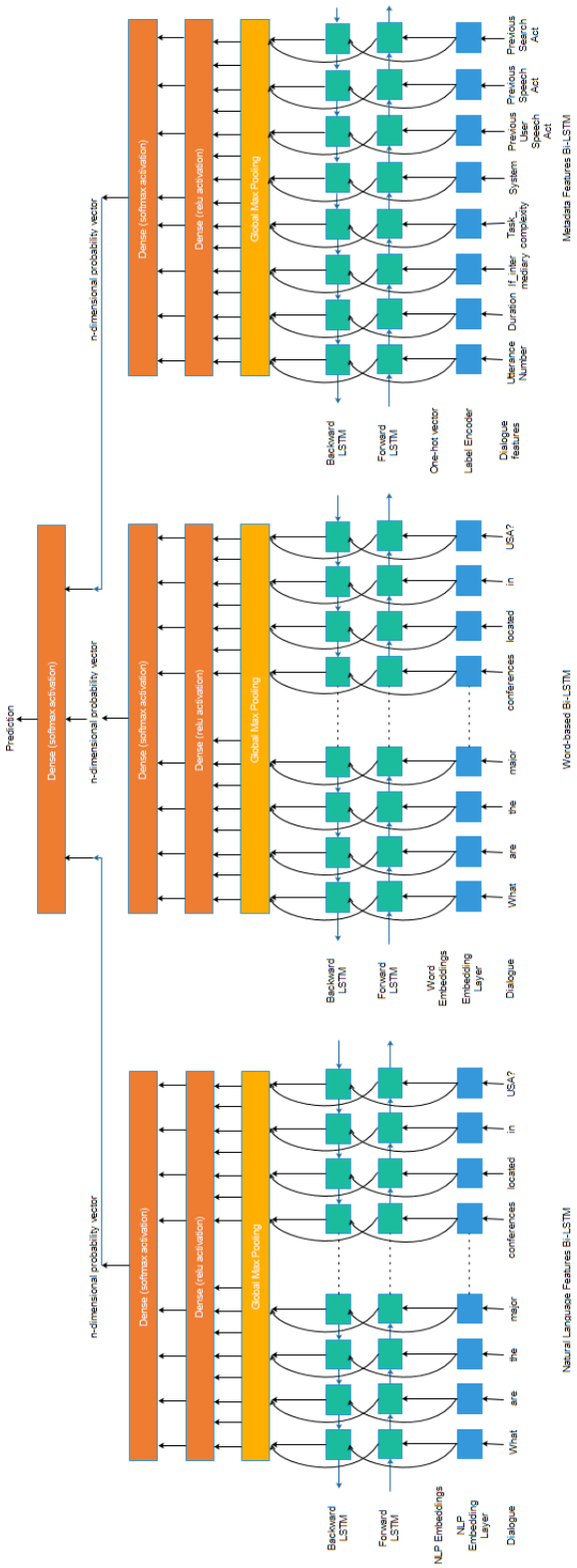


Figure 6.4: MDSC with Three Channels.

6.6 Chapter Summary

In this chapter, we have proposed MDSC, which is a multi-channel deep speech classification model for identifying the speech acts and the search actions in conversational search systems. Our classification model is simple but effective and incorporates different groups of features – word embeddings, lexical and syntactic features, and dialogue metadata – for the two prediction tasks. We provide a detailed description of how the features were generated and implemented on different channels. We also explain the model architecture and the implementation details, which include the hyperparameters used. The proposed model is modular and could be extended to most user study data which contains transcripts of searcher-intermediary conversations.

Chapter 7

Exploring the Role of Clarifications in User-Agent Information-seeking Dialogues

In this chapter, we present the results of our user study (described in Chapter 4) and connect them to the first research question:

> *For moderately complex tasks, can we determine the influence of explicit system-level clarification on the user's search experience?*

Our user study tests how explicit system-level clarifications (or explanation of the system's understanding of the user's problematic situation) influence the search experience of the user. Twenty-five participants took part in the experimental within-subjects design and performed three search tasks (the first a warm-up task, followed by one task each on experimental and control systems). In the following sections, we analyze statistically the feedback collected from the participants through questionnaires and interviews.

7.1 Descriptive Statistics

The data collected as part of the user study are based on the user responses to pre- and post-task questionnaires. The findings are further corroborated using comments made by the user during the exit interview. There were a total of 5 pre-task questions and 13 post-task questions, with all the questions being answered using 5-point Likert scales. During the user study, we reversed the scale for some of the questions randomly. This ensured that the participant read the questions carefully before marking the response. Once we finished with data collection, the responses were aggregated with a constant descriptive scale. The pre-task questionnaire (Table B.2) contained two questions on

topic knowledge and three on task complexity. Level 1 on the Likert scale represented the lowest value of the differential (for example, no knowledge, no familiarity, or not at all difficult). The post-task questionnaire (Table B.3) contained 3 questions on task experience and 10 on system experience. Level 1 on the Likert-scale denote complete agreement to the question asked while level 5 represents complete disagreement.

The distribution of the data is visualized in Figure D.1 (Pre-task responses), Figure D.2 and Figure D.3 (Post-task responses). Although all the samples were independent (obtained from different participants), all the variables were ordinal. Hence, we made no assumptions about the normality of the data and performed the Shapiro-Wilk test and Kolmogorov-Smirnov test (with Lilliefors Significance Correction) to test the normality of the distribution. The outcome of the tests, as presented in Table 7.1, indicates that the assumption of normality was violated for most of the variables (as the p-value was less than 0.05 in almost all the cases). Therefore, for the rest of the chapter, all the analyses involve non-parametric tests.

Table 7.1: Normality Tests

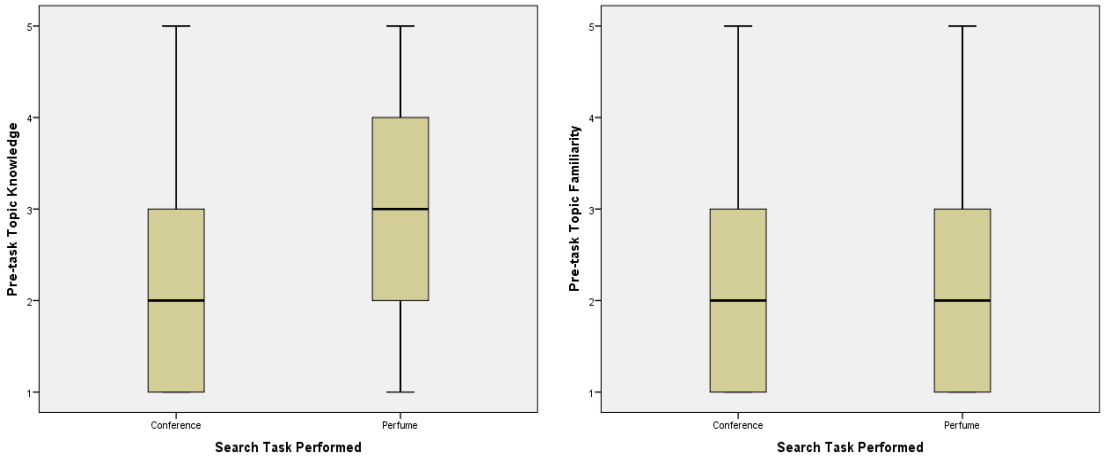
Tests of Normality					
Post-task Differentials	System	Kolmogorov-Sm. Statistic	Sig.	Shapiro-Wilk Statistic	Sig.
Task Complexity	No Clarification	.229	.002	.879	.007
	Clarification	.200	.011	.913	.035
Task Difficulty	No Clarification	.212	.005	.894	.014
	Clarification	.279	.000	.870	.004
Successful Completion of Task	No Clarification	.350	.000	.688	.000
	Clarification	.282	.000	.769	.000
Ease of using the System	No Clarification	.396	.000	.671	.000
	Clarification	.369	.000	.697	.000
Helpfulness of System	No Clarification	.372	.000	.626	.000
	Clarification	.391	.000	.679	.000
System Understood the Problem	No Clarification	.349	.000	.649	.000
	Clarification	.314	.000	.777	.000
System Communication Clear	No Clarification	.325	.000	.744	.000
	Clarification	.258	.000	.812	.000
System Response Adequate	No Clarification	.400	.000	.647	.000
	Clarification	.312	.000	.728	.000
Experience with System Satisfying	No Clarification	.420	.000	.593	.000
	Clarification	.289	.000	.759	.000
Would have done differently	No Clarification	.237	.001	.889	.011
	Clarification	.208	.007	.909	.029
System Found Right Information	No Clarification	.349	.000	.649	.000
	Clarification	.367	.000	.708	.000
Better Information was Available	No Clarification	.278	.000	.813	.000
	Clarification	.222	.003	.811	.000
Will Use the System Again	No Clarification	.434	.000	.610	.000
	Clarification	.321	.000	.753	.000

7.2 Before Task Perception of the Search Topic and Task

In this section, we discuss the user's perception of the search topic and task, as reported in the pre-task questionnaire. The responses were marked after the users read the search task description, and before they started the search session. The pre-task questions on the user's topic knowledge, topic familiarity, and task complexity were not directly linked to our research questions. However, it was important to assess if these factors (which were not predictors under experimental conditions) had any influence on the user's search experience. We also evaluated the effect of gender on all the pre-task differentials.

7.2.1 Topic Knowledge and Familiarity

Based on the self-reported responses, the users felt similar familiarity with both the search topics: Conference (Mean=2.04, SD=0.212) and Perfume (Mean=2.32, SD=0.236). However, the perceived knowledge in perfumes (Mean=3.04, SD=0.248) was more than that of Conference (Mean=2.36, SD=0.237). Although the variance in means for topic knowledge could be attributed to the gender imbalance in our participant population, we did not have enough evidence to make that claim. The effect of gender – main effect or interaction effect with system – on task knowledge and familiarity was not significant. Although no significant effects of gender were observed on topic knowledge, it would be interesting to check if that has any impact on the post-task assessments. Figure 7.1 shows the box-plot for the two differentials.

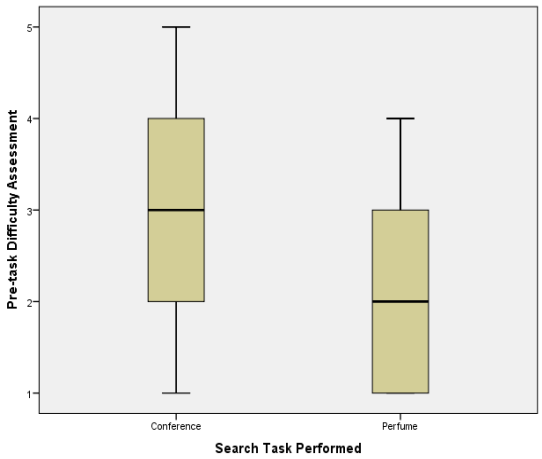


(a) Topic Knowledge. (b) Topic Familiarity.

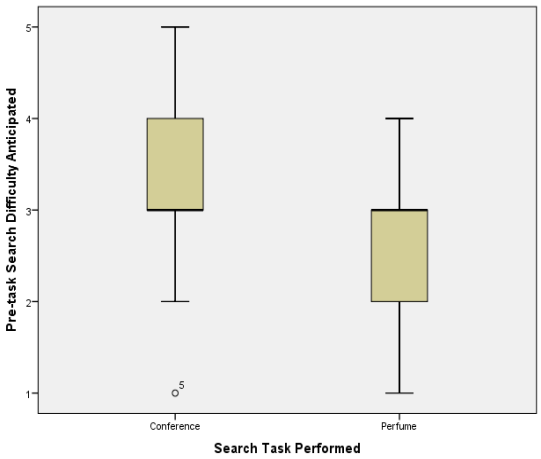
Figure 7.1: Topic Knowledge and Familiarity

7.2.2 Task Difficulty

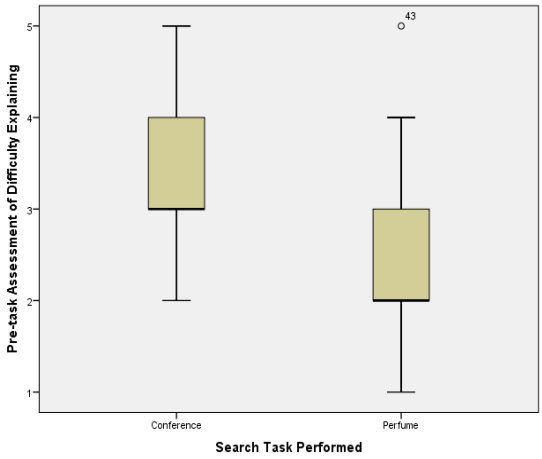
Figure 7.2 shows the boxplots for the anticipated task difficulty, search difficulty, and the difficulty explaining the search task to the agent. The perfume search task was considered less difficult (Mean=2.08, SD=0.182) compared to the conference search (Mean=2.84, SD=0.214). The users also felt that it would be harder to search for a conference (Mean=3.24, SD=0.194) than for a perfume (Mean=2.44, SD=0.183) and also more difficult (Mean=3.2, SD=0.183) to explain it to the agent (Mean=2.44, SD=0.183 for perfume search).



(a) Task Difficulty.



(b) Search Difficulty.



(c) Difficulty Explaining to Agent.

Figure 7.2: Task Difficulty

We observed that although a majority of our participants were students, the perfume search appeared friendlier and more natural compared to the conference search. This could be a function of the gender imbalance in our participants. We had a higher number of female participants (most of them younger students) and were more likely to be familiar with searching for perfume and describing it than males. We evaluated the effect of gender on different parameters for task difficulty. Although gender did not have a significant main effect on any parameter by itself, the interaction effect between gender and system is significant (at $p = 0.01$) for the overall pre-task difficulty assessment ("What is the level of difficulty of the given task?"). Another possible reason could be the wording of the conference search, which may have appeared more domain-specific and technical for most participants. Interestingly, the pilot testers assessed that searching for the conference is easier than searching for perfume.

7.3 Influence of Clarifications

The role of explicit system-level clarification was explored by comparing the experimental system to the control. The control system followed the regular flow of information-seeking conversation without clarifying the agent's model of the user. The experimental system provided explicit system-level clarifications on the queries formulated, information sources used, and general perception of the user's search problem. The clarification script used in the experimental system involved the following categories of system-level clarifications:

- Queries formulated by the system

  Example Script: System: Based on what you said, I am entering the query: <query used>

- System's perception of the user's search problem

  Example Script 1:

  System: To clarify, you are looking for <information object>. Is that correct?

  Example Script 2:

  System: Based on what you said, I understand that you are looking for <information

object>. Am I correct?

- Information Sources Used

  Example Script:

  I am reading the information from the: <name of the website>. Is that alright?

In this section and the next, we explore the effect of different variables in our data on the response (or outcome) variables. The different outcome variables were post-task feedback of the user on various differentials related to the systems, the tasks, and the quality of information. As we were investigating the role of clarification (using the experimental and control systems), the system used was our control (or predictor) variable. However, there were other variables (confounding variables) – task order, search tasks, user age, gender, and background, topic familiarity to name a few – which could have influenced the results. We evaluated the main (or direct) effect of all the variables (both control and confounding) on the outcome variables. We have also assessed the interaction effect of the confounding variables with the control variable (the system used) to check if they are significant.

7.3.1 User Perceptions of the Two Systems

Once the users completed the search tasks using the experimental and control systems, we asked them questions related to the task, the system, and the quality of information. The system-related differentials were: helpfulness, understanding, communication, response, overall experience with the system, and if the user believed that he would have used an alternate search strategy than that followed by the agent.

Table 7.2: System Used and User Feedback

Post-Task Responses	Pearson Correlation with System Used where (1=Conference, 2=Perfume)
System was Easy-to-use	.143
System was Helpful	.000
System Understood the Problem	.170
System Communicated Clearly	.194
System Response was Adequate	.057
Satisfied with Experience	.243
Would have done differently	0.000
Will Use the System Again	.200

The correlation values presented in Table 7.2 suggests that our data do not support a strong interrelationship between the system-oriented variables and the systems used. Two variables – *satisfaction with the experience* (r=0.243) and *will use the system again* (r=0.2) – show some correlation with the system used. This indicates that the use of experimental or control system may influence the user experience and system usability. Figure 7.3 and Figure 7.4 help in visualizing the relationship using boxplots. However, we needed to perform further tests to confirm the effect of clarification (in the experimental system).

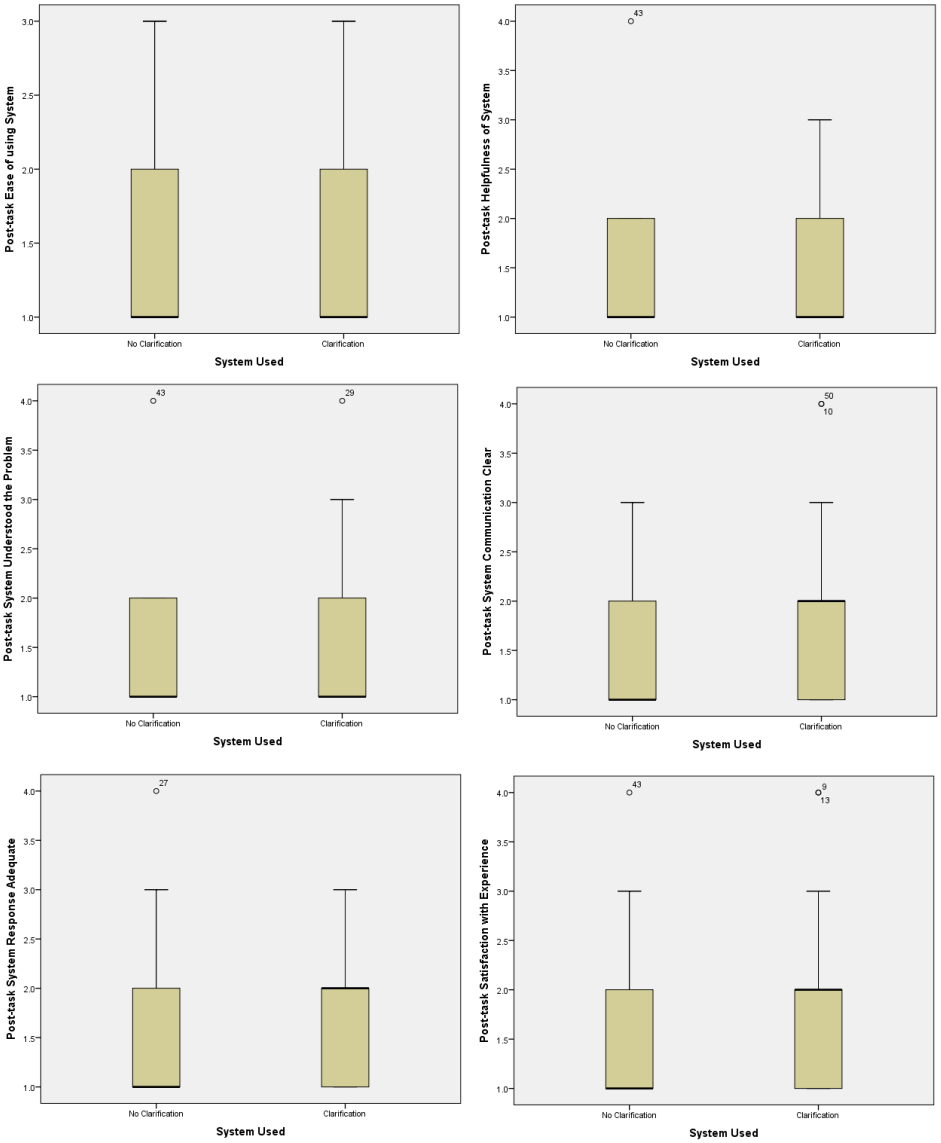


Figure 7.3: Post-task responses by System (part 1)

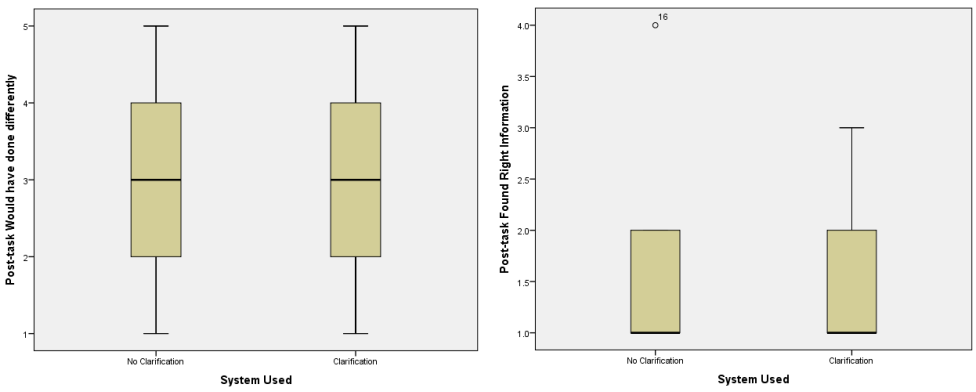


Figure 7.4: Post-task responses by System (part 2)

7.3.2 Testing for the Effect of Clarifications

As the assumption of normality was violated, we performed Wilcoxon Signed Ranks Test – a non-parametric test for paired data – to analyze the difference in user feedback for Experimental and Control systems. Table 7.3 shows the results from the test, which has been obtained by deducting the means of the control system from those of the experimental system.

Table 7.3: Wilcoxon Signed Ranks Test

Post-Task Responses (Experimental - Control) (1=completely agree, 5=completely disagree)	Test-statistic[a] (Z)	Asymp. Sig. (2-tailed)
Task was Complex	-0.171^b	0.86
Task was Difficult	-0.645^c	0.52
Task was Successful	-0.741^b	0.46
System was Easy-to-use	-1.291^b	0.20
System was Helpful	$+0.000^d$	1.00
System Understood the Problem	-1.224^b	0.22
System Communicated Clearly	-1.428^b	0.15
System Response was Adequate	-0.532^b	0.59
Satisfied with Experience	-2.066^b	0.04
Would have done differently	-0.247^c	0.80
Will Use the System Again	-1.933^b	0.05
Found Right Information	-0.312^b	0.75
Better Information was Available	-0.322^b	0.75

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.
c. Based on positive ranks.
d. Sum of negative ranks = Sum of positive ranks.

The Wilcoxon Signed Ranks Test evaluates if both the samples are from the same population, that is, if "the average signed-rank of two dependent samples is zero." It used the standard normal distributed z-value to test if the result was statistically significant.

Feedback on the System

Twenty-five participants were subjected to the Experimental and Control systems. The post-task questionnaire recorded the user feedback on a 5-point Likert Scale. For system-related responses, level 1 represented complete agreement (lowest in scale = highest in feedback).

Table 7.4: Wilcoxon Signed Ranks Test (Differentials Q4-Q8).

Post-task Responses (Experimental - Control) (1 = Completely Agree, 5= Completely Disagree)		N	Mean Rank
System was Easy-to-use	Negative Ranks	4	6.00
	Positive Ranks	8	6.75
	Ties	13	
	Total	25	
System was Helpful	Negative Ranks	4	4.50
	Positive Ranks	4	4.50
	Ties	17	
	Total	25	
System Understood the Problem	Negative Ranks	5	5.70
	Positive Ranks	8	7.81
	Ties	12	
	Total	25	
System Communicated Clearly	Negative Ranks	4	5.50
	Positive Ranks	8	7.00
	Ties	13	
	Total	25	
System Response was Adequate	Negative Ranks	5	8.90
	Positive Ranks	9	6.72
	Ties	11	
	Total	25	

Tables 7.4 and 7.5 show the distribution of positive and negative ranks. As the ranks were calculated as the difference between user agreement for experimental and control

systems. For example, in response to a system-level differential, if the user selected 1 (Completely Agree) for experimental system and 5 (Completely Disagree) for control system, then the rank will be negative. Therefore, a negative rank shows preference of the user towards the experimental system for that specific differential. The reverse is true for positive ranks, which shows that the users preferred the control system (based on the agreement scores)

Let us look at each of the evaluation parameters individually:

1. System was easy to use

   The results indicate that there were 13 ties, 4 negative ranks, and 8 positive ranks. Based on the test, the test-statistic is not significant (Z = -1.291, p = 0.20) and thus, we cannot reject the null hypothesis. We do not have conclusive evidence that the control system (without clarification) was perceived easier by the users. However, more users preferred the control system over experimental system.

2. System was Helpful

   The results show no difference between the experimental and control systems (17 ties, and 4 positive and negative ranks each). Both systems were equally helpful to the users.

3. System Understood the Problem

   The results seem to indicate that the control system was preferred by more users (12 ties, 8 positive, 5 negative ranks). The higher number of ties suggest that most users had no preference between the two systems. As the test statistic is not significant (Z = -1.224, p = 0.22), we do not have sufficient evidence to suggest that clarifications from the system decreased the user's perception that his problem was better understood by the system.

4. System Communicated Clearly

   Once again, the results seem to indicate that the experimental system was preferred by fewer users (13 ties, 8 positive ranks, and 4 negative ranks). However, the observed test-statistic is not significant (Z = -1.428, p = 0.15) and thus, we

cannot reject the null hypothesis. We do not have conclusive evidence that our experimental system (using clarification) was different from the control.

5. System Response was Adequate

   More users agree that the system response was adequate for the control system as compared to the experimental system (11 ties, 9 positive, and 5 negative). The test-statistic is not statistically significant (Z = -0.532, p = 0.59), hence, we cannot reject the null hypothesis.

6. Satisfied with Experience

   The results highlight that the experimental system was preferred by three users (negative ranks), while ten users preferred the control (positive ranks), and twelve were tied. Based on the Signed Ranks test, the observed test-statistic is statistically significant (Z = -2.066, p = 0.04), and thus, we can reject the null hypothesis. Therefore, we have enough evidence to suggest that the experimental system is different from the control system. However, the results are negative which indicate that the control system led to greater user satisfaction with the search tasks. In other words, since both the systems were identical in all aspects except clarifications, we can argue that system-level clarifications decreased the user experience while searching using a voice assistant.

7. Would have done differently

   There are 10 negative ranks, 9 positive ranks, and 6 ties. This indicates that although most users preferred one system over the other, the opinion was equally divided. The observed test statistic is not significant (Z = -0.247, p = 0.8), and thus, we cannot reject the null hypothesis.

8. Will use the System Again

   The results indicate that there were 18 ties, 6 positive ranks, and 1 negative rank. This suggests that while most users did not have any preference, a greater number of users preferred the control system over experimental. Based on the Signed Ranks test, the test-statistic is not statistically significant (Z = -1.933, p = 0.05), and thus, we cannot reject the null hypothesis.

Overall, the results of our user study suggests that the use of system-level clarifications (or explanations) did not produce any positive effect on the user's search experience. In fact, it has lowered the user's overall search experience. The result of the Wilcoxon Signed Ranks Test shows the differences between the control and experimental systems are significant ($p<0.05$) for only one of the parameters.

Table 7.5: Wilcoxon Signed Ranks Test (Differentials Q9-Q13).

Post-task Responses (Experimental - Control) (1 = Completely Agree, 5= Completely Disagree)		N	Mean Rank
Satisfied with Experience	Negative Ranks	3	6.00
	Positive Ranks	10	7.30
	Ties	12	
	Total	25	
Would have done differently	Negative Ranks	10	10.10
	Positive Ranks	9	9.89
	Ties	6	
	Total	25	
Will Use the System Again	Negative Ranks	1	3.00
	Positive Ranks	6	4.17
	Ties	18	
	Total	25	
Better Information was Available	Negative Ranks	6	5.83
	Positive Ranks	6	7.17
	Ties	13	
	Total	25	
Found Right Information	Negative Ranks	5	4.00
	Positive Ranks	4	6.25
	Ties	16	
	Total	25	

Feedback Quality of Information

Two questions in the post-task questionnaire obtain user feedback on the quality of information returned by the agent. Level 1 represented complete agreement (lowest in scale = highest in feedback), and Level 5 shows complete disagreement (Table 7.5).

1. Found Right Information

   The results show that there were 13 ties, and 6 positive and negative ranks each.

While most users had no preference, the remaining were equally divided on the preferred system. However, the test statistic is not statistically significant (Z = -0.312, p = 0.75), which means that there were no differences between the experimental and control systems.

2. Better Information was Available

   For this differential, there were 16 ties, 4 positive ranks, 5 negative ranks. This shows that most users had no preference between the two systems. The test statistic is not statistically significant (Z = -0.322, p = 0.75), so the null hypothesis could not rejected.

The agent was using the same information retrieval mechanism (Google search) in both the cases, and either returned the summary of top-results or read from within the documents. We would have been surprised if there were any differences in the perceived quality of information. Such a difference would have highlighted possible issues with the delivery of information or the results returned by the search engine itself.

Feedback on the Tasks

The pre-task responses suggested that the users possessed greater knowledge about perfumes and considered conference search to be more difficult to search, to explain, and overall. Therefore, it was essential to assess the post-task responses of the users. If one of the two tasks were significantly different from the other, it could affect the other experiential outcomes.

Table 7.6: Wilcoxon Signed Ranks Test (Differentials Q1-Q3).

Post-task Responses (Experimental - Control) (1 = Completely Agree, 5= Completely Disagree)		N	Mean Rank
Task was Complex	Negative Ranks	9	8.11
	Positive Ranks	8	10.00
	Ties	8	
	Total	25	
Task was Difficult	Negative Ranks	9	8.89
	Positive Ranks	7	8.00
	Ties	9	
	Total	25	
Task was Successful	Negative Ranks	4	5.13
	Positive Ranks	6	5.75
	Ties	15	
	Total	25	

There were a total of three task-related questions, and level 1 indicated complete agreement. The task-level parameters are presented in Table 7.6 and explained as follows:

1. Task Complexity

   The results seem to indicate that the tasks were considered equally complex for both the systems. There were 8 ties, 8 positive, and 9 negative ranks. The observed test-statistic was not statistically significant (Z = -0.171, p = 0.86) and the complexity of the search task was not perceived differently in either system.

2. Task Difficulty

   The results seem to indicate that the tasks were considered similarly difficult in both the systems. There are 9 ties, 7 positive ranks, and 9 negative ranks. Once again, the test statistic was not statistically significant (Z = -0.645, p = 0.52).

3. Task Success

   There were 15 ties, 6 positive ranks, and 4 negative ranks. This shows that most users did not consider task success to be different when using either systems.The test statistic was not statistically significant (Z = -0.741, p = 0.46), therefore, we cannot reject the null hypothesis.

Our results show that although the users may have perceived one search task as more difficult than the other, the task-system combination eliminated any influence of the search task topic on the experimental outcomes.

The interview results suggested that the participants were divided in their opinion of the search task. While some considered both tasks to be similar, the conference search task was judged to be tougher than the perfume search task by more participants. This was possibly due to the gender imbalance in the recruited participants, where we had a higher number of undergraduate female students. The age and gender of the participants could have been the reasons why they were more knowledgeable about perfumes and knew how to describe the different features to the Wizard. Gender did not have any direct effect on the user feedback, but significant interaction effects between gender and system used were observed for post-task success assessment ("Task was successful") and pre-task difficulty assessment ("Task was difficult") with $p<0.05$.

> "*I think they were very similar.*" – User 11

> "*In the first search where I was choosing a perfume for my dad, I had options to go in and [find] out different flavors... my budget was not fixed... I can go up and down. But in the second search which was related to my profession. I had to be specific about the area the about choosing the best conference.*" – User 2

> "*I think the perfume was a bit more of a complicated search, so it was helpful to use her rather than searching myself as there are a lot of like different aspects of it.*" – User 6

7.4 Other Observations

Next, we explored some of the other variables in our data and evaluated their influence on the outcome variables – post-task feedback of the user on various differentials related to the systems, the tasks, and the quality of information. As we were investigating the

7.4.1 Effect of Search Task on Post-task Responses

First, we checked the correlation between the experimental setting (the search task performed, the system used, and the order of task presentation) and the post-task responses. The correlation statistics are shown in Table 7.7. We have highlighted the only significant correlation in bold: Going from conference search to perfume search, participants disagree more that they would have done the task differently. We created a univariate general linear model to evaluate the effect of search task performed on the post-task responses. The results indicated significant effect (with p=0.013) of search task performed on one post-task parameter: "would have done differently".

As pre-task responses have shown that participants claim to know more about perfumes than conferences, it is surprising that users agree more that they would not have performed the perfume search differently. The users also agree more that the system obtained right information for the perfume search as compared to conference search. We believe that this could be because of the users awareness of the search space (related to the search topic), which made them believe that the search results solved his information problem. Based on post-task responses, users perceived perfume search to be less difficult, more successful (task-related feedback), and agreed more that they found the right information (quality of information). We investigated the effect of pre-task topic knowledge on all the post-task parameters. While no main effects were found, there were significant interaction effects between pre-task topic knowledge and system used for one system-related differential ("If I were searching, I would have done it differently"). This conforms with findings from previous studies (J. Liu, Liu, & Belkin, 2016; X. Zhang, Liu, & Cole, 2013) about the effect of topic knowledge on search.

As the participants had varying levels of topic knowledge (about perfume and conference), the search tasks might have performed had some influence on the post search task performed. Although we found a direct effect of the search task performed on only one parameter, we measured the interactions between the search task performed and the system used on all the outcome variables. As the search task was a confounding

variable, any such effect could explain the possible influence of clarifications in the experimental system or its lack thereof in the control system. Significant system-search task interaction effects (with $p<0$) were observed for the following responses: 'System was Easy-to-use,' 'System was Helpful,' 'System Communicated Clearly,' 'Satisfied with Experience,' 'Would have done differently,' 'Will Use the System Again,' and 'Better Information was Available.' These observations help in explaining the detrimental effect of clarification in the experimental system. As the users had varying levels of topic knowledge for the two search tasks, their experiences when using the experimental and control systems could have been influenced by the search tasks and not by the system alone. Therefore, more data should be collected, and further analysis should be done to validate the results.

Table 7.7: Effect of Experimental Settings on Post-task Responses

Pearson's Correlation Coefficients			
	Experimental Conditions		
Post-Task Responses (1=completely agree, 5=completely disagree)	Search Task (1=Conf, 2=Perfume)	Task Order	System Used (0=Control, 1=Experimental)
Task-related Feedback			
Task was Complex	.078	.140	.016
Task was Difficult	.263	.033	-.099
Task was Successful	-.258	-.052	.103
System-related Feedback			
System was Easy-to-use	-.143	.029	.143
System was Helpful	-.120	-.120	.000
System Understood the Problem	-.073	-.024	.170
System Communicated Clearly	.145	.097	.194
System Response was Adequate	-.171	-.057	.057
Satisfied with Experience	-.097	.049	.243
Would have done differently	**.349***	.127	0.000
Will Use the System Again	-.143	.143	.200
Quality of Information			
Found Right Information	-.200	-.086	.029
Better Information was Available	.129	.129	.043

7.4.2 Order of Tasks did not Influence User Experience

The order in which tasks were presented to the users did not have any significant correlation with any of the post-task response variables (Table 7.7). We evaluated the effect of task order on all post-task differentials and failed to observe any significant main effect. Also, there were no significant interaction effects of task order and system used on post-task user experiential feedback.

7.4.3 Pre- and Post-task Perceptions on Task Differentials

We have reported the correlation values in Figure 7.8. Significant correlations have been highlighted in the table. Level 1 represents lowest in pre-task differentials but complete agreement in post-task differentials. Therefore, a positive correlation shows an inverse relationship.

Table 7.8: Pre- and Post- Task-related Differentials

Pearson's Correlation Coefficients			
Pre-Task Differentials (1=lowest)	Post-task Responses (1=completely agree)		
	Task was Complex	Task was Difficult	Task was Successful
Topic Knowledge	.197	**.306***	**-.334***
Topic Familiarity	.199	.273	**-.289***
Task Difficulty	**-.350***	**-.429****	**.302***
Perceived Search Difficulty	-.157	-.193	.226
Perceived Difficulty Explaining	-.258	-.209	.249

The results – which conforms with findings from previous studies (J. Liu et al., 2016; X. Zhang et al., 2013) about the effects of topic knowledge on search success and experience – indicate the following:

1. Greater topic knowledge (increasing) made the user disagree that the task was difficult. Instead, the user agreed more that the search task was successful. This could be due to greater confidence in the search space, which allows the user to evaluate the search session and its success better.

2. More (increasing) topic familiarity leads to more agreement that the task was

successful.

3. If the users anticipated the task to be more difficult before the search session began, they expressed an identical opinion after the search session (a higher agreement that the task was complex and difficult). Also, for tasks perceived as less difficult, the users agreed more to the task being successful.

7.4.4 Effect of Gender on User Feedback

In our study, the recruited participants were not the perfect representative sample of the population. Out of the 25 users recruited for the study, 20 were female, and 5 were male. Gender was one of the confounding variables in this study, which could have affected the outcome variable (user experiential factors). Therefore, we evaluated the effect of gender on user feedback. While gender did not have a significant main effect on pre- and post-task differentials, we observed significant interaction effects of gender when combined with the system used. Significant interaction effects – between gender and system used – were observed for post-task success assessment ("Task was successful") and pre-task difficulty assessment ("Task was difficult") with $p<0.05$. The age and gender of the participants could have been the reasons why they were more knowledgeable about perfumes and knew how to describe the different features of the perfume to the Wizard. Knowledge of the search space made the female users agree on the lower difficulty and higher success of the perfume search task compared to the conference search.

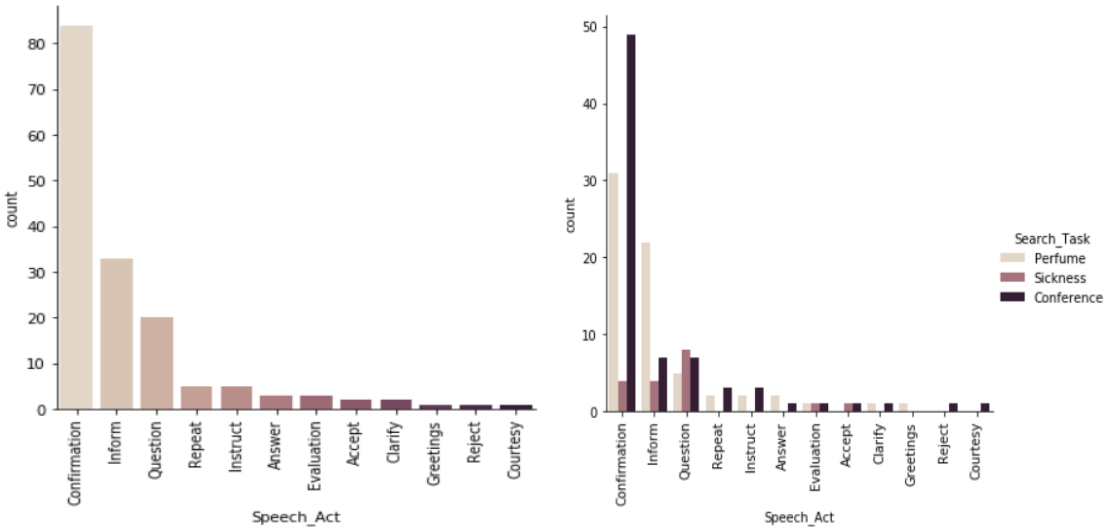
7.5 Discussion

In our study, we investigated the effect of explicit system-level clarifications on the search and interaction experience of the user when using spoken conversational systems. We recruited 25 users who used an experimental system (which offered clarifications) and a control system (without system clarifications) and answered questions related to their system-level experiences. Although we expected explicit system-level clarifications to improve the users' search experience, – system revealment property proposed

by Radlinski and Craswell (2017) – the analysis of the user study data showed that there were no positive effects. Instead, the user's post-task response suggested that satisfaction with the overall search experience was higher in the control system.

To explain the deviation of our findings from the theoretical models that this thesis is based on, we observed the influence of the confounding variables on the outcomes. We also investigated if there were interactions between the control variable (system used) and the confounding variables (gender, topic knowledge, task order, and search tasks). The gender imbalance in the data, high topic knowledge for one of the two tasks, the order in which the tasks are presented could have influenced the search experience. We analyzed the variance between the experimental and control systems and observed that most confounding variables did not any direct effect on the post-task responses. There were exceptions: gender had a significant effect on pre- and post-task difficulty assessments while the search task had a significant effect on the users' perception of how to search ('would have done differently'). However, we observed significant interactions between the search task performed and the system used on multiple post-task parameters.

While our statistical analysis provides some insight into the observed result, we use the interview data (user feedbacks during the exit interviews) to highlight the possible reasons why clarifications could produce a detrimental effect on the users' search experience. First, we look into how the two systems were operationalized. In a user-agent information-seeking interaction, the points of clarifications are random and depend on the intermediary. As our goal was to limit the cognition of the Wizard to a system feasible a few years from now, therefore, we had to develop some rules when clarifications should be offered. In the experimental system, every time the agent searched, it clarified the information sources used, the query terms, or the Wizard's broad-level interpretation of the user's information problem. This allowed the user to correct the agent if required. The user could also take the initiative, instruct the agent, and control the search strategy.



(a) Speech Acts. (b) Speech Acts (by Search Task).

Figure 7.5: User Speech Acts following System Clarification

In Figure 7.5, we present different speech acts (by the user), which immediately follows the system-level clarifications. Figure 7.5a shows that the users responded to more than half of the clarifying questions with confirmation. Few examples of the system-level clarifications were:

- I am entering the query <query words>. Is that ok?
- Would you like me to query <query>?
- I have another article <source>. would you like me to read from

In all the examples above, the user responded in affirmative: "Yes", "Yes, that is ok", "Yeah". The second frequent response was 'Inform' where the user added some additional keywords ("*for men*") or provided additional information ("*I prefer Bergamot but then I want to look around what is in the market*"). The third category of user response ('Question'), which had more than 20 instances, usually followed system elicitation ("*Can you tell me more about what you are looking for?*" or "*Do you have any specific preferences?*"). The user asked a follow-up question ("*What's the nearest one?*") or stated his preference to the system ("*under 100 dollars*"). There were a few instances where the user asked the system to repeat or instructed the system to search in a specific way ("*Now filter by date*"). The high number of affirmatives in

the user responses highlight an interesting fact: the system asked more clarifications because of how it was operationalized, many of which were answered with a single word (*yes* or *yeah*). If the system possessed enough cognition to determine the points of low confidence, clarifications could be limited to only those situations where the user's information need was unclear. The user response to such situations could involve correcting the query used, providing more information, or asking follow-up questions, all of which were observed during the study.

Therefore, the detrimental effect of clarification, as observed in the statistical analysis, could be attributed to the following factors related to the operationalization of the system:

- As the Wizard provides clarification to the user every time he searches and asks the user to confirm, it increases the number of turns and can be perceived as intrusive by the user.
- The frequency of such clarifications and revealments was high, which might have disrupted the usual search process, and therefore, produced a negative effect on the user's search and interaction experience.

User feedback during the exit interview suggests that a majority of the users recognized that the system provided some form of clarification in one of the tasks. We specifically inquired if that was helpful or was an impediment to the flow of conversation. While most users felt that the clarifications were helpful, it was frustrating for a few.

> "*I like how it tells you the query that it's going to give you before because I feel like with like Siri and Alexa they just kind of do their own thing you don't know exactly what they're doing, so it's nice that they tell you that.*" – User 11

> "*I guess it would depend on your audience too, yeah, but I just found it helpful because her reiterating it allowed me to know that she was understanding what I was asking.*" – User 17

The common themes which emerged were clarifications helped users "*double-check what [they] asked*" (User 1) and they appreciated knowing that "*the system is understanding where you're asking because you could be speaking to and if it doesn't understand then what's the point?*" (User 21). It also allowed them to correct the queries ("*fixed it when I restated what I wanted*" – User 11)

However, explicit clarifications were also frustrating for some users. Since the agent declared the queries it was using, few users found that redundant. For example, User 2 says, "*Johanna has to use her brain what she should put in the system*". Similarly, User 12 preferred if " *it [the agent] just would do the query and tell me.*"

> "*I don't know a lot of people look for like human interaction in like a voice assistant... I just want to get my answer you know and I just want to get my answers accurate and as fast as possible.* " – User 10

Some of the comments highlight that system-level clarification if provided, should be done judiciously and only at times when the confidence of the search system is low. This should reduce the number of conversational turns and the search task completion time. Also, since explicit clarifications are not practiced in human-human conversations, and are either implied or done tacitly, we should follow a similar approach during a human-system conversation. The last user comment suggests that the users may not be looking for human-like interactions with a system. However, it should be pointed out that the motivation behind conversational search systems is to provide accurate results to satisfy the user's information need, and conversation is one of the many ways to elicit the information need. Considering the detrimental effect that clarifications might cause on the users' search experience, the system-level clarifications should either be implicit and presented as part of an engaged conversation or provided only at times when the system fails to understand the user. Further research needs to be done to determine how to implement such clarifications without downgrading the search experience of the user.

7.6 Design Recommendations and Desired Functionalities

We designed the user study to answer our first research question (the role of explicit system clarifications) and to prepare the dataset for our second research question (natural language understanding of spoken conversational search agents). However, during the study, while interviewing the users about their experiences, we came across the numerous challenges that people face with conversational search systems. We identified the features which are highly desired (and are feasible), and therefore, we suggest some recommendations for future design and functionalities. We have supported our recommendations with user comments.

7.6.1 Advanced Search Capabilities

For many users, the ability to perform advanced and complex search mattered more than the speed of response.

> "*It's adding more steps, but it's making sure you get the right one every time.*" – User 22

> "*I liked the fact that it is able to give you more than just your yes or no and simple responses*" – User 21

For multiple returned results, the result should be presented as a list. But this allows the user to query by item number in the list instead of querying by name. For longer responses, the agent should present one item from the list at a time and ask the user for confirmation before reading out the next item.

7.6.2 Reporting the Findings: Say less but Save More

A majority of our users suggested that the longer response from the agent was harder to process cognitively. This is in line with previous research (Guy, 2016; Trippas et al., 2018; Turunen et al., 2012) in conversational search systems. However, most of the prior work (and ours) focused on result presentation over audio. Based on the user feedback during the interview, one of the key recommendations is to share the search

history and findings over email or text on request. Our study involved complicated proper nouns (names of perfumes and conferences), and a majority of users felt that while they preferred writing the names down, it was tedious to do so. One alternative would be to send the answer to the user's phone on request or send the link to the documents over email (on demand). Some of the users even wanted the transcript to be forwarded for future reference. Implementing these functionalities can lead to a better result presentation for conversational search systems.

Interviews with participants revealed that many of them preferred some mechanism through which they could save the information. Writing down the agent responses was cognitively demanding (in laboratory environment) and not feasible in real-life (for any of the use cases of conversational search systems). When we asked if the participants would be willing to share their emails or phone numbers for the system to forward the findings, almost all the participants agreed that it would be helpful.

> "*Personally like if I'm on the go and I need information, you know chances are I'm not gonna be able to like to write it down or something so it would be nice if it could have some type of way that it could store the information that I could access it later*" – User 11

> "*If [the agent] could like maybe repeat the search with like what I said and then just give a link to the website or give a few links to websites and not just one link* " – User 20

> "*when it comes to the articles and stuff like that a lot of the times what we do is we'll be saving the article. I was just writing down the information what it was but if I can ask her hey can you email me the article?*" – User 24

While some users preferred receiving just the final answer, others wanted the queries and URLs to be included as well. However, none of them wanted to look at the transcript of the conversation.

7.6.3 Faster Response Time

Although we requested the users to be more patient with our system (as it was a prototype and hence, slower), the slow response time was frustrating for many users.

> "*It felt difficult because she was testing my patience. [The agent] was taking a lot of time to reply me*" – User 2

> "*I don't know a lot of people look for like human interaction in like a voice assistant... I just want to get my answer you know and I just want to get my answers accurate and as fast as possible.* " – User 10

7.6.4 Pause and Control Speed of Utterance

For complex searches, some user suggested the option to pause and resume the playback. The users also preferred if the agent, on request, would break the utterance in parts or slow it down without causing the audio quality to go down (for example, robotic voice). The agent should also be able to spell the words when required, even if such words contain diacritics or accents.

> "*I think pausing would be good option because I'm like, you're writing something and then pause and then resume keep writing.*" – User 6

This was a recurring theme among users (as they had difficulty writing down the responses)

> "*Perhaps the system could go a little slower when it's giving results. I think that would be a good idea.*" – User 21

> "*...maybe slow down the pace should be an option*" — User 25

7.7 Chapter Summary

In this chapter, we explored the role of explicit system-level clarifications on the user's experience with the conversational search system. We used an experimental system

(which provided such clarifications) and a control system and analyzed the user feedback for both systems. The results of the Wilcoxon Signed Rank Test showed that the use of explicit system-level clarifications produced no positive effect on the user's search experience. There were no observed differences between the experimental and control systems for all but one system-level differential. When asked about the overall search experience, the user agreement was higher for the control system over experimental. The difference between the two systems was statistically significant ($Z = -2.066$, $p = 0.04$), but the direction was negative. As the experimental and control systems were similar in all aspects except system-level clarifications, it can be argued that system-level clarifications led to lower satisfaction with the overall search experience. For the task- and information-level differentials, the results did not indicate any difference between the two systems. Since our results did not conform to the theoretical frameworks, we checked the effect of confounding variables on the outcomes. We observed some interaction effects of the control variable with gender, pre-task knowledge, and search tasks, which explains why the findings deviated from expectations. Although the statistical results do not reveal any positive effect of system clarifications, the interview data provided some valuable insights on the preferences of the users. We identified some possible issues with how system revealment was operationalized and suggest how they could be corrected in the future. We report these observations and make recommendations for future design and functionality. However, more tests need to be conducted to confirm the generalizability of these findings and recommendations.

Chapter 8

Towards Natural Language Understanding of Spoken Conversational Search Systems

Our second research question was:

> *How can we automatically predict the different speech acts and the search actions in a user-system information-seeking conversation?*

In this chapter, we report the performance of our Multi-channel Deep Speech Classifier (MDSC) for the two prediction tasks: (i) speech acts (of the user and the agent), and (ii) search actions (of the agent). We have presented the prediction results for the two datasets:

1. The CONVEX dataset which was collected as part of our user study; and
2. The SCS dataset which was available publicly.

For each prediction task, we repeated the experiment 30 times. For each iteration, we picked different training and test sets randomly using different seed values. The classification models were trained from scratch for each iteration using the training data for that iteration. After collecting the accuracy values for 30 iterations, we reported the descriptive statistics for each prediction task. We also performed ablation analysis to evaluate the importance of different categories of features in the classification tasks. Additionally, we have used a collection of popular machine learning algorithms to compare the performance of our model against several off-the-shelf classifiers. The list of baseline classifiers used were as follows: (1) Dummy Classifier with random guesses (str); (2) Dummy classifier with most frequent class (mfq); (3) k-Nearest Neighbors classifier (knn); (4) Gaussian Naive Bayes (gnb); (4) Random Forest Classifier (rfc); (5)

AdaBoost classifier (ada); (6) Quadratic Discriminant Analysis (qda); (7) Support Vector Classifier (svc); (8) Multi-layer Perceptron classifier with stochastic gradient descent (mlp); (9) One-vs-the-rest (OvR) multiclass/multilabel strategy (ovr); (10) One-vs-one multiclass strategy (ovo); (11) Grid Search Cross Validation (gsc); and (12) Decision Tree Classifier (dct). The performance of MDSC models (with prefix *mdsc*) is reported for all the possible combinations of the channels. Channel 1 uses lexical and semantic features (which is also referred to as natural language processing features), Channel 2 uses word embeddings, and Channel 3 uses dialogue metadata features. The pairwise combinations of the channels are shown as *mdsc-12*, *mdsc-13*, and *mdsc-23*, while *mdsc-123* uses all the three channels together.

8.1 Predicting Speech Acts

Our first task was to predict the speech actions in both the datasets. The performance of MDSC classifier (with different configurations) can be visualized in Figure 8.1 and Figure 8.2. The prediction results are summarized and compared with the baseline machine learning classifiers in Table 8.1 (for the CONVEX dataset) and Table 8.2 (for SCS dataset).

8.1.1 CONVEX Dataset

We compare the performance of our prediction model, MDSC, with various machine learning classifiers. Figure 8.1 shows the boxplots for easy visualization while Table 8.1 shows the statistical details.

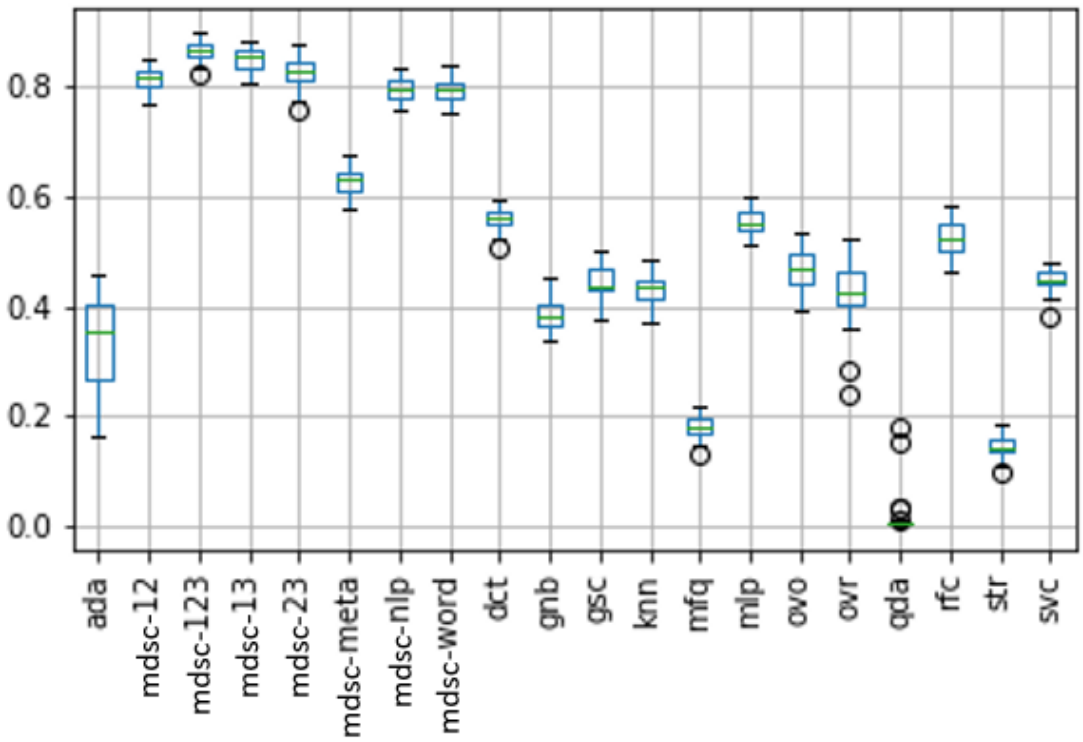


Figure 8.1: CONVEX dataset: Speech Act Prediction Accuracy.

Table 8.1: Predicting Speech Acts: Accuracy on CONVEX dataset

	mean	std	min	25%	50%	75%	max
ada	0.337	0.087	0.161	0.265	0.353	0.404	0.458
mdsc-12	0.816	0.019	0.771	0.801	0.822	0.828	0.853
mdsc-123	**0.868**	**0.016**	**0.823**	**0.859**	**0.869**	**0.877**	**0.902**
mdsc-13	**0.853**	**0.020**	**0.809**	**0.838**	**0.857**	**0.869**	**0.886**
mdsc-23	0.828	0.027	0.760	0.813	0.830	0.847	0.880
mdsc-meta	0.630	0.025	0.580	0.612	0.634	0.646	0.676
mdsc-nlp	0.796	0.020	0.757	0.782	0.796	0.811	0.834
mdsc-word	0.794	0.019	0.755	0.782	0.794	0.807	0.842
dct	0.560	0.021	0.510	0.548	0.563	0.572	0.597
gnb	0.384	0.028	0.338	0.364	0.381	0.403	0.452
gsc	0.443	0.030	0.376	0.428	0.439	0.468	0.501
knn	0.434	0.026	0.371	0.412	0.439	0.446	0.488
mfq	0.183	0.020	0.131	0.170	0.180	0.196	0.218
mlp	0.556	0.022	0.512	0.540	0.550	0.575	0.599
ovo	0.468	0.037	0.390	0.440	0.467	0.495	0.534
ovr	0.425	0.066	0.240	0.403	0.424	0.466	0.526
qda	0.017	0.041	0.000	0.003	0.005	0.005	0.177
rfc	0.526	0.029	0.463	0.504	0.522	0.550	0.583
str	0.144	0.019	0.098	0.134	0.143	0.157	0.183
svc	0.449	0.022	0.379	0.439	0.446	0.463	0.482

We have highlighted the two best performing models in Table 8.1 – MDSC with all the three channels (*mdsc-123*) and MDSC with NLP and metadata features (*mdsc-13*) with a median accuracy of 86.9% and 85.7% respectively.

8.1.2 SCS Dataset

Next, we evaluated the performance of MDSC using the SCS Dataset. The results, as shown in Figure 8.2 and Table 8.2, indicates that highest median accuracy of 64.1% is obtained by MDSC using channels 1 (NLP features) and 3 (Dialogue Metadata). MDSC using the three channels together is a close second with median accuracies of 63.9%. The highest accuracy of the best performing instances is 73.2% and 72.2%, respectively.

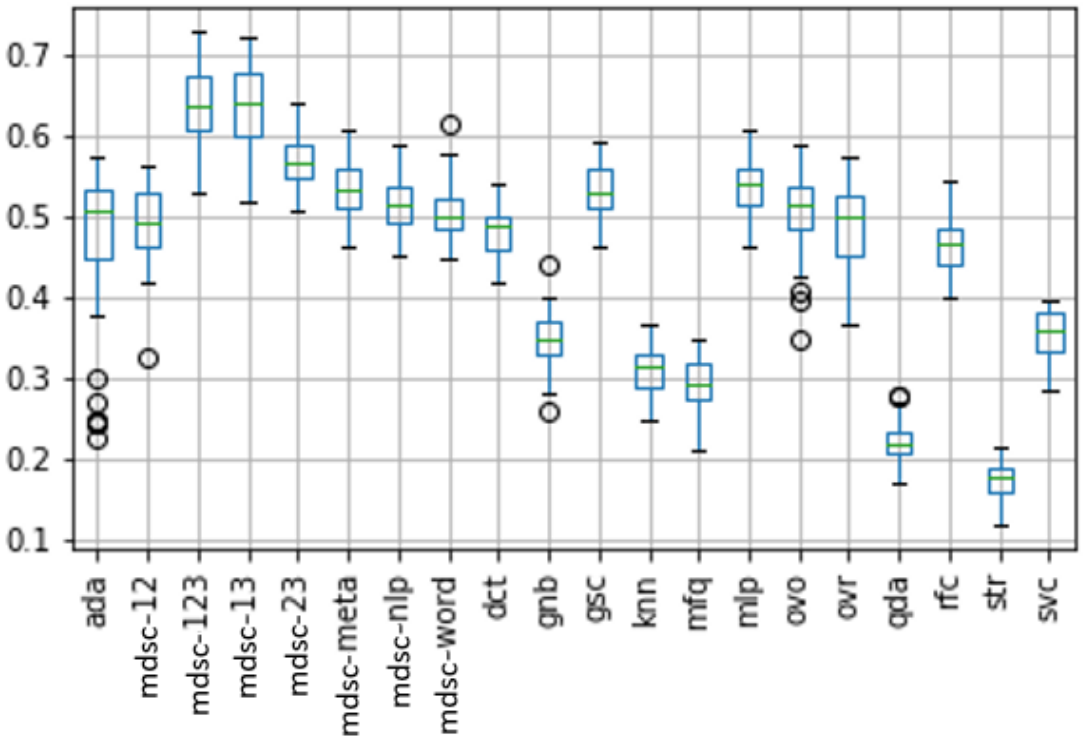


Figure 8.2: SCS dataset: Speech Act Prediction Accuracy.

Table 8.2: Predicting Speech Acts: Accuracy on SCS dataset

	mean	std	min	25%	50%	75%	max
ada	0.459	0.111	0.225	0.449	0.507	0.536	0.574
mdsc-12	0.493	0.052	0.325	0.464	0.495	0.531	0.565
mdsc-123	**0.634**	**0.051**	**0.531**	**0.608**	**0.639**	**0.675**	**0.732**
mdsc-13	**0.633**	**0.052**	**0.522**	**0.602**	**0.641**	**0.679**	**0.722**
mdsc-23	0.569	0.036	0.507	0.549	0.569	0.591	0.641
mdsc-meta	0.537	0.036	0.464	0.513	0.533	0.560	0.608
mdsc-nlp	0.517	0.034	0.455	0.494	0.517	0.539	0.589
mdsc-word	0.507	0.035	0.450	0.484	0.502	0.524	0.617
dct	0.484	0.028	0.421	0.459	0.490	0.501	0.541
gnb	0.349	0.041	0.258	0.331	0.349	0.372	0.440
gsc	0.536	0.032	0.464	0.513	0.531	0.560	0.593
knn	0.310	0.029	0.249	0.291	0.316	0.329	0.368
mfq	0.296	0.032	0.211	0.274	0.294	0.319	0.349
mlp	0.537	0.035	0.464	0.514	0.543	0.560	0.608
ovo	0.506	0.059	0.349	0.488	0.514	0.539	0.589
ovr	0.486	0.050	0.368	0.455	0.500	0.526	0.574
qda	0.224	0.026	0.172	0.207	0.220	0.234	0.278
rfc	0.465	0.034	0.402	0.443	0.467	0.488	0.545
str	0.174	0.023	0.120	0.159	0.179	0.190	0.215
svc	0.356	0.030	0.287	0.335	0.359	0.382	0.397

8.1.3 Discussion

For both the datasets, we performed ablation analysis by using each of the three channels individually, combining them in pairs, and all the three channels together. The different configurations of MDSC are shown in the Tables using prefix *mdsc*. The highest reported accuracy was 90.2% and 73.2% for CONVEX and SCS datasets, respectively. The results of ablation analysis indicate that the best performance is achieved using all the three channels, that is, the word embeddings (channel 2), the lexical and semantic features (channel 1), and the dialogue metadata (channel 3). However, for SCS dataset, the median performance of the *mdsc-13* (64.1% accuracy) is slightly better than *mdsc-123* (63.9% accuracy). The lower accuracy on SCS data can be attributed to the human-human nature of the dialogues. There are multiple occasions where the utterances resemble casual chit-chats and do not follow a specific set of rules. None of the three feature channels show comparable performance when used individually, and the largest drop in performance is observed on removing the metadata features (as

observed in *mdsc-12*), followed by natural language features (*mdsc-23*). The results are consistent for both the datasets.

		ACTUAL LABELS											
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
PREDICTED LABELS	S1	61	0	1	2	3	1	0	0	4	0	0	0
	S2	0	59	0	0	0	0	0	0	0	0	2	1
	S3	0	0	2	0	2	0	0	0	0	0	0	0
	S4	0	1	1	53	0	0	0	0	1	0	0	0
	S5	0	0	0	2	24	0	0	0	1	0	0	0
	S6	2	0	0	1	0	3	0	1	0	0	0	0
	S7	0	0	0	0	0	0	1	0	0	0	1	0
	S8	0	0	0	0	0	0	0	0	0	0	0	0
	S9	1	1	0	7	1	1	0	0	42	0	0	0
	S10	0	0	0	0	0	1	0	0	0	21	1	0
	S11	0	0	0	0	1	0	0	0	0	0	0	0
	S12	2	0	0	1	0	1	2	0	0	1	0	53

(a) CONVEX (MDSC-123).

		ACTUAL LABELS											
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
PREDICTED LABELS	S1	29	0	0	1	0	5	0	3	3	1	1	0
	S2	0	0	0	0	0	0	0	0	0	0	0	0
	S3	0	0	1	5	2	0	0	0	0	0	0	0
	S4	0	1	3	41	4	5	0	0	1	0	0	0
	S5	0	1	1	2	19	0	0	1	2	0	0	0
	S6	2	0	0	0	0	9	1	1	0	1	0	1
	S7	0	0	0	0	0	0	0	0	0	0	0	0
	S8	11	0	0	0	0	3	0	6	0	0	0	0
	S9	0	0	0	0	0	0	0	0	0	0	0	0
	S10	8	2	0	0	1	0	1	0	0	9	2	16
	S11	0	0	0	0	0	0	0	0	0	0	0	0
	S12	1	0	0	0	0	0	0	0	0	0	0	2

(b) SCS (MDSC-123).

Figure 8.3: Confusion Matrix (Speech): CONVEX and SCS (MDSC-123)

In Figure 8.3, we present the confusion matrix for speech prediction using *mdsc-123* (MDSC with three channels together). Figure 8.3a and Figure 8.3b are for CONVEX and SCS data respectively. The iterations have been selected at random. The green cells are the correct predictions while the red cells show the major misclassifications by the model.

For CONVEX speech prediction, four instances of S9 (Repeat) were classified as S1 (Question/Seek). A closer inspection into the dataset revealed that all the misclassifications occured when the user requested the agent to repeat the last utterance. The model wrongly assumed that the user utterance is a question instead of a request to repeat. Similarly, the answers by the agent were marked as S4 by the annotators for the first time and as S9 for subsequent repetitions. The model failed to recognize the rule followed by the human annotators and classified multiple instances of S4 (Answer

by the agent) as S9 (Repetitions by the agent). On two occasions, the model confused courtesy (S11) by the agent as accept (S2). Both of these instances involved only the word "Okay". In three cases, implicit requests for clarifications (S5) were wrongly classified as Question/Seek (S1). An example of correctly classified instance is: "Based on what you said I am running the query "top conferences in Artificial Intelligence" am I correct?" while a wrongly classified instance is: "It's a perfume for men or women?"

For SCS speech predictions, 11 instances of S1 (Question/Seek) were misclassified as S8 (Instruct). On observing the specific instances, we noticed that the wrongly classified instances were either towards the end of the conversation (utterance number > 30) or did not end with a question mark. All the correctly predicted instances (of S1) ended with question mark. Similar to our observations for CONVEX, three instances of S9 (Repeat) were classified as S1 (Question/Seek). It must be noted that S9 is a minority class in SCS dataset. The highest number of misclassification occured when predicting S12 (Greetings and Closing Rituals). 16 instances of S12 were predicted as S10 (Confirmation). A close inspection revealed that all the wrongly classified utterances had words like "Okay" or "Yeah" which were heavily used for confirmation.

Let us look at utterance number 5 (by the agent) for User 21:

> "*Here is what I have found: Deep Learning Summit San Francisco, Insurance AI and Innovative Tech, USA 2020; Deep Learning in Health Care Summit, Boston; and AI for CPG Summit. Would you like me to list some more?*"

Human annotators labeled this instance as S4 (Answer) and all the annotators were in complete agreement. The *mdsc-metadata* model classified this utterance as S5 (Clarify) and *mdsc-word* and *mdsc-nlp* models classified the utterance as S9 (Repeat). Machine Learning models depend on the confidence scores to make prediction. The *mdsc-metadata* model looked at different features like duration of the utterance, speaker of the utterance, utterance number, and previous speech act and search actions. But since *mdsc-metadata* did not use word-level meanings, it failed to get enough confidence to label the instance as S4. Similarly, *mdsc-word* used only word-embeddings and *mdsc-nlp*

used only lexical and semantic features without any without any contextual information like the speaker, utterance number, previous user speech act and previous agent search act. Therefore, both these models could not differentiate between the agent generated S4 and S9. Individually, none of the models (using different channels) achieved enough confidence to make the correct prediction. However, when all the channels were used together in MDSC-123, the combination of weak clues from each channel was able to push the confidence past the threshold, and resulted in accurate detection. This is a classic example of Gestalt principle (Koffka, 2013) where "the whole is greater than the sum of its parts."

Overall, for speech act prediction, MSDC outperforms all the traditional classification models by a large margin and shows improvements of 54.4% and 18.3% over the nearest baseline for CONVEX and SCS, respectively (Decision Tree with 56.3% median accuracy and Multilayer Perceptron with 51.4% median accuracy). Based on our findings, it can be argued that the dialogue metadata features like the relative position of the utterance in the conversation, or the duration of it are extremely important in predicting the speech acts. Also, different lexical and semantic features could be crucial in natural language understanding. Although word embeddings are important, we need a considerable amount of training data before the model can recognize underlying patterns.

Table 8.3: Statistical Significance using Wilcoxon Signed-Rank Test (Speech Act)

Classifier 2	Classifier 1 (mdsc-123)	
	CONVEX	SCS
	(If p<0.05?)	(If p<0.05?)
ada	Yes	Yes
dct	Yes	Yes
gnb	Yes	Yes
gsc	Yes	Yes
knn	Yes	Yes
mdsc-12	Yes	Yes
mdsc-13	Yes	**No**
mdsc-23	Yes	Yes
mdsc-meta	Yes	Yes
mdsc-nlp	Yes	Yes
mdsc-word	Yes	Yes
mfq	Yes	Yes
mlp	Yes	Yes
ovo	Yes	Yes
ovr	Yes	Yes
qda	Yes	Yes
rfc	Yes	Yes
str	Yes	Yes
svc	Yes	Yes

Next, we evaluated the significance of our classification results using statistical analysis. This helped us in assessing if the difference between MDSC and the other classifiers are real or by chance. As the training and test sets were chosen from the same dataset (for 30 iterations), the observations were not independent, and so we opted for a non-parametric Wilcoxon signed-rank test for pairwise comparisons. The best performing MDSC classifier for both the datasets – MDSC with three channels (*mdsc-123*) – was compared with all the baseline classifiers. The statistical significance is reported in Table 8.3. The results emphasize that for the CONVEX dataset, the MDSC model with three channels (*mdsc-123*) is significantly better than all the baseline classifiers. For the SCS dataset, the results were significantly different when compared to the baseline classifiers. However, the improvement was not significant when compared to *mdsc-3*, which is understandable as it is a different variation of the same architecture.

8.2 Predicting Search Actions

Next, we used the MDSC model to predict the search actions performed by the agent (Wizard). The search actions were initiated in response to the spoken utterances of the user. Therefore, instead of using the spoken utterance of the current turn, we have used the previous utterance by the user to generate the word-embeddings.

8.2.1 CONVEX Dataset

When predicting search actions for the CONVEX dataset, the best performing model was MDSC with channels 1 (NLP features) and 3 (dialogue metadata). The highest reported accuracy was 63.7%. However, the highest median accuracy was obtained using MDSC with channel 3 (dialogue metadata). Both the models are highlighted in Table 8.4. The boxplot can be seen in Figure 8.4.

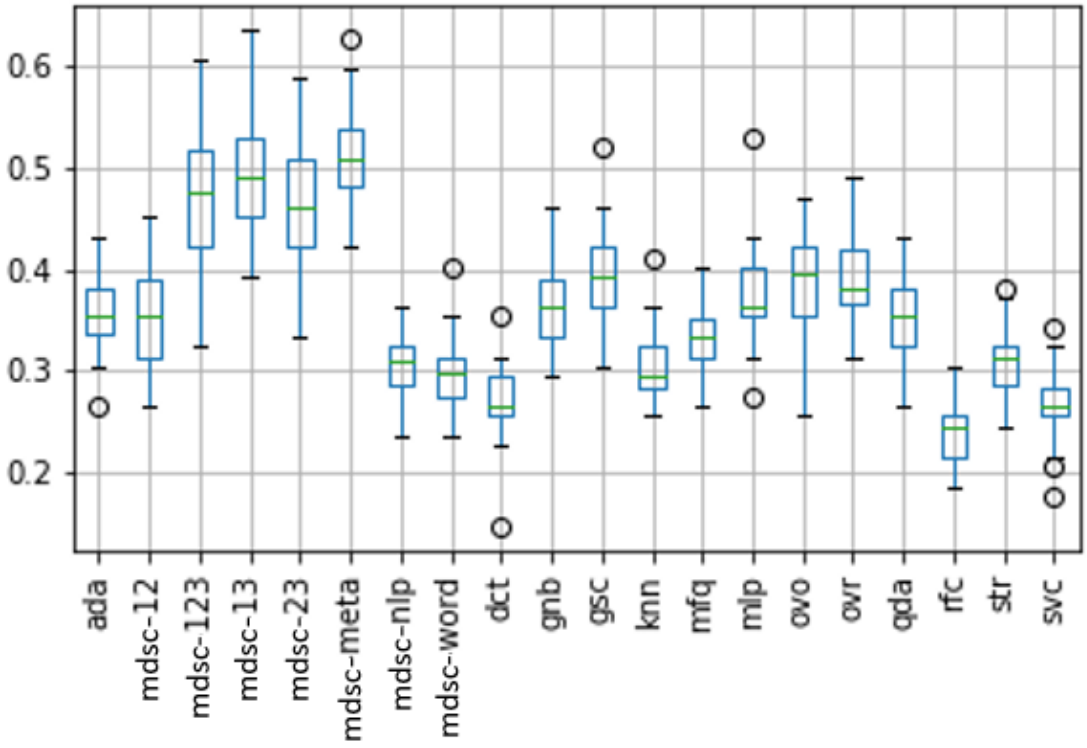


Figure 8.4: CONVEX dataset: Search Act Prediction Accuracy.

Table 8.4: Predicting Search Acts: Accuracy on CONVEX dataset

	mean	std	min	25%	50%	75%	max
ada	0.358	0.039	0.265	0.336	0.353	0.380	0.431
mdsc-12	0.351	0.053	0.265	0.314	0.353	0.390	0.451
mdsc-123	0.475	0.069	0.324	0.424	0.475	0.517	0.608
mdsc-13	**0.491**	**0.056**	**0.392**	**0.453**	**0.490**	**0.529**	**0.637**
mdsc-23	0.457	0.067	0.333	0.422	0.461	0.510	0.588
mdsc-meta	**0.510**	**0.048**	**0.422**	**0.483**	**0.510**	**0.539**	**0.627**
mdsc-nlp	0.308	0.027	0.235	0.287	0.309	0.324	0.363
mdsc-word	0.294	0.036	0.235	0.275	0.299	0.314	0.402
dct	0.269	0.037	0.147	0.255	0.265	0.294	0.353
gnb	0.365	0.044	0.294	0.333	0.363	0.390	0.461
gsc	0.394	0.048	0.304	0.363	0.392	0.422	0.520
knn	0.303	0.034	0.255	0.284	0.294	0.324	0.412
mfq	0.330	0.027	0.265	0.314	0.333	0.350	0.402
mlp	0.372	0.046	0.275	0.353	0.363	0.402	0.529
ovo	0.390	0.050	0.255	0.353	0.397	0.422	0.471
ovr	0.388	0.044	0.314	0.365	0.382	0.419	0.490
qda	0.356	0.042	0.265	0.326	0.353	0.380	0.431
rfc	0.238	0.031	0.186	0.216	0.245	0.255	0.304
str	0.309	0.037	0.245	0.287	0.314	0.324	0.382
svc	0.270	0.036	0.176	0.255	0.265	0.284	0.343

8.2.2 SCS Dataset

When predicting search actions on the SCS dataset, the best median accuracy was obtained using MDSC with metadata features (*mdsc-meta*). However, the highest accuracy was reported by the AdaBoost model (63.3%). The MDSC model had the highest accuracy of 60%. More details can be found on Figure 8.5 and Table 8.5.

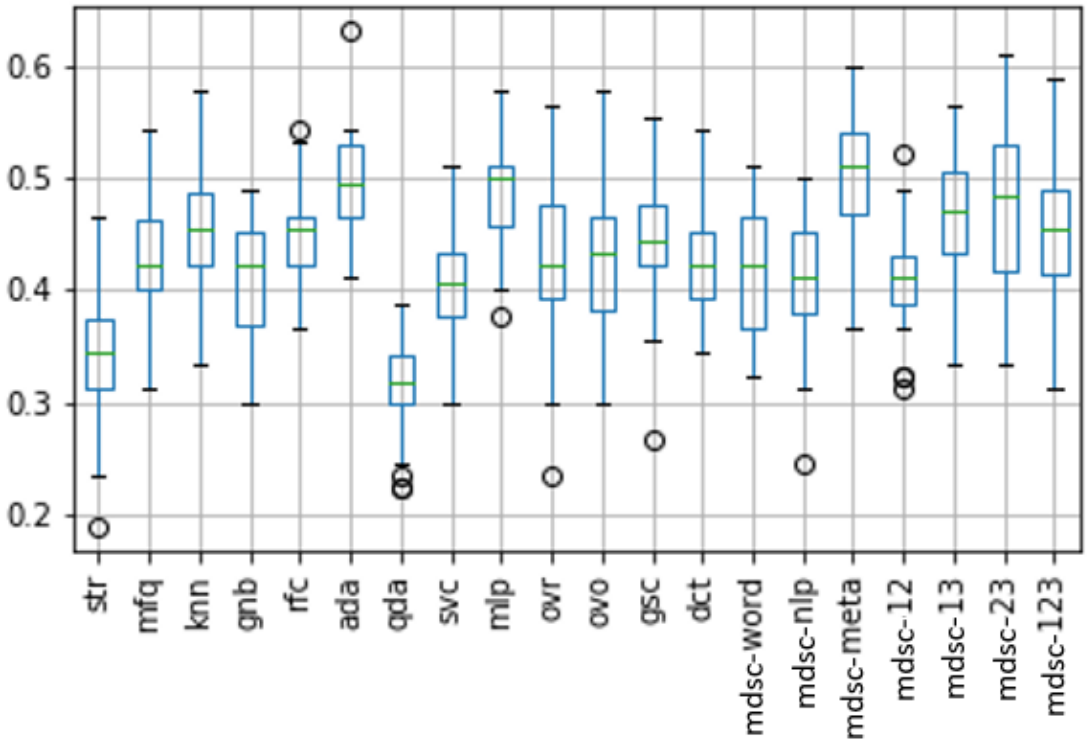


Figure 8.5: SCS dataset: Search Act Prediction Accuracy.

Table 8.5: Predicting Search Acts: Accuracy on SCS dataset

	mean	std	min	25%	50%	75%	max
ada	**0.494**	**0.047**	**0.411**	**0.467**	**0.494**	**0.531**	**0.633**
mdsc-12	0.410	0.047	0.311	0.389	0.411	0.431	0.522
mdsc-123	0.461	0.064	0.311	0.414	0.456	0.489	0.589
mdsc-13	0.464	0.060	0.333	0.433	0.472	0.506	0.567
mdsc-23	0.477	0.070	0.333	0.417	0.483	0.531	0.611
mdsc-meta	**0.503**	**0.060**	**0.367**	**0.469**	**0.511**	**0.542**	**0.600**
mdsc-nlp	0.409	0.057	0.244	0.381	0.411	0.453	0.500
mdsc-word	0.417	0.056	0.322	0.367	0.422	0.467	0.511
dct	0.427	0.056	0.344	0.392	0.422	0.453	0.544
gnb	0.410	0.052	0.300	0.369	0.422	0.453	0.489
gsc	0.443	0.056	0.267	0.422	0.444	0.478	0.556
knn	0.454	0.054	0.333	0.422	0.456	0.486	0.578
mfq	0.430	0.053	0.311	0.400	0.422	0.464	0.544
mlp	0.486	0.043	0.378	0.458	0.500	0.511	0.578
ovo	0.432	0.069	0.300	0.383	0.433	0.467	0.578
ovr	0.426	0.075	0.233	0.392	0.422	0.478	0.567
qda	0.313	0.043	0.222	0.300	0.317	0.342	0.389
rfc	0.453	0.045	0.367	0.422	0.456	0.467	0.544
str	0.340	0.059	0.189	0.311	0.344	0.375	0.467
svc	0.407	0.048	0.300	0.378	0.406	0.433	0.511

8.2.3 Discussion

Once again, we performed an ablation analysis for both the datasets. The highest reported accuracy was 63.7% and 63.3% for CONVEX and SCS datasets, respectively. To predict the appropriate search action for the agent, we used the user utterance in the last turn to generate the word embeddings. The results indicate that MDSC using metadata features slightly outperforms the baseline classifiers. MDSC with metadata channel (*mdsc-meta*) had the highest median accuracy for both the datasets with 51% and 51.1%, respectively. The improvements were 32.3% and 2.2% over the closest machine learning baselines (OVO with 39.7% accuracy for CONVEX and MLP with 50% accuracy for SCS). For both the datasets, metadata features were the most important for MDSC.

		ORIGINAL LABELS			
		SR1	SR2	SR3	SR4
PREDICTED	SR1	16	0	9	0
	SR2	5	18	7	2
	SR3	14	2	22	6
	SR4	0	1	0	0

(a) CONVEX (MDSC-Meta).

		ORIGINAL LABELS			
		SR1	SR2	SR3	SR4
PREDICTED	SR1	15	3	4	0
	SR2	13	25	12	0
	SR3	2	10	6	0
	SR4	0	0	0	0

(b) SCS (MDSC-Meta).

Figure 8.6: Confusion Matrix (Search Actions): MDSC-Meta

Unlike speech act prediction, combining all the channels led to a decline in performance when predicting search actions. One of the likely reasons is the significance of dialogue features in determining the search actions of the intermediary. The speech act of the user immediately preceding the intermediary's action governs the search tactic to be followed and therefore, could be used to develop an understanding of how to search in response to spoken utterances. Word-based features are necessary to build context, and in framing the queries, but they have limited influence (compared to speech acts)

while determining the search action. Second, search actions are performed only by the intermediary; therefore, the number of instances with the search actions in both the datasets was less than half of those with speech acts. We believe that as the training data decreased while predicting search actions, our model failed to identify underlying patterns in word embeddings and lexical and semantic features. As a result, both the channels showed poor performance and led to lower accuracy when combined.

In Figure 8.6, we present the confusion matrix for search act prediction using the MSDC with metadata channel (*msdc-meta*). Figures 8.6a and 8.6b are for CONVEX and SCS data respectively. Once again, the correct predictions are the diagonal elements (highlighted in green) and the major misclassifications have been highlighted in red. For CONVEX data, 14 instances of SR1 (Query Creation or Refinement) were wrongly classified as SR3 (Document Scanning). Although analysis of the data did not reveal any noticeable insights, we did notice that the correctly classified instances of SR1 were preceded by user utterances that had a mean duration of 23 seconds whereas the incorrectly classified instances were preceded by user utterance that had a mean duration of 12 seconds or less. There is a possibility that the model assumed that longer utterances involved the users stating (or restating) their information need and called for query creation/modification. Therefore, the model predicted document scanning for shorter utterances (which might involve identifying sections within document). Similarly, all the wrongly classified instances of SR3 were preceded by user utterances of duration longer than 10 seconds and lesser than 30 seconds. They were classified either as SR1 or SR2. Next, we noted that SR4 (Organizing Answer from Multiple Documents) was a minority class and none of the instances of SR4 were correctly classified. Six instances of SR4 were marked as SR3 and two as SR2.

In Figure 8.6b, we see the matrix for SCS data. The last column shows that SCS data did not contain any instances of SR4. As the data did contain details of the search actions performed but broad labels, it was not possible to recognize multi-document answers. We noticed that 13 instances of SR1 and 12 of SR3 were wrongly classified as SR2. One possible explanation is the higher number of SR2 labels in the training dataset. Since SR2 is the majority class (almost 43% of the search action labels are

SR2), our model is biased towards SR2. Training with a larger dataset with oversampled minority classes could improve the model performance.

Table 8.6: Statistical Significance using Wilcoxon Signed-Rank Test (Search Action)

Classifier 2	Classifier 1 (mdsc-meta)	
	CONVEX	SCS
	(If p<0.05?)	(If p<0.05?)
ada	Yes	Yes
dct	Yes	Yes
gnb	Yes	Yes
gsc	Yes	**No**
knn	Yes	Yes
mdsc-12	Yes	Yes
mdsc-123	Yes	Yes
mdsc-13	**No**	Yes
mdsc-23	Yes	Yes
mdsc-nlp	Yes	**No**
mdsc-word	Yes	Yes
mfq	Yes	Yes
mlp	Yes	**No**
ovo	Yes	Yes
ovr	Yes	Yes
qda	Yes	Yes
rfc	Yes	Yes
str	Yes	Yes
svc	Yes	Yes

The statistical significance is reported in Table 8.6. Once again, we used Wilcoxon signed-rank test – which is non-parametric – for pairwise comparisons. The best performing MDSC classifier for both the datasets – MDSC with metadata channel (*mdsc-meta*) – has been compared with the other classifiers. The results proved that for the CONVEX dataset, the MDSC model with the metadata channel (*mdsc-meta*) was significantly better than all the baseline classifiers. However, no significant improvement was observed when *mdsc-meta* was compared with *mdsc-13*. For SCS dataset, no significant difference was observed when the results of *mdsc-meta* was compared to *gsc*, *mlp*, and *mdsc-nlp*. For all other classifiers, the results were significantly different.

8.3 Chapter Summary

In this chapter, we proposed a deep neural approach to identify the speech acts and the search actions in human-agent information-seeking dialogues. Since speech acts convey the meaning of the utterances, by identifying the speech acts automatically, we can develop an understanding of what is being said by the user. As the utterance of the user should guide the search action performed by the agent, we can argue that the speech act of the user could be used as a feature to predict the search action which the agent should perform. Therefore, by predicting the speech acts and the search actions accurately, the conversational search system can decide on what the ideal response should be (as each label represents different categories of action). Overall, our research contributes towards the natural language understanding of spoken conversational search systems.

To a human intermediary, the content of the dialogue provides sufficient indication for the search actions necessary. However, for a non-human intermediary (as our system here), large amounts of training data are required to identify patterns from only the words in the conversational dialogue. By adding the different natural language and the metadata features, we optimize the prediction process and considerably lower the training and computation times. Feature engineering is important as training and deploying a massive deep neural model is not only expensive (both financially and computationally) but also produces emissions which have an adverse effect on the environment (Strubell, Ganesh, & McCallum, 2019). The neural model which we proposed here is simple and modular. We have tested our model on two datasets – the CONVEX dataset created by us and SCS dataset available publicly – and our model shows impressive accuracy while predicting speech acts (highest of 90.2% for CONVEX and 73.2% for SCS dataset). The accuracy while predicting search actions is lower (63.7% on CONVEX and 60% on SCS), but training on a larger dataset should improve the performance in the future. Overall, our classifier (MDSC) outperforms the existing machine learning baselines significantly and could be extended to other datasets with little or no additional effort. Our results are fully reproducible, and the code and data

will be made public on a major code sharing platform.

Chapter 9

Conclusion

In this chapter, we conclude the thesis by providing a brief summary of the research undertaken, the contributions and practical implications of our findings, and the possible directions for future research. Conversational systems are becoming increasingly popular in our everyday lives as they allow multitasking and handsfree operations. However, as is common with many emerging technologies, there are several challenges which the scientific community is trying to address, which range from understanding the user's utterances to result presentation and evaluation of such systems. In our work, we explored the two of the intelligent functionalities expected of conversational search systems. We concentrate on two of the many intelligent functionalities desired of conversational search systems:

- Natural language understanding (how to connect what is being said by the user to the search actions to be performed by the agent?);
- System-level clarifications (does explicit clarifications by the agent facilitate user-agent information-seeking conversations and, therefore, create a better search experience for the user?)

We focus on search systems specifically, with voice-only input and output, although many of our findings could be extended to multimodal search systems.

9.1 Thesis Summary

In this thesis, we presented an overview of the Conversational Search Systems as a research domain and proposed possible solutions to some of the existing challenges in such systems. While conversations are the natural mode of communication for humans,

search systems have traditionally been textual and browser-based. With the popularity of mobile and wearable devices, there is an increasing demand for voice-based search systems. Users can interact with these systems – also called digital personal assistants – using natural language dialogues. However, as we highlighted in Chapters 2 and 3, numerous challenges must be overcome in different stages of information retrieval. Current and future research in this domain is focused on various aspects of information organization, human-system interaction, result presentation, and evaluation strategies.

In our work, we have focused on two of the intelligent functionalities that are expected from conversational search systems. The first explores system-level clarifications that the system could provide to the user. Prior research (Hollnagel, 1979) has suggested that successful communication between the user and the system can only be achieved when both have developed clear models of each other. The act of model building could involve user- and system-revealment (Radlinski & Craswell, 2017) during the search session. As the search task increases in complexity, it requires a greater effort to understand the user's information problem. The effort could be in the form of higher cognition (as in a human-human conversation), more contextual awareness (through better knowledge representation), or more extended conversations. Therefore, in our research, we have assessed the influence of explicit system-level clarification on the user's search experience when performing moderately complex search tasks. We performed a within-subjects Wizard-of-Oz experiment – with two systems, experimental and control – to answer our research question. In the experimental system, the agent clarified its model every time it searched for information. The clarification included the query terms, information sources, or the agent's understanding of the user's information need. Our analysis of the user study data revealed no significant insight into the effect on system-level clarification. There were no observed differences between the experimental and control systems for all but one post-task response. When asked about satisfaction with the overall search experience, the user agreement was higher for the control system over experimental. The difference between the two systems was statistically significant ($Z = -2.066$, $p = 0.04$), but the direction of the result was negative.

As the observations were contrary to our expectations, we investigated for possible

reasons why clarifications did not improve the users' search experience. We assessed if the confounding variables (gender, topic knowledge, task order, and search tasks) had any direct effect on the user responses or any interaction effect with the control variable (system used). While most confounding variables did not any direct effect on the post-task responses, gender had a significant effect on pre- and post-task difficulty assessments while the search task had a significant effect on the users' perception of how to search ('would have done differently'). Significant interactions were observed between the search task performed and the system used on multiple post-task parameters. In the future, further investigations are required with a larger number of participants and different search tasks. Our analysis of the exit interviews provides some insights into how the operationalization of the system could have influenced the findings and possible alterations for future studies. For example, the Wizard provided clarification to the user every time the Wizard searched and asked the user for confirmation. This increased the number of turns and might have been perceived as intrusive by the user. Also, the frequency of such clarifications and revealments were high and might have disrupted the usual search process, and therefore, produced a negative effect on the user's search and interaction experience. For future work, we suggest that the system-level clarifications should either be implicit and presented as part of an engaged conversation or provided only at times when the system fails to understand the user. Further research needs to be done to determine how to implement such clarifications without downgrading the search experience of the user. We analyzed the user interviews and reported some other observations from the study. We also proposed some recommendations about the design and functionality expected in future systems.

The second research question explored the natural language understanding of conversational search systems. During a search session, the spoken utterances of the user guide the actions of the agent. Depending on the utterance of the user, the agent may or may not perform a search action. In either case, the agent responds in natural language, either furthering the conversation or responding with an answer (if it performs a search action). In our work, we have used the concept of Speech (or Dialogue) Acts from Linguistics to develop a natural language understanding model for the search agent.

Speech acts convey the meaning of the utterance on a functional level. By identifying the speech acts automatically, we can develop an understanding of what is being said by the user and what should be the response of the agent. The speech act of the user should also influence the search action of the agent. Therefore, we develop two predictive models: one to predict the speech act of the user and the agent and the other to predict the search action performed by the agent. To build the machine learning classifiers, we use the CONVEX dataset (developed by us) and the SCS dataset (available publicly). We performed a thematic analysis of the data to develop a set of qualitative codes. Next, we use these codes to annotate each utterance in the two datasets with the search acts and corresponding search actions. Finally, we developed a multi-channel deep neural classifier (MDSC) to perform the prediction tasks. We trained our classifier on three different categories of features – word-embedding of the utterances, lexical and semantic features, and the dialogue metadata. We reported the accuracy of our classifier along with the confusion matrices. We also tested the statistical significance of our model when compared with baselines, and performed ablation analysis to show the importance of each category of features. The results indicate that the MDSC model (with best configurations) achieved the highest accuracy of 90.2% for CONVEX and 73.2% for the SCS dataset (when predicting speech acts). The accuracy of prediction for search actions was 63.7% for CONVEX and 60% for SCS data. Metadata features were most important for both the prediction tasks, while word-embeddings were least effective. For speech act prediction, the best performing model used all the three channels together. Overall, our classifier (MDSC) outperformed the existing machine learning baselines significantly. To a human intermediary, the content of the dialogue provides sufficient indication for the search actions necessary. However, for a non-human intermediary (as our system here), large amounts of training data are required to identify patterns from only the words in the conversational dialogue. By adding the different natural language and the metadata features, we optimize the prediction process and considerably lower the training and computation times. Also, the simple and modular nature of our model ensures that it could be extended for similar user study datasets with few changes.

9.2 Contributions

Our contributions focus on the design and development of more user-centered spoken conversational search systems. In the following subsections, we try to present an overview of the significant contributions of this thesis.

9.2.1 Detailed Survey of Prior Literature

First, as part of this thesis, we explored the literature on conversational systems from both human- and algorithmic- perspectives. We presented a detailed account (to date) of the key papers in the last decade and beyond that contributes to the development of theory and understanding of conversational information retrieval as a domain. This includes experimental user study designs to understand users and their preferences and applications of machine learning to build such systems. However, our review of literature is comprehensive up to the time of writing the thesis. Also, the reviewed papers are mostly from IR-specific conferences and journals. Although we have tried to include relevant papers from conferences in artificial intelligence and natural language processing, our review is not exhaustive.

9.2.2 Development of New Gold Standard Dataset

Many of the prior research works have focused primarily on agent-user interactions and information-seeking dialogues. While the search actions of the agent were shared as broad themes, there was no publicly available record of the queries used, web pages visited, time spent on each page, and so on. Therefore, a major goal of our user study was to collect the search activities (performed by the agent) in addition to the spoken dialogues (by the user and the agent). We have collected four different types of data – the dialogues between the user and the agent, the search activities of the agent, and the user feedback (collected through survey and interview). We also annotated each utterance in the dataset with labels for speech acts and search actions. The CONVersation with EXplanation (CONVEX) dataset developed as part of this thesis will be made publicly available for further research, analysis, and evaluation purposes. We presented

the details of the instruments used, how the data was cleaned and processed, and how the utterances were identified. This methodology could be used for developing other datasets in the future.

9.2.3 Development of Themes and Data Annotation

We provide a detailed description of the thematic analysis performed to identify the different speech acts (performed by the user and the agent) and the search actions (performed by the intermediary). To begin with, we started with a large set of themes identified in previous studies. Through multiple rounds of annotation (using independent annotators) and inter-annotator reliability assessments, we finalized 12 themes for speech acts and four for search actions. The details of the qualitative coding process and the codebook have been shared in Chapter 5. The themes developed in this thesis could be used as-is or refined for annotating other conversational datasets. We have annotated two datasets – the CONVEX data collected by us and the SCS dataset available publicly – using the revised set of themes for speech acts and search actions.

9.2.4 Predictive Model for Natural Language Understanding

In our work, we have used the concept of Speech (or Dialogue) Acts from Linguistics to develop a natural language understanding model for the search agent. The actions of the conversational search agent – which could either be a spoken utterance or a search action – are governed by the spoken utterances of the user. In Chapter 6, we presented the details of the Multi-channel Deep Speech Classifier (MDSC) – a simple, modular, and deep neural classifier – which we developed to automatically predict the speech acts and the corresponding search actions in both CONVEX and SCS datasets. We discussed the architecture of the model, the different hyperparameters, and the different groups of features – word embeddings, lexical and syntactic features, and dialogue metadata – used by our model for the two prediction tasks.

While our model has been developed for information-seeking conversations, the modular nature of the proposed model ensures that it could be extended to conversational search datasets (which contains transcripts of searcher-intermediary conversations) with

minimal fine-tuning of hyperparameters. The proposed model furthers the natural language understanding in spoken conversational search systems. The code will be made available to the community for further analysis and applications.

9.2.5 Recommendations for New Functionalities

A brief analysis of the user interview data presented some interesting ideas for new features. We suggested alternate ways of presenting the search results on user request (which can mitigate the linear and transient nature of speech), proposed turn-based, contextual handling of list results, and called for variations in the speed of agent utterance (particularly for non-native speakers of English).

9.3 Limitations of our Work

We would like to point out some of the limitations of the work undertaken in this thesis. First, the data collected as part of our user study involved 25 participants. While the number of users was predetermined to obtain requisite statistical power, it could always be extended for a more comprehensive analysis. Also, the distribution of participants was not a perfect sample of the population (or end-users of conversational search agents). A majority of participants were female and undergraduate students in a public university. While it reduced the variance in the test subjects, the feedback received, and results obtained apply more to a younger student population. We do not have sufficient evidence to claim that the results would generalize for the entire target population.

Second, the results of the study raised some questions about the operationalization of some of the system components. In our study, we made some choices regarding when and how system revealments should be provided to the user. For example, the Wizard clarified his model of the user (queries and information sources) every time he searched. Also, the system-level clarifications were overt and not sought implicitly during the conversation. As the clarifications were frequent, it increased the number of turns and the duration of the search task and, therefore, could have potentially affected

the search and interaction experience of the users. We designed the Wizard of Oz study and developed behavioral protocols and the scripts for the Wizard. While the results suggest that certain aspects of the study design and behavioral protocols need to be altered for future work, it also provides essential insights into how the different components of the system should be implemented.

Next, the number of utterances in each of the two datasets are less than ideal for building deep neural models involving multi-class predictions. While CONVEX contained 1834 instances of speech acts, SCS contained 1043 instances. Search actions are performed only by the intermediary; therefore, the number of instances with the search actions in both the datasets was less than half of those with speech acts (509 in CONVEX and 447 in SCS). As there were 12 output labels for speech acts, and 4 for search actions, the performance of the model suffered when predicting minority classes with insufficient training data. Therefore, our model failed to identify underlying patterns in word embeddings and lexical and semantic features. As a result, both the channels showed poor performance (for search actions) and led to lower accuracy when combined.

Next, the themes for speech acts and search actions were developed by analyzing CONVEX data. We noticed some ambiguities when labeling the SCS dataset with the same set of themes. This showed that the themes might not conform to all datasets and search situations. While we firmly believe that the SCS dataset is not an ideal representation of human-system conversation, we will need to validate our themes using other datasets and make revisions as required.

The lexical and syntactic features used in one of the channels were generated using SpaCy. While these features have been used in previous research works to gain insight into the properties of written text, our data contains transcripts of spoken conversations. As such, the utility of these features was limited to the quality of the transcription. For example, even if the uttered word is a number, the transcribed text contained it as a word (a collection of alphabets and not digits), therefore, rendering the feature ineffective. The textual features are also missing important speech components like pitch, loudness, prosody, or emphasis.

Lastly, our natural language model addresses a small subset of a bigger problem, which is open domain conversation. We focused on information seeking dialogues which are task-focused and follows a specific pattern.

9.4 Directions for Future Research

As pointed out in the limitations, our study contains data from 25 users, with a high concentration of female students from a public university. We performed statistical analysis of the user study data to answer our first research question. The results highlight some interaction effect of the search tasks and participant's gender on the outcome variables. To make the results more generalizable, we plan on recruiting more participants in the future. While creating a better representative sample of the end-users is one option, we will also explore tasks of different complexities and from different topic-domains. The results from our study also provide some important insight into the different factors of study design and system operationalization; for example, how frequent system-level clarifications may deteriorate user satisfaction. For future work, we intend to make some changes in how the system revealments are operationalized and implemented. The goal would be to develop a better model for system revealment and system dialogues that can accurately portray an automated agent and will be deemed essential by the users.

We have collected a wide variety of data as part of our user study, which includes user-system dialogues, search activities performed by the Wizard, questionnaires containing user feedback, and exit interview data. We have used the user-agent interaction data and user feedback to answer our second and first research questions, we plan on using the search logs of the agent for further analysis. This could include a cost-benefit analysis of successive turns, or developing newer metrics for evaluation. Next, we plan on performing conformance analysis with the set of themes and other publicly available conversational datasets. This should help us create a uniform set of codes and frameworks for conversational search data. As many of the theoretical frameworks – which formed the foundation of this study) – were developed in the early 90s when spoken searches were not as popular, it is essential to develop new theoretical models which

would account for the current patterns of user-system interaction, user expectations, and communication strategies. We plan on aggregating our user study data with other publicly available user-system conversational data to revise the earlier frameworks for modern-day conversational search systems.

We would like to evaluate the performance of our model on other publicly available datasets that contain human-human and human-machine dialogues. The modular nature of our model could be exploited by adding additional channels or categories of data. For example, audio-based features like prosody, pitch, and loudness are significant in spoken dialogue (but were not explored in this task). Creating profiles of users, searches, and tasks could also enable us to implement them as newer categories of features. We would like to explore further how to engineer features in real-time and implement them in existing state-of-the-art systems. A possible direction for the future could be to use other pre-trained models like BERT or ELMo for contextualized embeddings and to build more transparent and explainable models that could provide necessary clarifications to the users in real-time.

Conversational search systems need to overcome a lot of potential challenges involving both the users and algorithmic development. Some of the major research directions could be to enhance long- and short-term context, profile users and tasks, develop query reformulation techniques over voice, present results in a multimodal fashion, and to develop evaluation strategies specific to conversational search. While a futuristic search agent will possess all the above-mentioned intelligent functions, this thesis explored only two 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

Pre-study Documentations

A.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: 10/11/2019
Approval Date: 9/9/2019
Expiration Date: N/A

eIRB Notice of Approval for Initial Submission # Pro2019001950

STUDY PROFILE

Study ID: Pro2019001950
Title: Exploring Intelligent Functionalities of Spoken Conversational Search Systems

Principal Investigator:	Souvick Ghosh	**Study Coordinator:**	Souvick Ghosh
Co-Investigator(s):	Katherine Ognyanova	**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
Subjects:	Unlimited	**Specimens:**	N/A	**Records:**	N/A

CURRENT SUBMISSION STATUS

Submission Type:	Research Protocol/Study	**Submission Status:**	Approved
Approval Date:	9/9/2019	**Expiration Date:**	N/A

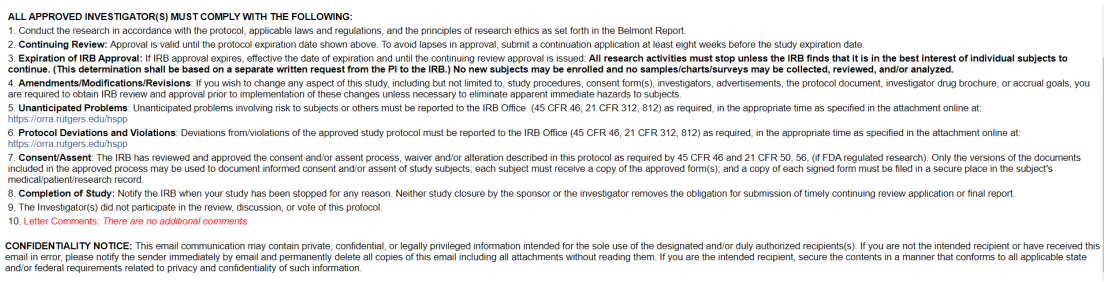
Pregnancy Code:	No Pregnant Women as Subjects	**Pediatric Code:**	No Children As Subjects	**Prisoner Code:**	No Prisoners As Subjects

Protocol:	Research Protocol Post-task Questionnaire Pre-Task Questionnaire Exit Questionnaire Entry Questionnaire	**Consent:**	Consent Form.pdf	**Other Materials:**	Recruitment Flyer

*** Study Performance Sites:**

Other Rutgers Site School of Communication and Information (Library and Information Science)

There are no items to display



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.*

CONFIDENTIALITY NOTICE: This email communication may contain private, confidential, or legally privileged information intended for the sole use of the designated and/or duly authorized recipients(s). If you are not the intended recipient or have received this email in error, please notify the sender immediately by email and permanently delete all copies of this email including all attachments without reading them. If you are the intended recipient, secure the contents in a manner that conforms to all applicable state and/or federal requirements related to privacy and confidentiality of such information.

Figure A.1: IRB Approval

A.2 Recruitment Letter

Table A.1: Recruitment Letter

Come take part in the paid user study at SC&I

Hello,

My name is Souvick Ghosh, and I am a doctoral candidate in the Department of Library and Information Science at RU School of Communication and Information (SC&I). My colleagues and I are conducting a study on a prototypical conversational system (you all have heard or used Siri, Cortana, Alexa, how cool it is to try something newer?!). You are invited to participate in this study! So you get to try something new and get paid for it!!!

In order to participate in this study, you have to be an adult who is proficient in speaking and listening to English. You also need to know how to perform basic searchers online. This study consists of one laboratory session, in which you will be required to complete one warmup task and two search tasks using our newest prototype. In this task, you will interact with (or perform searches using) a voice-based search system for all the tasks. We will ask you to answer questions related to your search experience and preferences.

The total time you will spend for this study will be approximately one hour, including a 10-minute exit-interview. You will receive $20 in cash for your participation upon completion of the study, and there will also be three prizes given to the three most active participants worth $50, $30 and $20.

Participants will be required to perform the tasks in laboratory using a voice-based search device, so it is mandatory for all the participants to schedule a study slot with the researcher. Taking part in this study will help to advance our understanding of how to design future versions of conversational search systems.

Here are the basic requirements to take part in this study:

- You must be at least 18 years old to participate.
- Proficiency in English is required.
- Intermediate typing and online search skills are required.
- Familiarity with voice-based assistants preferred.

Once you register for your participation, you will receive further instructions regarding how to go about taking part in the user study. Participation is purely voluntary. Choosing or declining to participate in this study will not affect any of your classes or grades at Rutgers. This study has been approved by the Rutgers Institutional Review Board (IRB Study # 2019001950) and will be supervised by Dr. Katherine Ognyanova (katya.ognyanova@rutgers.edu) at the School of Communication and Information. For more information about this study, please email Souvick Ghosh at souvick.ghosh@rutgers.edu. You can also contact Souvick Ghosh to ask questions or get more information about the project.

Thank you for your interest! I look forward to hearing from you. To sign-up*, please fill up this online link:

https://forms.gle/EPD6L6rRwWN7ZhpY6

*Signing up for the study does not guarantee your participation to the study due to limitation of availability. Once your signup has been confirmed, you will receive further instructions on how to proceed with the study.

A.3 Consent Form

RUTGERS
School of Communication and Information

CONSENT TO TAKE PART IN A RESEARCH STUDY

Title of Study: Exploring Intelligent Functionalities of Spoken Conversational Systems
Principal Investigator: Souvick Ghosh, PhD Candidate

STUDY SUMMARY: In this research, we explore how a conversational search agent (voice assistants) can be improved to provide better response to user commands. We automatically predict the search activities which the assistant should perform in response to the user's spoken words. We propose a voice assistant with improved responses and investigate the user's preferences while interacting with such systems.

The **purpose of the research** is to: investigate the intelligent functionalities of spoken conversational agents. If you take part in the research, you will be asked to perform three search tasks using a prototypical voice-based conversational search agent. Your time in the study will take is approximately 15 minutes per task and a total of one hour in a laboratory setting.

Possible harms or burdens of taking part in the study may be None and possible benefits of taking part may be to interact with a prototypical voice-based conversational agent.

An alternative to taking part in the research study Your alternative to taking part in the research study is not to take part in it.

The information in this consent form will provide more details about the research study and what will be asked of you if you choose to take part in it. If you have any questions now or during the study, if you choose to take part, you should feel free to ask them and should expect to be given answers you completely understand. After your questions have been answered and you wish to take part in the research study, you will be asked to sign this consent form. You are not giving up any of your legal rights by agreeing to take part in this research or by signing this consent form.

Who is conducting this study?
Souvick Ghosh [Ph.D. Candidate] is the Principal Investigator of this research study under Dr. Katherine Ognyanova (katya.ognyanova@rutgers.edu), Assistant Professor, SC&I. A Principal Investigator has the overall responsibility for the conduct of the research. However, there are often other individuals who are part of the research team.

Souvick Ghosh may be reached at sg1223@scarletmail.rutgers.edu,
Work Address:
Room 303
School of Communication and Information
4 Huntington Street
New Brunswick
NJ 08901

Phone: +1-848-237-9980

The Principal investigator or another member of the study team will also be asked to sign this informed consent. You will be given a copy of the signed consent form to keep.

Figure A.2: Consent Form (Pg-1)

RUTGERS
School of Communication and Information

Why is this study being done?
The study aims to improve the existing state-of-the-art voice-based personal agents through investigation of intelligent functionalities in such systems.

Who may take part in this study and who may not?
Anyone who has basic search skills and is fluent in speaking and listening North American English is invited to take part in the study.

Why have I been asked to take part in this study?
If you have been invited to take part in this study, you meet all the selection criteria and could help in testing a prototypical conversational search system.

How long will the study take and how many subjects will take part?
Your time in the study will take is approximately 15 minutes per task and a total of one hour in a laboratory setting.

What will I be asked to do if I take part in this study?
If you take part in the research, you will be asked to perform three search tasks using a prototypical voice-based conversational search agent. You will need to fill up an entry and exit questionnaire before and after the study respectively. In addition to that, you will have to perform a warm up task and two main tasks, each of them will not take more than 15 minutes. Those 15 minutes would include the search time and the time required to fill the pre- and post-task questionnaire. You will also have to write a response to the task question in a few lines or a paragraph. Your interaction with the search agent (the audio commands issued by you) and the video of you while issuing commands will be recorded for analysis. Audio recording is mandatory but you may refuse video recording if you wish.

What are the risks of harm or discomforts I might experience if I take part in this study?
There is no foreseeable risk and/or discomfort to participation in this study. We do not believe that this study could cause any potential harm or discomfort to you.

Are There Any Benefits To Me If I Choose To Take Part In This Study?
The benefits of taking part in this study may be an insight into future voice-based search systems and the associated novelty factor. However, it is possible that you may not receive any direct benefit from taking part in this study.

What Are My Alternatives If I Do Not Want To Take Part In This Study?
Your alternative is not to take part in this study.

How Will I Know If New Information Is Learned That May Affect Whether I Am Willing To Stay In The Study?
During the study, you will be updated about any new information that may affect whether you are willing to continue taking part in the study. If new information is learned that may affect you after the study or your follow-up is completed, you will be contacted.

Will There Be Any Cost To Me To Take Part In This Study?
There will not be any cost for you to participate in the study.

Will I Be Paid To Take Part In This Study?
You will receive $ 20.00 for taking part in this study. To get paid, you must either complete all the tasks or spent at least one hour in the laboratory engaging with the search agent to complete the task.

Figure A.3: Consent Form (Pg-2)

RUTGERS
School of Communication and Information

How Will Information About Me Be Kept Private Or Confidential?
All efforts will be made to keep your personal information in your research record confidential and private. The research records will include unique identifiers assigned to participants and not their name or email address. We will store the name and email address during the data collection period which will be then removed prior to data analysis.
There will be no linkage between participants' identities and the responses in the research as we will remove any identifiable data before the analysis. We will keep the responses anonymous and confidential by limiting individual's access to the research data and keeping it in a secure location on our password-protected servers.
The research team and the Institutional Review Board at Rutgers University are the only parties that will be allowed to see the data, except as may be required by law. If a report of this study is published, or the results are presented at a professional conference, only group results will be stated. All efforts will be made to keep your personal information in your research record confidential, but total confidentiality cannot be guaranteed.

What Will Happen If I Am Injured During This Study?
Not applicable. The study does not involve any possible threat to physical or emotional harm.

What Will Happen If I Do Not Wish To Take Part In The Study Or If I Later Decide Not To Stay In The Study?
It is your choice whether to take part in the research. You may choose to take part, not to take part or you may change your mind and withdraw from the study at any time.

If you do not want to enter the study or decide to stop taking part, your relationship with the study staff will not change, and you may do so without penalty and without loss of benefits to which you are otherwise entitled.

You may also withdraw your consent for the use of data already collected about you, but you must do this in writing to Souvick Ghosh at 4 Huntington Street, New Brunswick, NJ 08901 or email at sg1223@scarletmail@rutgers.edu

However, once the study is complete and you have been compensated for your participation, we will remove the identifiers, anonymize the data, and perform group-based analysis. It may not be possible to remove the data at that point of time.

Who Can I Contact If I Have Questions?
If you have questions about taking part in this study, you can call the Souvick Ghosh (sg1223@scarletmail@rutgers.edu) or Dr. Katherine Ognyanova (katya.ognyanova@rutgers.edu) at Library and Information Science, SC&I, 4 Huntington Street, New Brunswick, NJ 08901.

If you have questions about your rights as a research subject, you can call the IRB Director at: New Brunswick/Piscataway Art Sci IRB (832)235-2866 or the Rutgers Human Subjects Protection Program at (973) 972-1149.

PERMISSION (AUTHORIZATION) TO USE OR SHARE HEALTH INFORMATION THAT IDENTIFIES YOU FOR A RESEARCH STUDY

The study does not collect or perform any research which has health implications.

Figure A.4: Consent Form (Pg-3)

RUTGERS
School of Communication and Information

AGREEMENT TO PARTICIPATE

Subject Consent:

I have read this entire consent form, or it has been read to me, and I believe that I understand what has been discussed. All of my questions about this form and this study have been answered. I agree to take part in this study.

Subject Name (Print):_____

Subject Signature:_____ Date:_____

Signature of Investigator/Individual Obtaining Consent:

To the best of my ability, I have explained and discussed all the important details about the study including all of the information contained in this consent form.

Investigator/Person Obtaining Consent Name (Print):_____Souvick Ghosh_____

Signature:_____ Date: _____

FOR NON-ENGLISH SPEAKING SUBJECTS:
Translation of the consent form (verbal or written) must have prior approval by the IRB. Find Guidance at https://orra.rutgers.edu/hsppguidance or contact your IRB office for assistance.

SURROGATE OR LEGALLY AUTHORIZED REPRESENTATIVE CONSENT:
Use of a surrogate or legally authorized representative to consent for a research subject must have prior approval by the IRB. Find guidance at https://orra.rutgers.edu/hspp or contact your IRB office for assistance https://orra.rutgers.edu/contactus.

CONSENT ADDENDA:
Investigators seeking consent to audio or visually record aspects of the research, take photographs, or store information or biospecimens for future research secondary to a main study will find consent addenda language at https://orra.rutgers.edu/formsandtemplatesirb.

Figure A.5: Consent Form (Pg-4)

A.4 Information Sheet for Participants

Table A.2: Information Sheet for Participants

Guidelines:
1. You can begin the search sessions by using the trigger word: "Hi Joanna" OR "Hey Joanna"
2. To end the search session, you can say: "Bye, Joanna"
3. To make the assistant repeat what it just said, you can say: "Joanna, can you repeat?"
4. To stop the assistant, you can say: "Stop, Joanna"
5. To restart from the beginning, you can say: "Joanna, start over"
6. Our system should support longer discussions and long questions. However, please try not make it too long.
7. We are using an advanced prototype that performs huge computations in the background. Also, it is running on the university server which is slower than commercial servers. So the response might be slower than your regular voice assistants.
8. Please excuse small glitches because we are still in prototype/research mode on this. You can note any unusual or unsatisfactory interaction and let us know at the end of the study.

A.5 Instructions for Wizard

A.5.1 Pre-study Guidelines

Table A.3: Instruction Sheet for Wizard

Guidelines:
1. Try to follow the script as much as possible. This will ensure that all sessions are standardized and consistent. 2. Make sure that the initial response is ready when the user begins with key phrase "Hi, Joanna" 3. Every time the user asks something, acknowledge and ask him to wait. 4. Make sure of the distinctions between the two experimental systems. a. System 1: we follow the script without providing any explicit clarification (emphasized in the script) b. System 2: Clarifications 5. No response should be longer than four sentences or 50 words, whichever is lower. No need to count, just an estimate. 6. Your response should be typed into the text-to-speech system. 7. A faster response is preferable to a more detailed one. 8. End the call at the end of the study. 9. Some simple commands which the user will use: "Hi Joanna" OR "Hey Joanna" –This is to start the search session "Bye, Joanna" –This is to end the search session "Joanna, can you repeat?" –This is to repeat what the assistant just said. Replay whatever you just played. "Stop, Joanna" –This is to stop the assistant. Stop the voice from playing by muting the microphone "Joanna, start over" –This is to restart from the beginning. Start by "Hi, this is Joanna, how may I help you today?

A.5.2 Pre-study Checklist

Table A.4: Pre-study Checklist for Wizard

Checklist:
1. Check the devices: a. Is the microphone working? b. Test the green (ON)/red(OFF) light. c. Make sure the speaker is working. 2. Check if Kaltura is running in the background 3. Open Amazon TTS 4. Open Google Voice 5. Do you know which system it is? Please confirm. 6. Before I transfer to the user, do you have any questions? 7. Remember: The user will begin the conversation with “Hello Joanna”, so have your first response ready. 8. Let me know when you are ready to transfer!

Appendix B

Questionnaires

B.1 Pre-Test Questionnaire

Table B.1: Pre-Test Questionnaire

Questions
Demographic Questions:
1. Please enter your age: 2. How would you identify your gender?
Language Proficiency:
3. Native language: English /Other (please specify): 4. How would you rate your English-speaking proficiency level? (1=Novice, 5=Expert) 5. How would you rate your English listening proficiency level? (1=Novice, 5=Expert) 6. How would you rate your English reading proficiency level? (1=Novice, 5=Expert) 7. How would you rate your English writing proficiency level? (1=Novice, 5=Expert)
Search Experience:
8. How often do you search the Web per day? (Never / 1-3 searches per day / 4-6 searches per day/ 7-10 searches per day/ 10+ searches per day) 9. How often do you use personal assistants (Siri, Alexa, Bixby, etc.) for searching? (Never / 1-3 searches per week / 4-6 searches per week / 7-10 searches per week / 10+ searches per week) 10. How would you rate your level of online searching skills? (1=Novice, 5=Expert) 11. How would you rate your success while using voice-based personal assistants? (1=Low, 5=High)

B.2 Pre-Task Questionnaire

Table B.2: Pre-Task Questionnaire

Questions
Topic Knowledge:
1. How much do you know about this topic of the task? (1=nothing, 3=somewhat, 5=I know a lot) 2. What is your familiarity with the given topic? (1=not at all familiar, 3=somewhat familiar, 5=extremely familiar)
Task Complexity:
3. What is the level of difficulty of the given task? (1=not at all difficult, 3=somewhat difficult, 5=extremely difficult) 4. How difficult do you think it will be to search for information for this task using a search engine? (1=not at all difficult, 3=somewhat difficult, 5=extremely difficult) 5. How difficult do you think it will be to explain the task to the search agent? (1=not at all difficult, 3=somewhat difficult, 5=extremely difficult)

B.3 Post-Task Questionnaire

Table B.3: Post-Task Questionnaire

Questions
Task Experience:
1. The task which I performed was complex [in general] (1=completely agree, 5=completely disagree) 2. The task which I performed was difficult [to me] (1=completely agree, 5=completely disagree) 3. I completed the task successfully. (1=completely agree, 5=completely disagree)
System Experience:
4. The system was easy to use. (1=completely agree, 5=completely disagree) 5. The system was helpful. (1=completely agree, 5=completely disagree) 6. The system understood my problem. (1=completely agree, 5=completely disagree) 7. The system communication was clear. (1=completely agree, 5=completely disagree) 8. The response provided by the system was adequate. (1=completely agree, 5=completely disagree) 9. My experience with the system was satisfying. (1=completely agree, 5=completely disagree) 10. If I was searching, I would have done it differently. (1=completely agree, 5=completely disagree) 11. I would use the system again. (1=completely agree, 5=completely disagree)
Quality of Information:
12. I have succeeded in finding the right information. (1=completely agree, 5=completely disagree) 13. I think better information was available, which the system failed to provide. (1=completely agree, 5=completely disagree)

B.4 Exit Interview

Table B.4: Exit Interview Questions

Questions
1. Did you notice any difference between the two search experiences (your interaction with the system)? If yes, what were the differences?
2. In which of the two searches did you find the system to be more helpful? Why?
3. In which of the two searches did you find the system easier to use? Why?
4. Which search experience (not the search topic but your interaction with the system) did you like more? Why?
5. If I told you that there were two different systems, which one would you like to use more in your daily life?
6. Any other feedback? Merits or shortcomings?

Appendix C

Post-study Documentation

C.1 Compensation Receipt

Table C.1: Compensation Receipt

Compensation Receipt

Principal Investigator: Souvick Ghosh (souvick.ghosh@rutgers.edu)
Faculty Advisor: Dr. Katherine Ognyanova (katya.ognyanova@rutgers.edu)
Dept. of Library and Information Science
School of Communication and Information

Project Title: "Exploring Intelligent Functionalities of Spoken Conversational Systems"

IRB Study Number: #2019001950

I acknowledge receipt of $20 for participating in the research study.

Name of the Research Participant:

Signature of the Research Participant:

Name of the researcher who paid the participant: Souvick Ghosh

Signature of the Researcher:

Date:

Appendix D

Statistics

D.1 Pre-task Responses

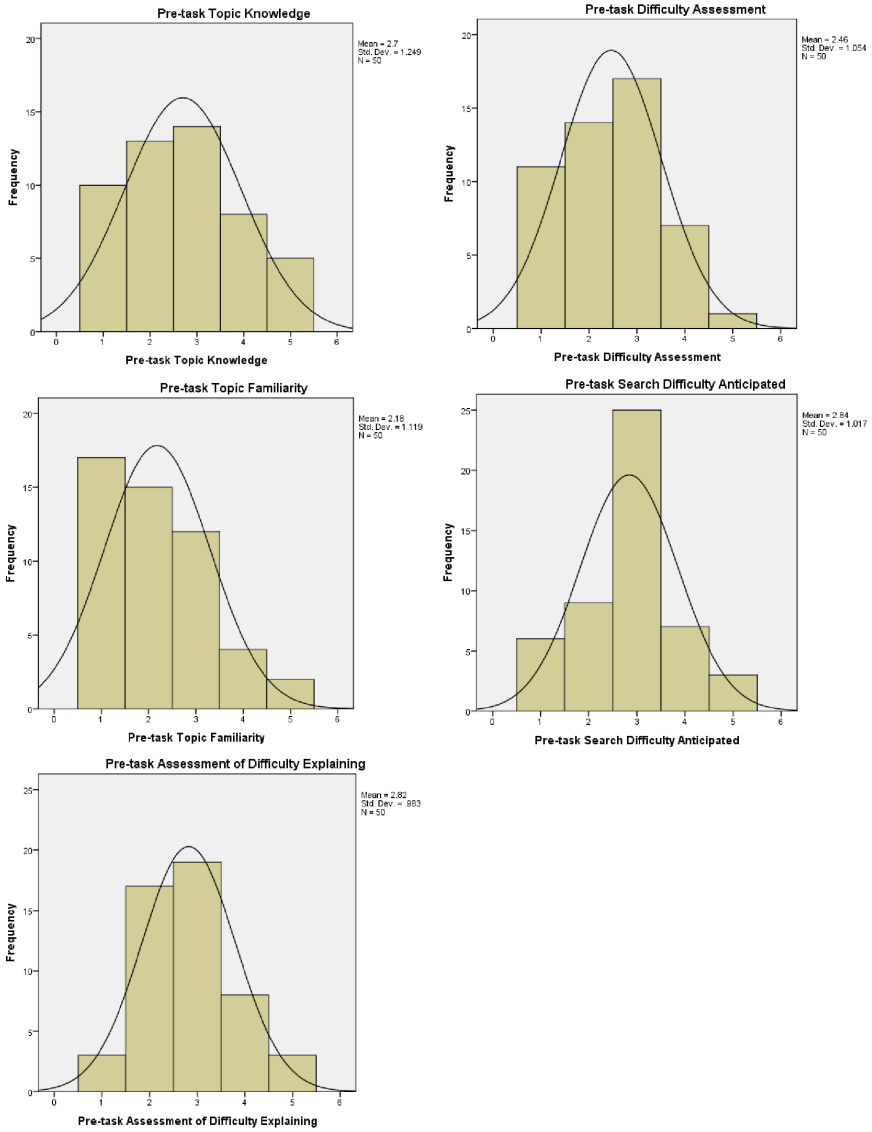


Figure D.1: Pre-task Responses

D.2 Post-task Responses

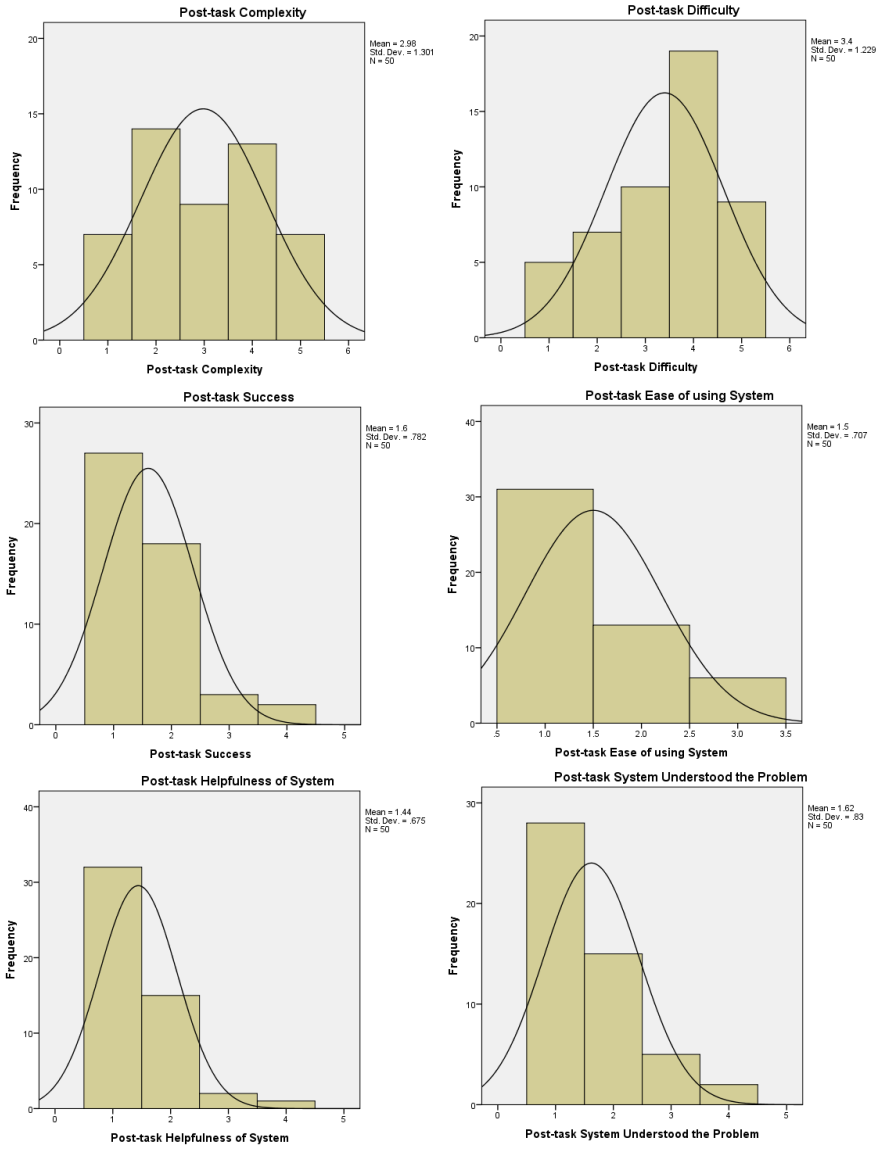


Figure D.2: Post-task Responses (Q1 - Q6)

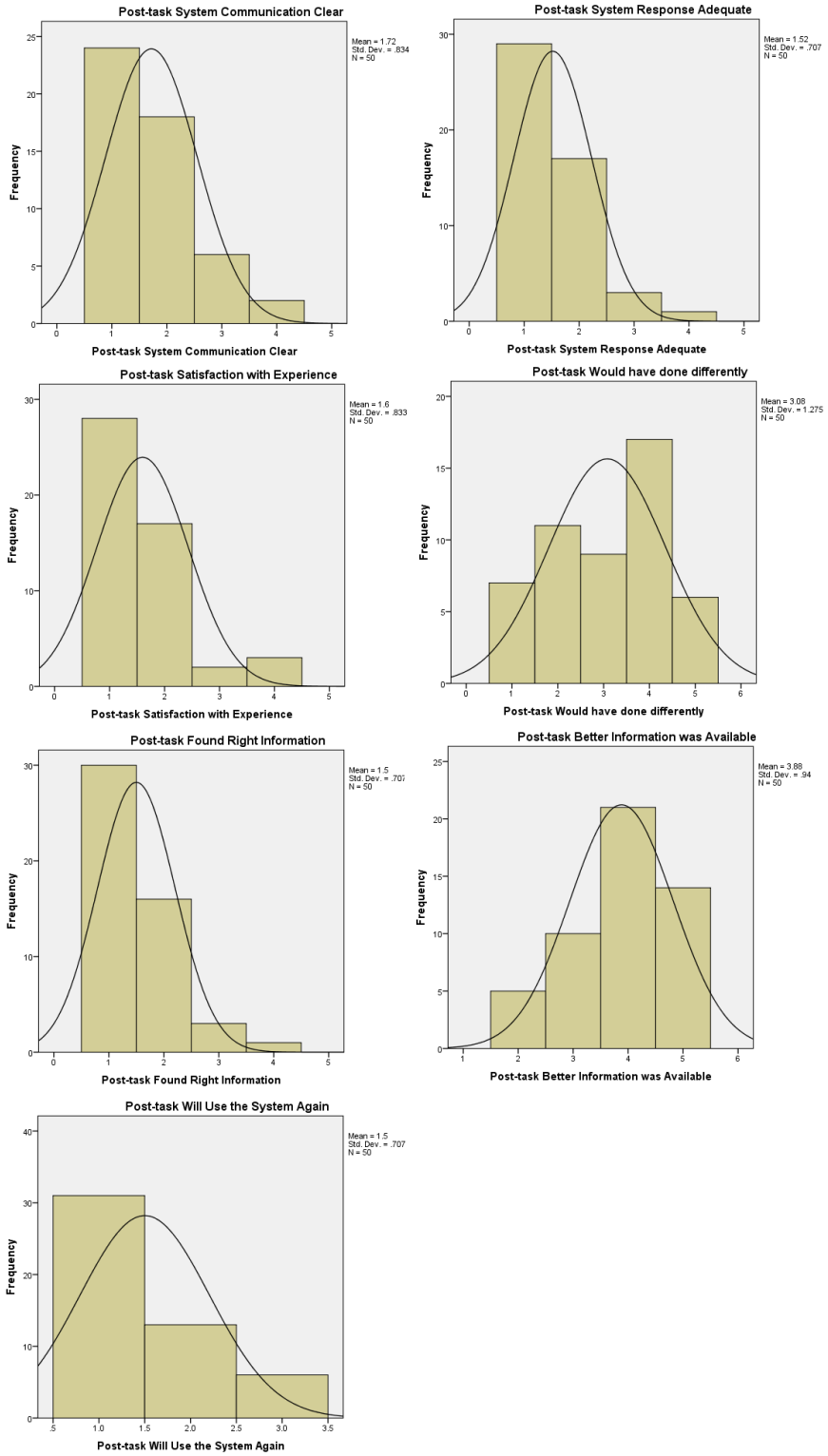


Figure D.3: Post-task Responses (Q7 - Q13)

References

Ahmadvand, A., Choi, J. I., & Agichtein, E. (2019). Contextual dialogue act classification for open-domain conversational agents. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 1273–1276).

Ajmera, J., Joshi, A., Mukherjea, S., Rajput, N., Sahay, S., Shrivastava, M., & Srivastava, K. (2011). Two-stream indexing for spoken web search. In *Proceedings of the 20th international conference companion on world wide web* (pp. 503–512).

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).

Allen, J., & Core, M. (1997). Damsl: Dialogue act markup in several layers (draft 2.1). In *Technical report, multiparty discourse group, discourse resource initiative.*

Altinok, D. (2018). An ontology-based dialogue management system for banking and finance dialogue systems. *arXiv preprint arXiv:1804.04838*.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... Wittrock, M. C. (2001). A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives, abridged edition. *White Plains, NY: Longman*.

Arguello, J., Choi, B., & Capra, R. (2017). Factors affecting users' information requests. In *Sigir 1st international workshop on conversational approaches to information retrieval (cair'17)* (Vol. 4).

Arguello, J., Choi, B., & Capra, R. (2018). Factors influencing users' information requests: Medium, target, and extra-topical dimension. *ACM Transactions on Information Systems (TOIS)*, *36*(4), 1–37.

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).

Azzopardi, L., Dubiel, M., Halvey, M., & Dalton, J. (2018). Conceptualizing agent-human interactions during the conversational search process. In *The second international workshop on conversational approaches to information retrieval.*

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review*, *13*(5), 407–424.

Begany, G. M., Sa, N., & Yuan, X. (2015). Factors affecting user perception of a spoken language vs. textual search interface: a content analysis. *Interacting with Computers*, *28*(2), 170–180.

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

Belkin, N. J. (1988). On the nature and fuction of explanation in intelligent information retrieval. In *Proceedings of the 11th annual international acm sigir conference on research and development in information retrieval* (pp. 135–145).

Belkin, N. J., Brooks, H. M., & Daniels, P. J. (1987). Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies*, *27*(2), 127–144.

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.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, *35*(12), 29–38.

Belkin, N. J., Seeger, T., & Wersig, G. (1983). Distributed expert problem treatment as a model for the analysis and design of information provision mechanisms. *Journal of Information Science*.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer

learning. In *Proceedings of icml workshop on unsupervised and transfer learning* (pp. 17–36).

Bernsen, N. O., Dybkjaer, H., & Dybkjaer, L. (1996). Principles for the design of cooperative spoken human-machine dialogue. In *Proceeding of fourth international conference on spoken language processing. icslp'96* (Vol. 2, pp. 729–732).

Bi, K., Ai, Q., Zhang, Y., & Croft, W. B. (2019). Conversational product search based on negative feedback. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 359–368).

Bickmore, T., & Cassell, J. (2005). Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems* (pp. 23–54). Springer.

Bloom, B. S., Krathwohl, D., & Masia, B. (1956). Taxonomy of educational objectives (new york, mckay). *Google Scholar*.

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Bordes, A., Boureau, Y.-L., & Weston, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation*, *56*(1), 71–90.

Borlund, P. (2002). Evaluation of interactive information retrieval systems.

Borlund, P. (2003). The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal*, *8*(3).

Boroş, T., & Dumitrescu, S. D. (2015). Robust deep-learning models for text-to-speech synthesis support on embedded devices. In *Proceedings of the 7th international conference on management of computational and collective intelligence in digital ecosystems* (pp. 98–102).

Bowden, K. K., Oraby, S., Wu, J., Misra, A., & Walker, M. (2017). Combining search with structured data to create a more engaging user experience in open domain dialogue. *arXiv preprint arXiv:1709.05411*.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *International conference on internet science* (pp. 377–392).

Brooks, H. M., & Belkin, N. J. (1983). Using discourse analysis for the design of information retrieval interaction mechanisms. In *Acm sigir forum* (Vol. 17, pp. 31–47).

Bunt, H. (1999). Dynamic interpretation and dialogue theory. *The structure of multimodal dialogue*, *2*, 1–8.

Bunt, H. (2009). The dit++ taxonomy for functional dialogue markup. In *Aamas 2009 workshop, towards a standard markup language for embodied dialogue acts* (pp. 13–24).

Bunt, H. C. (1989). Information dialogues as communicative actions in relation to partner modelling and information processing. *The Structure of Multimodal Dialague*, 47–73.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information processing & management*, *31*(2), 191–213.

Cai, R., Zhu, B., Ji, L., Hao, T., Yan, J., & Liu, W. (2017). An cnn-lstm attention approach to understanding user query intent from online health communities. In *2017 ieee international conference on data mining workshops (icdmw)* (pp. 430–437).

Cassell, J., Sullivan, J., Churchill, E., & Prevost, S. (2000). *Embodied conversational agents*. MIT press.

Chang, E., Seide, F., Meng, H. M., Chen, Z., Shi, Y., & Li, Y.-C. (2002). A system for spoken query information retrieval on mobile devices. *IEEE Transactions on Speech and Audio processing*, *10*(8), 531–541.

Chen, Y.-N., Hakkani-Tür, D., & He, X. (2015). Learning bidirectional intent embeddings by convolutional deep structured semantic models for spoken language understanding. *Proceedings of NIPS-SLU*.

Chen, Z., Yang, R., Zhao, Z., Cai, D., & He, X. (2018). Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 225–234).

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., ... Zettlemoyer, L. (2018). Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

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).

Cohen, P. R., & Oviatt, S. L. (1995). The role of voice input for human-machine communication. *proceedings of the National Academy of Sciences*, *92*(22), 9921–9927.

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.

Daniels, P. J., Brooks, H. M., & Belkin, N. J. (1985). Using problem structures for driving human-computer dialogues. In *Recherche d'informations assistée par ordinateur* (pp. 645–660).

Demberg, V., & Moore, J. D. (2006). Information presentation in spoken dialogue systems. In *11th conference of the european chapter of the association for computational linguistics*.

Deng, L. (2016). Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, *5*.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Qualitative research in information management*, *9*, 61–84.

Dhingra, B., Li, L., Li, X., Gao, J., Chen, Y.-N., Ahmed, F., & Deng, L. (2016). Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777*.

Dubiel, M., Halvey, M., Azzopardi, L., & Daronnat, S. (2018). Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The second international workshop on conversational approaches to information retrieval*.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal*

of documentation, *45*(3), 171–212.

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of documentation*, *53*(4), 384–403.

Fleming, N. D. (2001). *Teaching and learning styles: Vark strategies*. IGI Global.

Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, *5*(1), 81–99.

Frummet, A., Elsweiler, D., & Ludwig, B. (2019). Detecting domain-specific information needs inconversational search dialogues.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. (2015). Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 795–798).

Gao, J., Galley, M., Li, L., et al. (2019). Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, *13*(2-3), 127–298.

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

Ghosh, S. (2019b). Informing the design of conversational ir systems: Framework and result presentation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 1454–1454).

Ghosh, S. (2019c). Investigating result presentation in conversational ir. In *Proceedings of the 2019 conference on human information interaction and retrieval* (pp. 421–424).

Ghosh, S., Rath, M., & Shah, C. (2018). Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 22–31).

Gibbon, D., Moore, R., & Winski, R. (1997). *Handbook of standards and resources for*

spoken language systems. Walter de Gruyter.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (`http://www.deeplearningbook.org`)

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Griol, D., Carbo, J., & Molina, J. M. (2013). Bringing context-aware access to the web through spoken interaction. *Applied Intelligence*, *38*(4), 620–640.

Gupta, A. K. (2017). Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865*.

Guy, I. (2016). Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp. 35–44).

Guy, I. (2018). The characteristics of voice search: comparing spoken with typed-in mobile web search queries. *ACM Transactions on Information Systems (TOIS)*, *36*(3), 1–28.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hearst, M. (2009). *Search user interfaces*. Cambridge university press.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., ... Matsuo, Y. (2014). Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 928–939).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., ... others (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97.

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780. Retrieved from `http://dx.doi.org/10.1162/neco.1997.9.8.1735` doi: 10.1162/neco.1997.9.8.1735

Hollnagel, E. (1979). The relation between intention, meaning and action.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of documentation*, *52*(1), 3–50.

Jadeja, M., & Varia, N. (2017). Perspectives for evaluating conversational ai. *arXiv preprint arXiv:1709.04734*.

Ji, Z., Lu, Z., & Li, H. (2014). An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Joho, H., Cavedon, L., Arguello, J., Shokouhi, M., & Radlinski, F. (2018). Cair'17: First international workshop on conversational approaches to information retrieval at sigir 2017. In *Acm sigir forum* (Vol. 51, pp. 114–121).

Kelly, D., Arguello, J., Edwards, A., & Wu, W.-c. (2015). Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 international conference on the theory of information retrieval* (pp. 101–110).

Kenter, T., & de Rijke, M. (2017). Attentive memory networks: Efficient machine reading for conversational search. *arXiv preprint arXiv:1712.07229*.

Khatri, C., Goel, R., Hedayatnia, B., Metanillou, A., Venkatesh, A., Gabriel, R., & Mandal, A. (2018). Contextual topic modeling for dialog systems. In *2018 ieee spoken language technology workshop (slt)* (pp. 892–899).

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., ... others (2018). Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Kim, J., Chern, G., Feng, D., Shaw, E., & Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the workshop on web content mining with human language technologies at the 5th international semantic web conference* (pp. 5–9).

Kiseleva, J., & de Rijke, M. (2017). Evaluating personal assistants on mobile devices. *arXiv preprint arXiv:1706.04524*.

Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Predicting user satisfaction with intelligent assistants. In

Proceedings of the 39th international acm sigir conference on research and development in information retrieval (pp. 45–54).

Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., & Wang, A. (2000). Suede: a wizard of oz prototyping tool for speech user interfaces. In *Proceedings of the 13th annual acm symposium on user interface software and technology* (pp. 1–10).

Koffka, K. (2013). *Principles of gestalt psychology* (Vol. 44). Routledge.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents* (pp. 329–343).

Kotti, M., Papangelis, A., & Stylianou, Y. (2017). Will this dialogue be unsuccessful? prediction using audio features.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218.

Krathwohl, D. R., & Anderson, L. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives (abridged ed.).* New York: Longman.

Krathwohl, D. R., & Anderson, L. W. (2009). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman.

Lai, J., & Yankelovich, N. (2002). Conversational speech interfaces. In *The human-computer interaction handbook* (pp. 698–713).

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.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... others (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258.

Larson, M., Jones, G. J., et al. (2012). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends® in Information Retrieval*, *5*(4–5), 235–422.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Liu, C., Xu, P., & Sarikaya, R. (2015). Deep contextual language understanding in spoken dialogue systems. In *Sixteenth annual conference of the international speech communication association.*

Liu, J., Liu, C., & Belkin, N. J. (2016). Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology*, *67*(11), 2652–2666.

Liu, Y., Sun, C., Lin, L., & Wang, X. (2016). Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Liu, Z., Niu Z.and Nie, J., Wu, H., & Wang, H. (2017). Conversation in ir: its role and utility. *Proceedings of the SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17) (Vol. 4)*..

Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

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).

Mallios, S., & Bourbakis, N. (2016). A survey on human machine dialogue systems. In *2016 7th international conference on information, intelligence, systems & applications (iisa)* (pp. 1–7).

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

Marchionini, G., & White, R. (2007). Find what you need, understand what you find. *International Journal of Human [# x02013] Computer Interaction*, *23*(3), 205–237.

McDuff, D., Thomas, P., Czerwinski, M., & Craswell, N. (2017). Multimodal analysis

of vocal collaborative search: a public corpus and results. In *Proceedings of the 19th acm international conference on multimodal interaction* (pp. 456–463).

McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the qualitative interview: Data preparation and transcription. *Field methods*, *15*(1), 63–84.

McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface* (Vol. 6) (No. 94). Springer.

Mehrotra, R., Awadallah, A. H., Kholy, A., & Zitouni, I. (2017). Hey cortana! exploring the use cases of a desktop based digital assistant. In *Sigir 1st international workshop on conversational approaches to information retrieval (cair'17)* (Vol. 4).

Michalski, V., Charlin, L., & Pal, C. (n.d.). Towards deep conversational recommendations.

Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Bioulac, S., Sauteraud, A., & Philip, P. (2016). Acceptability of embodied conversational agent in a health care context. In *International conference on intelligent virtual agents* (pp. 416–419).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mikolov, T., & Zweig, G. (2012). Context dependent recurrent neural network language model. In *2012 ieee spoken language technology workshop (slt)* (pp. 234–239).

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American documentation*, *2*(1), 20–32.

Morel, M.-A. (1989). Computer-human communication. *The Structure of Multimodal Communication*, 323–330.

Najjar, L. J., Ockerman, J. J., & Thompson, J. C. (1998). User interface design guidelines for speech recognition applications. In *Proc. of vrais* (Vol. 98).

Nogueira, R., & Cho, K. (2017). Task-oriented query reformulation with reinforcement learning. *arXiv preprint arXiv:1704.04572*.

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69.

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

Peng, Y., Fang, Y., Xie, Z., & Zhou, G. (2019). Topic-enhanced emotional conversation generation with attention mechanism. *Knowledge-Based Systems*, *163*, 429–437.

Petrik, S. (2004). Wizard of oz experiments on speech dialogue systems. *Design and Realisation with a New Integrated Simulation Environment. Masters. Graz University of Technology, Graz. Institute of Signal Processing and Speech Communication*.

Price, D. E., Dahlstrom, D., Newton, B., & Zachary, J. L. (2002). Off to see the wizard: using a" wizard of oz" study to learn how to design a spoken language interface for programming. In *32nd annual frontiers in education* (Vol. 1, pp. T2G–T2G).

Pu, H.-T. (2010). User evaluation of textual results clustering for web search. *Online Information Review*, *34*(6), 855–874.

Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., ... Chu, W. (2017). Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 498–503).

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).

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).

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent

conversational agents. *arXiv preprint arXiv:1704.04579*.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., ... others (2018). Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Reithinger, N., & Maier, E. (1995). Utilizing statistical dialogue act processing in verbmobil. *arXiv preprint cmp-lg/9505013*.

Ren, G., Malik, M., Ni, X., Ke, Q., & Bhide, N. (n.d.). Conversational/multiturn question understanding.

Richards, M., & Underwood, K. (1984). Talking to machines: How are people naturally inclined to speak. *Contemporary ergonomics*, 62–67.

Riesbeck, C. K., & Schank, R. C. (1991). *From training to teaching: techniques for case-based its.* Lawrence Erlbaum Associates Pub.

Sa, N., & Yuan, X. (2019). Examining users' partial query modification patterns in voice search. *Journal of the Association for Information Science and Technology*.

Sahib, N. G., Al Thani, D., Tombros, A., & Stockman, T. (2012). Accessible information seeking. *Proc. of Digital Futures*, *12*.

Sahib, N. G., Tombros, A., & Stockman, T. (2012). A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the American Society for Information Science and Technology*, *63*(2), 377-391. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21696` doi: 10.1002/asi.21696

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the annual meeting-american society for information science* (Vol. 34, pp. 313–327).

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).

Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of "way of life". *Library & information science research*, *17*(3), 259–294.

Schank, R. C., Kass, A., & Riesbeck, C. K. (2014). *Inside case-based explanation.*

Psychology Press.

Schaul, T., Antonoglou, I., & Silver, D. (2013). Unit tests for stochastic optimization. *arXiv preprint arXiv:1312.6055*.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Schutz, A., & Luckmann, T. (1973). *The structures of the life-world* (Vol. 1). northwestern university press.

Searle, J. R., & Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Serban, I. V., Klinger, T., Tesauro, G., Talamadupula, K., Zhou, B., Bengio, Y., & Courville, A. (2017). Multiresolution recurrent neural networks: An application to dialogue response generation. In *Thirty-first aaai conference on artificial intelligence.*

Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., ... others (2017). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

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.*

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., & Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-first aaai conference on artificial intelligence.*

Shah, C. (2018). Information fostering-being proactive with information seeking and retrieval: Perspective paper. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 62–71).

Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Shiga, S., Joho, H., Blanco, R., Trippas, J. R., & Sanderson, M. (2017). Modelling information needs in collaborative search conversations. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 715–724).

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

Sitter, S., & Stein, A. (1996). Modeling information-seeking dialogues: The conversational roles (cor) model. *RIS: Review of Information Science (online journal)*, *1*(1), 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*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stein, A., Gulla, J. A., & Thiel, U. (1999). User-tailored planning of mixed initiative information-seeking dialogues. *User Modeling and User-Adapted Interaction*, *9*(1-2), 133–166.

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

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Sugiyama, H., Meguro, T., Higashinaka, R., & Minami, Y. (2013). Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proceedings of the sigdial 2013 conference* (pp. 334–338).

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

Tannen, D., et al. (2005). *Conversational style: Analyzing talk among friends*. Oxford University Press.

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

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In

Proceedings of the sigchi conference on human factors in computing systems (pp. 415–422).

Thomas, P., Czerwinski, M., McDuff, D., Craswell, N., & Mark, G. (2018). Style and alignment in information-seeking conversation. In *Proceedings of the 2018 conference on human information interaction&retrieval* (pp. 42–51).

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).

Tian, Z., Yan, R., Mou, L., Song, Y., Feng, Y., & Zhao, D. (2017). How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 231–236).

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. (2017a). A conversational search transcription protocol and analysis. In *Proc of sigir 1st international workshop on conversational approaches to information retrieval (cair'17), cair* (Vol. 17).

Trippas, J. R., Spina, D., Cavedon, L., & Sanderson, M. (2017b). 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). New York, NY, USA: ACM. doi: 10.1145/3020165.3022144

Trippas, J. R., Spina, D., Cavedon, L., & Sanderson, M. (2017c). 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. (2015a). Results presentation methods for a spoken conversational search system. In *Proceedings of the first international workshop on novel web search interfaces and systems* (pp. 13–15).

Trippas, J. R., Spina, D., Sanderson, M., & Cavedon, L. (2015b). Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 991–994).

Turunen, M., Hakulinen, J., Rajput, N., & Nanavati, A. A. (2012). Evaluation of mobile and pervasive speech applications. *Speech in Mobile and Pervasive Environments*, 219–262.

Tuuri, K., Eerola, T., & Pirhonen, A. (2011). Design and evaluation of prosody-based non-speech audio feedback for physical training application. *International journal of human-computer studies*, *69*(11), 741–757.

Utama, P., Weir, N., Binnig, C., & Çetintemel, U. (2017). Voice-based data exploration: Chatting with your database. In *Proceedings of the 2017 workshop on search-oriented conversational ai.*

Vakulenko, S., Markov, I., & de Rijke, M. (2017). Conversational exploratory search via interactive storytelling. *arXiv preprint arXiv:1709.05298*.

Vakulenko, S., Revoredo, K., Di Ciccio, C., & de Rijke, M. (2019). Qrfa: A data-driven model of information-seeking dialogues. In *European conference on information retrieval* (pp. 541–557).

Varges, S., Weng, F., & Pon-Barry, H. (2009). Interactive question answering and constraint relaxation in spoken dialogue systems. *Natural Language Engineering*, *15*(1), 9–30.

Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

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).

Wang, H., Lu, Z., Li, H., & Chen, E. (2013). A dataset for research on short-text conversations. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 935–945).

Wang, M., Lu, Z., Li, H., & Liu, Q. (2015). Syntax-based deep matching of short texts. In *Twenty-fourth international joint conference on artificial intelligence.*

Wang, W., Huang, M., Xu, X.-S., Shen, F., & Nie, L. (2018). Chat more: Deepening and widening the chatting topic via a deep model. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 255–264).

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., ... Young, S. (2016). A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

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

White, R. W., Jose, J. M., & Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information processing & management*, *42*(1), 166–190.

Wildemuth, B. M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the symposium on human-computer interaction and information retrieval* (p. 4).

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., & Zweig, G. (2016). End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

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

Winograd, T., Flores, F., & Flores, F. F. (1986). *Understanding computers and cognition: A new foundation for design.* Intellect Books.

Winterboer, A. K., Tietze, M. I., Wolters, M. K., & Moore, J. D. (2011). The user model-based summarize and refine approach improves information presentation

in spoken dialog systems. *Computer Speech & Language*, *25*(2), 175–191.

Write out a script with conversational turns. (n.d.). `https://developer.amazon.com/en-US/docs/alexa/alexa-design/script.html`. (Accessed: 2020-04-13)

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., & Ma, W.-Y. (2017). Topic aware neural response generation. In *Thirty-first aaai conference on artificial intelligence.*

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5934–5938).

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).

Yan, R., & Zhao, D. (2018). Smarter response with proactive suggestion: A new generative neural conversation paradigm. In *Ijcai* (pp. 4525–4531).

Yan, R., Zhao, D., & E, W. (2017). Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 685–694).

Yang, F., Mukherjee, A., & Dragut, E. (2017). Satirical news detection and analysis using attention mechanism and linguistic features. *arXiv preprint arXiv:1709.01189*.

Yang, L., Qiu, M., Qu, C., Guo, J., Zhang, Y., Croft, W. B., ... Chen, H. (2018). Response ranking with deep matching networks and external knowledge in

information-seeking conversation systems. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 245–254).

Yankelovich, N., Levow, G.-A., & Marx, M. (1995). Designing speechacts: Issues in speech user interfaces. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 369–376).

Yi, J., & Maghoul, F. (2011). Mobile search pattern evolution: the trend and the impact of voice queries. In *Proceedings of the 20th international conference companion on world wide web* (pp. 165–166).

YIN, J.-j. (2019). A compression-based bilstm for treating teenagers' depression chatbot. *DEStech Transactions on Computer Science and Engineering*(ammso).

Yuan, X., Belkin, N., & Sa, N. (2013). Speak to me: A wizard of oz study on a language spoken interface. *HCIR 2013*.

Yuan, X., Belkin, N. J., 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.

Zamora-Martínez, F., Espana-Boquera, S., Castro-Bleda, M., & De-Mori, R. (2012). Cache neural network language models based on long-distance dependencies for a spoken dialog system. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4993–4996).

Zarisheva, E., & Scheffler, T. (2015). Dialog act annotation for twitter conversations. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue* (pp. 114–123).

Zhang, X., Liu, J., & Cole, M. (2013). Task topic knowledge vs. background domain knowledge: Impact of two types of knowledge on user search performance. In *Advances in information systems and technologies* (pp. 179–191). Springer.

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).

Zhao, T., & Eskenazi, M. (2016). Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.